Explainable Course Recommendation: Connecting College Education to Knowledge and Careers Through Skills

by

Hung Kim Chau

M.S., Computer Science, University of Information Technology, VNU-HCM, 2015

B.S., Computer Science, University of Information Technology, VNU-HCM, 2011

Submitted to the Graduate Faculty of

the Department of Informatics and Networked Systems, School of Computing and Information in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

DEPARTMENT OF INFORMATICS AND NETWORKED SYSTEMS

This dissertation was presented

by

Hung Kim Chau

It was defended on

November 3rd 2023

and approved by

Peter Brusilovsky, School of Computing and Information, University of Pittsburgh

Morgan R. Frank, School of Computing and Information, University of Pittsburgh

Daqing He, School of Computing and Information, University of Pittsburgh

Zachary A. Pardos, Graduate School of Education, University of California, Berkeley

Dissertation Director: Peter Brusilovsky, School of Computing and Information, University of Pittsburgh

Explainable Course Recommendation: Connecting College Education to Knowledge and Careers Through Skills

Hung Kim Chau, PhD

University of Pittsburgh, 2024

Academic choice and exploration are essential aspects of undergraduate education in the United States, allowing students to select courses with minimal restrictions. However, students often face challenges in navigating the complex academic landscape, hindered by limited information, insufficient guidance, and an overwhelming number of choices. Time constraints from the academic calendar and high demand for popular courses make a thorough evaluation of options difficult. Although academic institutions provide career guidance counselors or advisers, the number of advisers is still limited. Course recommendation systems aim to offer personalized suggestions based on students' academic backgrounds, preferences, skills, and career goals. However, there is a lack of research on students' perceptions of recommendations and the provision of explanations to help them evaluate course relevance. Moreover, the majority of course recommender systems concentrate only on the context of learning in higher education. Despite the importance of career goals, none have attempted to establish a connection between learning and work by incorporating job information into course recommendation and explanation.

This dissertation explores the development of an advanced course recommendation system in higher education, linking academic courses to career paths using deep learning and natural language processing techniques. It begins by examining various methods for representing and recommending courses, utilizing institutional big data and combining content-based and collaborative models for better performance. A central aspect of this dissertation is the development of skill-based explanations for course recommendations. This is achieved by employing a deep concept extraction model that utilizes BERT and BI-LSTM-CRF architectures. This model effectively extracts concepts from course descriptions, enhancing the recommendation process. Using this concept extraction model, I investigate the impact of skill-based explanations in a serendipitous course recommendation system, which was tested

using the AskOski system at the University of California, Berkeley. The findings indicate that these explanations not only increase user interest, particularly in courses with high unexpectedness, but also bolster decision-making confidence. To achieve greater personalization in course recommendations, the future of this field extends beyond academics by incorporating insights from the job market to align with students' career aspirations. This dissertation introduces an innovative approach for integrating skill-related data into course recommendation systems, effectively bridging the gap between academic pursuits and career aspirations. Finally, I develop an explainable, personalized course recommendation system that incorporates insights from the job market. This system tailors course suggestions based on students' academic histories and career preferences. Its objective is to enhance course exploration in higher education, assisting students in navigating their educational paths and acquiring the essential skills required for their chosen majors and future careers. A user study conducted at the University of Pittsburgh demonstrates that the recommendations were generally perceived as valuable, with explanations playing a pivotal role in aiding students to assess their interest in the recommended courses. This underscores the significance of integrating skill-related data and explanations into educational recommendation systems.

Table of Contents

List of Tables

List of Figures

Preface

This thesis represents not only the peak of years of dedicated research and study but also a journey of personal and academic growth. It embodies the challenges faced and the knowledge acquired during my pursuit of a deeper understanding in Computing and Information Science.

I want to express my sincere gratitude to my advisor, Dr. Peter Brusilovsky, for his invaluable guidance and expertise, which played a pivotal role in shaping both the direction and execution of this research. Dr. Brusilovsky not only provided an exceptional and nurturing environment but also offered opportunities for me to collaborate with other outstanding researchers in the field, fostering my learning and growth. I am deeply thankful to my dissertation committee for their insightful feedback, which was instrumental in refining my work. In particular, I would like to extend my appreciation to my co-advisor, Dr. Morgan Frank, and my mentor, Dr. Zachary Pardos, for their substantial and direct contributions to this dissertation. Lastly, I would like to acknowledge Dr. Daqing He, who served as a mentor on several projects in which I was involved, equipping me with the knowledge and skills essential for this research.

I am deeply grateful to my esteemed collaborators whose expertise and dedication have enriched this work in numerous ways. Special thanks to Dr. Zach Pardos, whose mentorship was not just invaluable but also a source of constant inspiration. His insightful guidance and collaborative spirit significantly shaped the projects in Chapter 3.

My sincere appreciation goes to Marshall Zhao, whose meticulous work in processing course datasets for our first study in Chapter 3 was foundational. Similarly, Run Yu's creativity and precision in both the design and execution of our user studies, especially in our second project in Chapter 3 and the AskOski system user study in Chapter 5, were vital.

The user interface and frontend development of the AskOski system in Chapter 5 owe much to Run Yu's dedication and skill in data collection. His contributions were pivotal. Furthermore, I am grateful for Dr. Pardos's collaborative vision in designing our research and for the CAHL lab team's relentless efforts in backend deployment and system testing,

which were critical to our project's success.

In Chapter 6, Baptiste Bouvier's diligent processing of the OSP data was instrumental. I also extend heartfelt thanks to Dr. Sarah Bana and Dr. Morgan Frank for their groundbreaking ideas in conceptualizing the research design and their partnership in co-authoring the related paper.

Finally, I extend my gratitude to Alireza Javadian Sabet for his significant role in designing the Qualtrics surveys and preparing for our user study in Chapter 7. His meticulous approach and innovative thinking were invaluable to our research endeavors.

To my family and friends, who have provided unwavering support and patience, this achievement is as much yours as it is mine. Your belief in my abilities and your constant encouragement have been the pillars of my strength.

This thesis is not just a reflection of my academic pursuit but also a testament to the collaborative effort and support of the community around me. It is my sincere hope that this work will not only contribute to academic discourse but also inspire future research in the field.

1.0 INTRODUCTION

1.1 Motivation

Education plays a critical role in economic growth and social progress by supporting the upward mobility of individuals [1] and strengthening America's position of leadership in the global economy [2]. College degrees are generally associated with higher potential lifetime earnings, larger professional networks, and more adaptable careers [1, 3]. However, there are several indications that higher education is not successfully uplifting everyone. The symptoms include inconsistent student achievement, barriers for students who want to transfer from 2-year community colleges to 4-year degree-granting universities [4, 5], high drop-out rates (i.e., exit before degree completion) [6], and unsatisfactory career outcomes [1] (e.g., graduates are unemployed and underemployed) [7, 8]. According to the National Center for Education Statistics (NCES) in 2019,[2] roughly 40% of individuals who enroll in a four-year postsecondary program do not finish their degree within six years.

Academic exploration and choice are fundamental to undergraduate education in the United States, as colleges and universities allow students to select many of their courses with minimal restrictions [4]. In particular, multidisciplinary and liberal arts majors value and require broad intellectual exposure to various fields of knowledge. However, critics of the elective model argue that it poses risks to students who navigate a complex academic landscape with limited information and insufficient guidance [9]. Risks are particularly pronounced at community colleges with "cafeteria style" curriculums and minimal information about the sequential relationships between courses, educational pathways, and occupational goals [4]. Even at selective 4-year colleges with ample advising, the level of guidance is still limited with a national adviser-to-student ratio of one to 400 [10].

The centrality of the elective process to the organization of U.S. undergraduate education poses the difficulties that students face when making decisions about courses including

[1] Employment and Unemployment Rates: https://nces.ed.gov/programs/coe/indicator_cbc.asp
[2] Undergraduate retention and graduation rates. https://files.eric.ed.gov/fulltext/ED594978.pdf

an abundance of choices, insufficient information, incomparable alternatives, and students' limited familiarity with making academic decisions [4, 11]. These challenges are especially profound for disadvantaged groups including low-income, minority, and first-generation college students who may have limited access to college networks. In addition, time constraints imposed by the academic calendar and registration period, and high demand for popular courses render a comprehensive evaluation of all available options a formidable task.

Academic institutions are often equipped with career guidance counselors or advisers who possess a considerable amount of experience in the field, the number of advisers is still limited though [10]. However, the dynamic nature of curriculum reviews and the need to modify course structures necessitates that these advisors acquire proficiency in all forthcoming changes. While faculty advisers possess deep knowledge in their specific area of research, they may sometimes become overly focused or biased toward their field of study. On the other hand, non-faculty academic advisers have a broader familiarity with various courses, but this may come at the expense of depth. Additionally, the task of academic advising demands considerable investment of time and mental effort and is frequently overwhelmed by a deluge of student inquiries, due to the surge in both student enrollment and course offerings.

Despite these issues, there is a shortage of course guidance and information systems to support advisers and students in higher education. An example of a deployed system is Stanford's CARTA platform which is designed to enable students to search for and view detailed information on specific courses. For each course, the platform surfaces historical grade distributions, course sequences and evaluations, and common courses taken before and after a course [12, 13]. Another example is the Degree Planner[3] at the University of Pittsburgh. The system is designed to help students stay on track, make informed decisions, and achieve timely graduation while minimizing scheduling conflicts and maximizing their academic experience. It provides access to degree requirements, course offerings, prerequisites, and options for customizing the course plan based on individual preferences, goals, and course availability. Although both systems could help students effectively plan their course schedules, they lack a personalized guidance mechanism to assist students in explor-

[3]https://www.registrar.pitt.edu/degree-planner

ing courses that align with their interests and career objectives.

As technology has made it easier to collect and analyze large amounts of educational data, it has allowed for more in-depth exploration of student behavior in a variety of contexts including course enrollment [14, 15, 16]. Course recommendation systems are gaining popularity in educational institutions due to their ability to enhance the learning experience of students and augment the support provided by academic advisors. These systems aim to provide personalized suggestions based on students' academic backgrounds, preferences, skills, and career goals, thus facilitating informed decision-making. Moreover, by suggesting courses that match students' skills and interests, a recommendation system can potentially contribute to better academic performance and overall success, which ultimately leads to better student retention and graduation rates, and job opportunities.

One of the early studies on course recommendation is CourseAgent [17], a community-based recommender system. CourseAgent uses a social navigation approach to offer course recommendations by leveraging students' evaluations of a course's relevance to their career goals. In essence, it utilizes course ratings provided by students after completing courses to help future students identify the most relevant courses for their needs. Parameswaran et al. [18] developed a course recommender system called CourseRank that used two main components; i.e., students' expressed major preferences and courses they have taken in the past to make recommendations. Additionally, CourseRank considered the requirements and constraints existing for the recommended courses and recommended sets of courses rather than just individual courses.

In recent years, deep learning has gained a considerable amount of interest in course recommendations. One notable example is the implementation of the AskOski system at the University of California, Berkeley, which leverages historical enrollment data and a collaborative-based mechanism with deep learning to recommend relevant courses across the campus tailored to individual students' interests. Furthermore, it integrates with the campus degree audit system to offer personalized course suggestions that address the unfulfilled graduation requirements of the students [19]. Jiang et al. [15] developed an innovative recommendation system based on recurrent neural networks, designed to provide course suggestions aimed at preparing students for specific target courses. This system takes into

account the students' estimated prior knowledge and their zone of proximal development to generate personalized recommendations.

Although the popularity of course recommendations has increased, there is a lack of research that considers human factors. Specifically, few studies have focused on how students perceive recommendations and the provision of human-understandable explanations to aid them in better evaluating the relevance of courses. This issue is particularly significant when students are faced with high-stakes and complex decisions regarding course selection. Students must weigh the opportunity cost of selecting one course over another, and a negative experience in an introductory-level course may deter them from pursuing an entire field of study [20]. Simply reading the course titles or descriptions provided in course catalogs may not be sufficient for students to select their preferences. Providing more information about recommended courses, including why they are recommended and how they match students' skills and interests, could equip them better to evaluate the utility of recommended courses, be more confident in making decisions, and less likely to dismiss them due to unfamiliarity. This is particularly important for serendipity-focused course recommendation systems [21], which aim to recommend courses that are unexpected or novel yet still relevant and thus likely to be adopted by students. Additionally, providing explanations could increase user perception of system transparency and build trust, leading to an increased perception of relevance and acceptance of recommendations [22, 23].

Moreover, The future of course recommendations isn't just about suggesting what to study, but also integrating insights from the job market. By doing this, students can see the real-world applicability of their courses. Yet, the current course recommender systems concentrate only on the context of learning in higher education. They utilize students' enrollment history, declared majors, prerequisite information (either explicitly or implicitly), or course requirements to suggest relevant courses for students to achieve success in college (e.g., better student retention and graduation rates). One of the primary objectives for students attending college is to secure a desirable job and establish adaptable careers in the future. Existing studies either support students in selecting courses (course recommendation) or assist graduates in finding jobs (job recommendation). There is a disconnect between work and learning in the US; higher education can fail to meet the skill demands of

the labor market, and graduates may struggle to get their dream jobs. Notwithstanding the importance of career goals, to the best of my knowledge, none have attempted to establish a connection between learning and work by incorporating job information into course recommendations and explanations. Many professionals acquire skills through higher education that subsequently shape their careers. Discrepancies between skills demanded, taught, and researched have been identified by analyzing job advertisements, course syllabi, and research publications in Computer Science [24].

Furthermore, knowledge and skills have consistently emerged as pivotal elements in various educational AI systems and computational socio-economic research over the past few decades. The literature provides a broad spectrum of definitions for skills, encompassing conceptual knowledge, which pertains to the understanding of fundamental concepts, ideas, and principles underlying a specific domain or subject, as well as procedural knowledge, which relates to the expertise in executing specific tasks. For instance, in studies like those by [24, 25], the term 'skill' is employed to denote keywords, concepts, or topics in a domain, such as 'Experiments,' 'Data Warehousing,' or 'Machine learning.' Conversely, in works like those by [26, 27, 28], the term 'skill' is used to describe workplace activities, such as 'operating computer systems' or 'writing computer programming code.' In the context of this dissertation, I will adopt a comprehensive view of skills, utilizing the term to encompass various forms of knowledge, including concepts that are automatically extracted from course descriptions and workplace activities as defined by the U.S. Department of Labor.

Skills not only enable recommender systems to make reasonable decisions but also serve as one of the most intuitive ways to explain the content of documents. Most of the content-based methods for course recommendation simply use bag-of-word (BOW) representation for course matching, resulting in the limitation of using those simplistic single words for presenting and explaining recommended courses [21, 29, 30]. Comprehensive keyphrases could help to improve the recommendation and, more importantly, better communicate the underlying semantics. Keyphrases have been successfully employed to explain recommendations [31, 32, 33] and have demonstrated potential in improving user comprehension compared to unigrams [34, 35]. Skills can be identified from course catalog descriptions and job postings [24]. The capacity to connect courses to jobs through skills can empower recommender systems to

guide students towards specific courses they could take to acquire the necessary skills for their future careers, or skills that may, for instance, increase their income. In addition, this connection can support stakeholders, from a macro perspective (such as policymakers), to implement appropriate modifications for their institutions in response to the changing and evolving skills in the recent labor market.

1.2 Main Directions of Work and Contributions

This dissertation addresses the challenges previously outlined by examining the relationship between higher education and career paths. It focuses on knowledge extraction, course recommendation, and the importance of providing explanations for these recommendations. The research highlights the value of delivering personalized course suggestions that consider students' academic histories, preferences, skills, and career goals. By including skill-based justifications for these suggestions, the study aims to improve user engagement and the adaptability of the course recommendation process. Furthermore, this dissertation introduces an innovative method for incorporating skill-related data into course recommendation systems, effectively connecting academic endeavors with career objectives. Ultimately, this work seeks to enhance the course selection experience for students, guiding them in their educational journey and equipping them with the vital skills necessary for their chosen majors and future professions.

As a preliminary work [36], I explored various methods for representing and recommending higher education courses using natural language processing and deep learning. Collaborating with researchers at the University of Berkeley, we found that institutional big data can address course articulation issues. We used two datasets from a 4-year university and a 2-year community college, incorporating course enrollments and descriptions. We applied word2vec [37] to enrollment sequences and content-based models using course catalog descriptions. Our analysis revealed promising results, with the content-based model performing as well as or better than the collaborative model. Combining both models produced even better results, indicating the collaborative model's additional valuable information. In addi-

tion, our second study [38] delved into the realm of explanation within course recommender systems. We tested three distinct explanation strategies employing unigrams: 'inferred' keywords, 'anchored' keywords, and 'taken' keywords. Our user study findings indicated that leveraging students' prior knowledge was an effective strategy in generating explanations. However, it was also noted that relying solely on unigram-based skills was insufficient for crafting effective explanations.

The insights gained from these two studies have been instrumental in shaping my dissertation, which focuses on the development of skill-based explanations for course recommendations. To facilitate this, I have developed an advanced concept extraction model capable of distilling concepts from course descriptions. Skills are now represented as concepts, rather than unigrams as in the preliminary study. My method treats concept extraction as a sequence labeling task and leverages cutting-edge deep learning architectures, namely BERT and BI-LSTM-CRF. This model has been trained on various public datasets, and I have employed a stacking ensemble technique to enhance its efficacy. The results underscore the efficiency of the combined BERT and BI-LSTM-CRF models in extracting concepts from course descriptions. Furthermore, expert evaluations affirm the high quality of the concepts extracted, highlighting the model's potential for real-world applications in the educational domain.

With the effective deep concept extractor, I am able to pioneer skill-based explanations in a deep learning-based course recommendation system for higher education. I investigate the impact of skill-based explanations on a serendipitous course recommendation system. I use the AskOski system[4] at the University of California, Berkeley, which is powered by the adaptation PLAN-BERT of BERT4Rec [39], a state-of-the-art deep neural network model for top N recommendation. The serendipitous course recommendation system aims to recommend courses that are unexpected yet still relevant, with the expectation that students will be more likely to adopt the recommendations. Maximizing both aspects is challenging, especially in the university setting where unexpected, yet still relevant, courses might be found in departments outside of the student's disciplinary focus. These courses might employ unfamiliar terminology in their catalog descriptions, making them less appealing to

[4]https://askoski.berkeley.edu/

students. In this context, my hypothesis suggests that augmenting course recommendations with explanations can improve the value of the recommendation. Specifically, I propose that providing students with more information about a course and why it is recommended will equip them to assess its utility more effectively and reduce the likelihood of ignoring it due to unfamiliarity. To test this hypothesis, I conduct a user study using the AskOski system, in collaboration with the CAHL lab at the University of California, Berkeley. While our overall findings didn't show a clear impact of the explanation on course recommendations generated by PLAN-BERT, they did reveal a significant increase in participant interest in courses that exhibited high levels of unexpectedness under the proposed diversification strategy. It is apparent that individuals who received explanatory information had a positive attitude toward the usefulness of these explanations in influencing their interest in the recommendations. Furthermore, the study uncovered another crucial aspect: the significant role of explanations in bolstering users' confidence in their decision-making process. Consequently, this reduced their tendency to provide 'neutral' opinions. A thorough statistical analysis highlighted a noteworthy interaction between participants' major declaration status and the presence of explanations. Specifically, among participants who had not declared a major, the absence of explanations was linked to an increase in their likelihood to express neutral opinions, and this difference was statistically significant.

In our previous study, we utilized students' enrollment histories to find relevant courses, diversified by department information, and enhanced it with concepts extracted from course descriptions for explanation. While the future of course recommendations extends beyond academics and incorporates insights from the job market to benefit students' careers, a critical question arises: Do the concepts derived from course descriptions truly align with the skills sought after in the labor market? I aim to go beyond that and pursue more personalization for course recommendations that will be useful for students' future careers. Therefore, I am motivated to build an explainable recommendation system that aligns academic courses with real-world career goals. However, none has established a connection between college courses and graduates' careers via skills. This leads me to examining how the granular workplace activities designed and produced by the U.S. Department of Labor to describe the US workforce (i.e., O*NET Detailed Work Activity (DWA) taxonomy) could frame the knowl-

edge offered in a course, field-of-study and university, thereby connecting higher education to work. I apply word embeddings [37] and document similarity techniques from natural language processing to represent each DWA and syllabus as continuous vectors distributed in the same pre-trained language embedding space. Language embedding models enable me to describe the semantic similarity between two textual documents or sentences; here, I compare syllabus course descriptions to DWAs. As a result, syllabi are represented based on their relationships with the DWAs (called the DWA-based syllabus representation). This skill-based course representation framework is assessed through two predictive tasks: firstly, forecast the evolution of taught skills in fields of study, and secondly, predicting variations in graduates' earnings. The findings demonstrate that integrating workplace skills into the course model effectively extracts features crucial for these prediction tasks. Consequently, this approach holds promise for creating explainable, personalized course recommendation systems.

With the promising outcomes of these skill and course representation methods and the effective tools we have developed, the ultimate goal of my dissertation is to develop an explainable personalized course recommendation system that incorporates job information and skills to improve student achievement and career prospects. This system aims to guide students to specific courses that will equip them with the necessary skills for their future careers. It tailors course suggestions based on students' enrollment history and career preferences and provides explanations for the recommendation. The study is the first of its kind to utilize actual job information for career-oriented explainable course recommendations using advanced NLP techniques. It aims to assist users in making well-informed decisions and increase their trust and acceptance of recommendations. Additionally, this research investigates the two distinct approaches for representing skills within course recommendation systems. The first approach involves the automatic extraction of concepts from course descriptions using the developed deep concept extractor which also employs in the first study on explainable course recommendations. On the other hand, the second approach relies on O*NET DWAs that are manually constructed by experts to describe work in the US labor market. These DWAs are used to build the aforementioned knowledge framework, which connects college courses to graduates' careers. I developed a career-oriented explainable course recommender mech-

anism and conducted a user study at the School of Computing and Information, University of Pittsburgh. Both recommender systems showed promising results, as indicated by user feedback. Participants generally considered the suggestions valuable and showed interest in enrolling in the recommended courses. Furthermore, the explanation was shown to have a positive effect on the recommendation.

In summary, the majority of participants found the recommendations useful. They explicitly agreed or strongly agreed that the explanations provided were instrumental in helping them gauge their interest in the courses recommended. This highlights the vital role of integrating skill-related data into the system. Furthermore, it emphasizes the significance of providing explanations in educational recommendation systems.

1.3 Research Questions

My dissertation systematically investigates four primary research questions that encompass diverse domains including exploring the association between higher education and graduate career, knowledge extraction, course recommendation, and explanation. A comprehensive elaboration of the underlying motivations is provided in Section 1.1. In this section, I briefly explain the rationale behind posing each research question and explicate how each question contributes to the investigation of course recommendations and explanations in college education.

[RQ 1] Is it possible to develop an effective, automatic concept extraction model for course descriptions without requiring manually labeled data for skill-based explainable course recommendations? This question highlights the challenges in building educational taxonomies and ontologies from text. Manual construction of ontologies is an extremely time- and cost-consuming process. Automatically constructing ontologies is a complex task that necessitates advanced technology in related fields, such as natural language processing or text mining, along with training labels. The current performance of existing automatic keyphrase and concept extraction methods remains underwhelming, particularly in educational domains where annotated training data is scarce, also making it difficult to

scale and keep updated. Enhancing automatic concept extraction is not only beneficial for immediate downstream tasks such as student modeling and content-based course recommendations and explanations, but it also represents a step forward in achieving the goal of automatic educational ontology construction. Developing automatic concept extractors that do not rely on manually labeled data for model training could potentially enable the scaling of proposed recommendation and explanation approaches while simultaneously facilitating the maintenance and updating of domain knowledge.

[RQ 2] Do skill-based explanations help to improve serendipitous course recommendation systems? As explained earlier, serendipity-focused course recommendation systems aim to suggest courses that are unexpected or novel, while still maintaining relevancy. Achieving this balance can be quite challenging. By providing explanations for the recommended courses using *complete*, *comprehensive* skills achieved from RQ1, students are better equipped to assess their utility more efficiently and are less likely to dismiss these suggestions due to unfamiliarity.

[RQ 3] Can the granular workplace activities designed and produced by the Department of Labor to describe the US workforce be utilized to represent courses in college education – connecting work and learning? This question establishes a foundation for connecting colleague education to jobs. A labor market information system where work skills are shared across entities, connecting education to work, could help students know what skills they need, educators know what skills to instruct for, employers know what skills workers have, and policymakers more effectively impact workforce development. Increasingly personalized course recommendations can identify relevant topics based on students' predefined career goals (e.g., maximizing job opportunities in Business Intelligent Analytics). For instance, recommending Computer and Information Science courses that incorporate "Prompt Engineering" skills may proactively equip today's students to meet the growing demand for Business Intelligent Analytics in the labor market.

[RQ 4] Are explainable career-oriented course recommendations using job information effective in helping students explore courses relevant to their career goals? This question is the main focus of my dissertation, which examines the effectiveness of a novel course recommendation system that utilizes actual job information to guide students

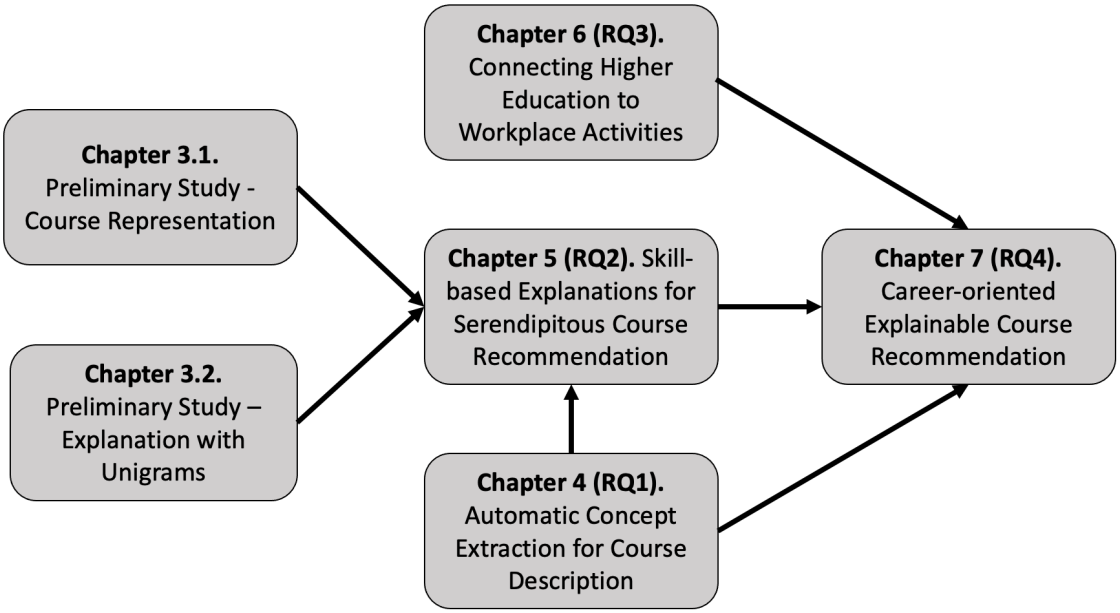


Figure 1: The dissertation structure answering the four research questions.

in selecting courses relevant to their career goals. Drawing on findings from course modeling methods explored in one of my preliminary works and RQ3, I will demonstrate how to connect courses to jobs through skills in order to develop a course recommender system. Furthermore, the ability to extract skills from text, as addressed in RQ1, enables me to experiment with the skill-based explanation for course recommendations. I will also evaluate the impact of course recommendations that are based on career goals, as well as the explanations provided, on student experiences in course selection within higher education.

1.4 Dissertation Organization

The chapters in my dissertation are organized as follows:

Chapter 2 presents a thorough literature review on the background and related works of my dissertation including automatic concept extraction, course recommendation, and explainable recommendation.

Chapter 3 reports my preliminary investigations into various methods for representing

and recommending courses in higher education using natural language processing and deep learning techniques. I examine the usefulness of institutional big data sources, including course enrollments for a collaborative-based approach and course catalog descriptions for several content-based approaches. These course representation models are evaluated using a course-to-course articulation dataset from a 4-year University of California and a 2-year California Community College. The second part of the chapter presents an early investigation into explanation in serendipity-enhanced course recommender systems. We use unigrams as skill components and provide three types of explanations for the recommended courses. We conducted a user study at the University of California, Berkeley, to evaluate the effectiveness of our proposed explanation approaches.

Chapter 4 (RQ1) presents a deep learning methodology for extracting fine-grained skills presented directly in course descriptions without requiring manually labeled course data for model training. I approach concept extraction from course descriptions as a sequence labeling task with the state-of-the-art deep learning architectures, BERT and BI-LSTM-CRF. I train multiple concept extraction models on several public datasets and then apply a stacking ensemble technique to improve model effectiveness. I conduct an offline evaluation with a small set of 50 course descriptions. The performance of the models is measured in terms of precision, recall, and F1 score. In addition to the offline evaluation, an expert evaluation is conducted to ensure the quality of extracted concepts for course recommendation applications.

Chapter 5 (RQ2) presents the design of skill-based explanations for serendipitous course recommendation systems. The system aims to provide students with comprehensive information about a course, including how it aligns with their prior knowledge and the novel knowledge it offers, students will be better equipped to evaluate its relevance and more confident in making choices. This chapter also describes an online user study conducted in collaboration with the CAHL lab at the University of California, Berkeley, utilizing the AskOski system powered by PLAN-BERT. PLAN-BERT is an adaptation of BERT4Rec, a state-of-the-art deep neural network model for top-N recommendations.

Chapter 6 (RQ3) introduces a framework for connecting higher education to work using the Detailed Work Activities (DWAs) designed and produced by the U.S. Department of

Labor to describe the US workforce. This framework represents course syllabi based on their relationships with the DWAs, referred as the DWA-based course representation. First, using earnings of graduates from the College Scorecard earnings data from the U.S. Department of Education, I demonstrate how differences in taught skills both within and between college majors correspond to earnings differences among recent graduates. Furthermore, I utilize the co-occurrence of taught skills across all of academia to predict the skills that will be taught in a major moving forward.

Chapter 7 (RQ4) introduces the framework for a personalized course recommender system that focuses on career orientation and skill-based explanations. The system aims to assist students in exploring courses that provide them with the knowledge and skills necessary for their future careers. It suggests courses based on the student's enrollment history and career preferences and provides explanations for the recommendations. I develop two recommender engines, one using DWA skill taxonomy, and the other using the concepts extracted from the course descriptions. This chapter also presents a user study with undergraduate students at the School of Computing and Information, University of Pittsburgh, to validate the importance of job information in course recommendations and test the hypothesis that explanations can improve user perception of recommendations.

Chapter 8 summarizes the key findings and conclusions derived from all the studies and analyses. Within this chapter, I will articulate the profound contributions that my dissertation brings to the forefront of the field. Additionally, I will discuss the limitations encountered during the research and outline promising avenues for future investigations.

2.0 RELATED WORK

2.1 Automatic Concept Extraction

The task of keyphrase extraction is to automatically extract a set of representative phrases from a document that concisely summarizes its content. Compared to named entities, keyphrases are more abstract, usually have vague definitions, and are heavily domain-specific, which makes it hard to obtain large-scale, high-quality annotated datasets to train and evaluate machine learning models. Nonetheless, it is important in many NLP systems including information retrieval, machine translation and recommendation. Automatic keyphrase extraction has been extensively studied and applied in many domains such as scientific articles, educational materials and bio-medical documents. There is a wide range of approaches such as rule-based, supervised learning, unsupervised learning or deep neural networks. Typically, automatic keyphrase extraction systems consist of two parts: (1) preprocessing data and extracting a list of candidate keyphrases using lexical patterns and heuristics; and (2) determining which of these candidates are correct keyphrases based on some ranking scores or trained classifiers. Other systems attempted keyphrase extraction as a sequence labeling task to capture long-term dependencies and semantic relationships of words.

Feature-based keyphrase extraction: The goal of extracting the candidate keyphrase list is to obtain all potential candidates while keeping the number of candidates as small as possible. Several studies extract candidates from words with certain part-of-speech (POS) tags [40, 41, 42, 43]. Others extract n-grams with simple filtering rules [44, 45] or only allow those matching Wikipedia article titles [46, 47]. More complex approaches extract noun phrases and apply predefined lexico-syntactic patterns [48, 49]. The next step is to score each candidate based on some features that indicate how likely that candidate is a keyphrase in the given document. Typical features for a feature-based keyphrase extraction system include; for example, statistical features (e.g., term frequency, inverse document frequency), positional features, linguistic features, context features, and external resources. Machine learning approaches to this scoring task can be grouped into supervised or unsupervised:

Unsupervised learning approaches: Mihalcea and Tarau [40] and Bougouin et al. [41] propose graph-based approaches that consider a candidate keyphrase as important if it is related to a large number of candidates and those candidates are also important in the document. Candidates and the relations between them form a graph for the input document. A graph-based ranking (e.g., PageRank) is applied to give a score to each node. Finally, the top-ranked candidates are selected as keyphrases for the input document. Unsupervised topic-based clustering methods [42, 47] attempt to group semantically similar candidates in a document as topics. Keyphrases are then selected based on the centroid of each cluster or the importance of each topic.

Supervised learning approaches: commonly frame this task as binary classification using logistic regression, SVM, tree-based models etc. A variety of features have been used for training supervised classification models including statistics-based features, title-based features, linguistics-based features or external resources such as Wikipedia [50, 51, 52, 53, 54].

Sequence tagging-based keyphrase extraction: Recent studies attempted to frame keyphrase extraction as a sequence labeling task similar to part-of-speech tagging and sequence tagging-based NER to capture long-term dependencies and semantic relationships of words. Bhaskar et al. [55] is one of the very first studies using a feature-based CRF model to extract keyphrases in scientific articles. More recently, deep neural-based sequence tagging techniques [56] applied to POS tagging and NER tasks have been adopted to keyphrase extraction problems [57, 58]. These supervised and semi-supervised techniques have been showing improvements over the traditional state-of-the-art methods. They make use of transfer learning techniques from pre-trained language models, overcome the burden of the feature engineering step and are more robust and flexible to employ in different domains. Park and Caragea [59] use pre-trained language models BERT and SciBERT with intermediate task transfer learning such as sequence tagging related tasks (e.g., POS tagging) to mitigate the lack of large amounts of labeled data. To improve extraction models beyond language modeling, the latest work on deep learning-based keyphrase extraction in open web domain has leveraged and incorporated visual features (e.g., position, font size or style) with ELMO and BERT language models into the same network for prediction [60, 61]. Another DL approach built a deep keyphrase generation with an encoder-decoder framework [62]. They applied an

RNN-based generative model to predict keyphrases.

Educational concept extraction: While many general keyphrase-extraction approaches exist, few have focused on an educational domain and almost none have considered course syllabus corpus. There are a number of projects that apply book concepts to achieve a specific target; for example, building concept hierarchies for textbooks [46] or separating prerequisite and outcome concepts [63]. However, they did not focus on advanced concept extraction and instead use existing data [63] or lightweight extraction approaches such as using Wikipedia titles and books' table of contents or index [46, 64, 65]. A related line of work has focused on building educational ontology from texts. Manual construction of ontologies is an extremely time-consuming and costly process [66, 67]. Automatically constructing ontology is a complicated task that requires advanced technology in related areas, such as natural language processing or text mining. It requires the recognition of not only concepts described in texts but also the relationships between them. The attempts to build ontologies from texts usually use existing technologies, such as NLPs or simple heuristic rules, to extract ontological concepts; for example, Shamsfard and Barforoush [66] use a simple morphological and syntactic analysis to extract primary concepts in Persian texts; Zouaq et al. [68] use a Stanford parser and KEA, a simple keyphrase extraction method, Wong et al. [67] summarize a list of studies using different strategies (e.g., statistics-based, linguistics-based or logic-based); Conde et al. [69] consider index items from a book as domain topics; and *litewi* [70] combines several unsupervised term extraction approaches and uses Wikipedia to provide additional information. However, to the best of my knowledge, the performance of existing automatic term and concept extraction methods remains underwhelming, especially in educational domains. Improving automatic keyphrase extraction is not only useful for immediate downstream tasks such as student modeling in intelligent textbooks and course recommendations, but is also a step forward in accomplishing the task of automatic educational-ontology construction.

FACE [71], a supervised feature-based machine learning method for automatic concept extractions from textbooks, is proposed and evaluated with a newly constructed dataset. The model is engineered and experimented with a highly encompassing feature set for machine learning to extract the annotated concepts; the feature set spans both linguistic features

and features encoding relative corpus statistics. It is systematically evaluated and compared with a number of keyword extraction models proposed in the literature. The results show that FACE outperforms several traditional state-of-the-art keyphrase extraction methods.

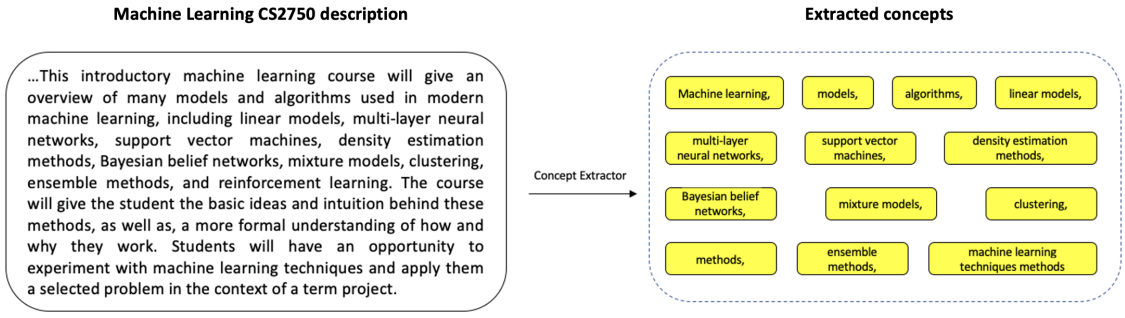


Figure 2: Given a course catalog description that contains a chunk of texts, the task of concept extraction is to identify a list of concepts presented in the description. This example shows the description of a Machine Learning course and a concept extractor expected to provide a list of concepts that represent the content of the course.

Although, the FACE framework shows promising results for intelligent textbooks. However, labeling data to train traditional supervised concept extraction models is very time-consuming and error-prone [72], especially for a large, multi-subjects corpus like course syllabi. Moreover, course descriptions do not have the rich structures as textbooks, rather just contain a chunk of unstructured texts (see Figure 2). We need a more general framework that is able to intrinsically capture the semantic and syntactic structures of concepts mentioned in a sentence or paragraph of a course description. There are several techniques and resources are shared and overlap between named-entity recognition (NER) and concept extraction. The advance of deep learning and transfer learning in NLP makes it possible to apply advanced sequence tagging and labeling models such as Bi-directional Long Short Term Memory with Conditional Random Field (Bi-LSTM-CRF) or Bidirectional Encoder Representations from Transformers (BERT), which have shown great success for NER, to concept extraction for course syllabi. In this thesis, I will build concept extraction models for course descriptions using Bi-LSTM-CRF [56] and BERT [73] with pre-trained word embeddings and language models.

2.2 Course Recommendation

Course recommendation is an area within personalized, adaptive systems that has recently attracted much attention from researchers. Course recommendation systems have become an essential component of higher education, especially in the context of multidisciplinary and Liberal Arts programs, where students are often faced with a vast array of course options to choose from with limited guidance. Based on a survey conducted on 81 students at Kyushu University, Japan, the main underlying reasons for student's course selection include *interest*, *high-grade*, *learning goal and career plan*, *social aspect*, and *popularity*. The primary objective of these systems is to offer personalized course suggestions that meet students' academic backgrounds, preferences, learning and career goals. By doing so, course recommendation systems can enhance the overall learning experience and contribute to improved academic outcomes and job opportunities for students. However, the reasons behind course selection are manifold and there are challenges in building such course guidance and information systems [74, 75]. The problems related to course selection can differ significantly depending on the context and specific circumstances. For instance, students are often faced with a wide range of courses to choose from, which can lead to difficulties if they possess limited knowledge about the various options available. Moreover, the course selection process should ideally consider individual student's career preferences in order to provide tailored recommendations that equip them with necessary skills for their future jobs. Furthermore, it is important for a course recommendation system to adapt to changes in course content and structures and the unique circumstances of individual students. This may include updates to course curricula, new course offerings, or shifts in a student's academic focus or interests. Additionally, academic institutions often present complex course structures, with some courses featuring overlapping content or requiring prerequisite courses to be completed beforehand.

For next course recommendation, studies utilized students' preferences, interests and performance [76, 77, 78], students' assessment of course relevance and ratings [17, 79, 80], course content [81, 29, 82], and sequential enrollment histories [83, 19, 84] to employ predictive models. The early course recommendation systems primarily relied on content-based

filtering and collaborative filtering methods.

Content-based recommender systems use the features and attributes of items to provide personalized recommendations to users. These systems typically rely on the similarity between items to suggest relevant content based on a user's preferences, interests, or past behavior. Course content-based recommender systems used the similarity between course features (i.e., descriptions and subject areas) and student preferences (e.g., majors and previous course subjects) to make recommendations. Some studies used student input query as an approximate for their subject interest, and returned a list of recommended courses in the Information Retrieval fashion. While Gulzar et al. [81] enhanced the recommendation list by using the knowledge of a Computer Science ontology, Morsomme and Alferez [82] showed a warning for courses too advanced based the student's past performances and provided suitable preparatory courses. Other content-based methods used course features such as the subject area, the contents, the professors and the competencies, each course content is represented as a word frequency vector; and cosine similarity is applied to the course vectors to find similar courses to those the students have already taken [29, 85]. Course categories were also utilized in content-based systems. Students' preferences and interests are estimated based on the categories of the courses they took in the past, represented as *preference* vectors. Based on the preference vectors and the student's majors, the user interest-based scores of the courses taken by similar students were computed and used to make recommendations to the target student along with the information about the timing and popularity of courses, and predicted performance of students Ma et al. [78]. K-Nearest Neighbor and Naïve Bayes classifiers were used to design a content-based recommender system by Neamah and El-Ameer [86]. The authors built students' user profiles based on their prior knowledge and actions such as enrolling and rating courses, and compared the user profiles with course attributes to generate recommended courses. Other studies also utilized course catalog descriptions [21, 30, 36]. The course description is simply represented with a bag-of-words model. These course vectors with word frequency or td-idf are used to find the similarity between courses.

Collaborative filtering-based recommendation systems, on the other hand, predict user preferences based on their similarity to other users. These systems analyze the

behavior of many users and identify patterns or similarities in their choices, preferences, and interests. The collaborative filtering algorithm collects user preferences data from a dataset and identifies users with similar preferences to the target user. It then suggests items that those similar users have liked or preferred in the past, to the target user. Ray and Sharma [76] presented one of the early works that extended the collaborative filtering approach to develop a course recommendation system. Their approach involved utilizing past student performance data to predict grades for elective courses. Specifically, the authors applied both user-based and item-based collaborative filtering techniques for recommending courses. The results demonstrated the potential of collaborative filtering in developing course recommendations. Houbraken et al. [87] also utilized the grades of a large group of students to discover course-competency requirements and student-competency levels using matrix decomposition. The authors showed that hidden features are responsible for observed grades, which are then translated into human understandable competencies by matching computed values to expert input. Their approach also enabled personalized study planning and student guidance through grade prediction for unobserved student-course combinations. Backenköhler et al. [88] proposed a collaborative filtering-based course recommendation system that combined grade prediction from student course performance data with statistical methods based on course orderings. The model aimed to capture the expected performance and preparedness of students for a given course, allowing for meaningful course recommendations for both new and senior students.

Instead of using past course performance data, Elbadrawy and Karypis [89] utilized course enrollment history and the popularity of courses. They experimented with various techniques, including neighborhood-based user collaborative filtering, matrix factorization, and popularity-based ranking, to address the task of ranking the top N courses. The authors analyzed the enrollment patterns and demonstrated the impact of student and course academic features on these patterns. To account for this, they established multi-granularity groups for both students and courses. The authors incorporated these groups into their user collaborative filtering, matrix factorization, and popularity ranking algorithms. The results indicated that incorporating these groups led to lower grade prediction errors and more accurate top N course rankings. Only using course enrollment data, Houbraken et al.

[87] presented a Markov-based collaborative filtering model for course recommendation to students based on the sequence of courses they have taken in the previous semesters. The proposed model incorporated the dynamics of student course-taking behavior and provided personalized course recommendations to students at each semester. In contrast, Bakhshinategh et al. [80] proposed a course recommendation system for students based on their "graduate attributes", which are values that students develop throughout their studies. The system utilized a collaborative filtering algorithm, where students rate the improvement in their graduate attributes after completing a course, and then recommended courses taken by other students who rated similarly. The authors experimented with the proposed method using a synthetic dataset and demonstrated the importance of considering the time dimension of student ratings for more accurate recommendations.

Knowledge-based recommender systems utilize explicit knowledge about the user's preferences and the characteristics of items to generate personalized recommendations to users. Unlike other types of recommender systems that rely on user data and behavior, a knowledge-based recommender system typically relies on domain expertise, ontologies, taxonomies, or other knowledge representation techniques to make recommendations. These systems can provide more explainable and interpretable recommendations, as they take into account explicit user preferences and domain knowledge. Knowledge-based systems are often combined with other approaches to improve the recommendation. In their work, Ibrahim et al. [90] presented a framework for a personalized course recommendation system that utilizes a hybrid-filtering approach and is based on ontologies. The goal of this system was to improve both the efficiency and user satisfaction of the recommendation process by integrating information from multiple sources, and by utilizing a hierarchical ontology similarity measure to provide students with appropriate recommendations. Gulzar et al. [81] integrated both content-based and knowledge-based methodologies in the development of their course recommender system. They developed an ontology to model knowledge for the Computer Science domain. Utilizing this ontology, the system was able to search for relevant courses and improve the accuracy of recommendations in response to students' query inputs.

Other hybrid course recommender systems aim to integrate the benefits of content-based and collaborative signals with expert knowledge, ultimately resulting in more accurate

and personalized recommendations. Bydžovská [77] incorporated various factors into their course recommendation system, such as course popularity, courses taken by students with similar profiles, and courses chosen by the student's friends (which highlights the social navigation aspect). This system offered recommendations for selective and optional courses, taking into account students' abilities, knowledge, interests, and available time slots in their schedules. Furthermore, the system provided cautionary alerts regarding challenging courses and reminded students of their mandatory academic duties. Esteban et al. [85] proposed a hybrid approach that combines collaborative filtering and content-based filtering. Their approach utilizes multiple criteria related to both student and course information to recommend the most suitable courses to students. To optimize the performance of their recommender system, they employed a genetic algorithm that automatically discovered the optimal configuration, including the most relevant criteria and configuration parameters. Morsy and Karypis [91] proposed grade-aware recommendation techniques that take into account the predicted grades of students when recommending courses. In one of their approaches, the authors leveraged collaborative-based recommendation methods and combined them with grade prediction models to enhance the accuracy of course rankings. In their study, Pardos et al. [36] explored various methods for representing courses, including collaborative-based models (i.e., course2vec) and content-based models (such as bag-of-words models with term frequency and TF-IDF, as well as word2vec). The authors then assessed the effectiveness of each model using course-to-course articulation data from two institutions. The results indicated that the combination of the course2vec and DescVec models yielded superior performance, outperforming all other models.

Other traditional recommendation methods were also applied to course recommender systems. Xu et al. [92] introduced a methodology including two algorithms: a forward-search backward-induction algorithm that selects course sequences to decrease the time required for graduation while considering prerequisite requirements and course availability, and a multi-armed bandit algorithm that recommends a course sequence to both decrease graduation time and increase overall academic performance. The system dynamically learned how students with different contextual backgrounds perform for given course sequences, and then recommended an optimal course sequence for new students. Scholars

Walk proposed in [84] was a random-walk-based approach that captures the sequential relationships between the different courses. The system leveraged both the "wisdom of the crowd" and students' prior course histories to generate a short list of recommended courses for the upcoming semester. The experimental evaluation conducted by the authors demonstrated that Scholars Walk outperformed other collaborative filtering and popularity-based approaches. CourseAgent was a community-based system for recommending courses. It was developed by leveraging course ratings provided by a community of students [17]. By using this approach, the system transformed the process of course rating into a valuable activity for users, helping them to better track their progress toward their career goals. As a result, CourseAgent provided students with useful rating information about courses to consider when enrolling. Non-model-based course recommender systems were also introduced; for instance, displaying analytics to students drawn from aggregate course evaluations [12], or presenting scheduling interfaces for accommodating course enrollment requirements [93]. Parameswaran et al. [18] focused on degree requirements and constraint satisfaction as priorities for the recommendation.

Deep learning-based recommender systems have gained immense popularity in the field of course recommendations in recent years. Jiang et al. [15] developed a novel recurrent neural network-based recommendation system for suggesting courses to help students prepare for target courses of interest, personalized to their estimated prior knowledge background and zone of proximal development. Wong [94] proposed a solution that also used recurrent neural networks to develop a sequence based course recommender to deal with challenges including sequence and concurrency, constraints, context and concept drift. In their work, Pardos et al. [19] leveraged a combination of course catalog descriptions, student enrollment histories, and student grades to develop a neural network-based system for personalized course guidance. Specifically, they implemented recurrent neural networks and skip-gram models to enhance the scrutability of the system. This approach was shown as an evolution of content tagging and provided a means for the recommender system to balance inferred user preferences with those explicitly specified by the user. A recent study [95] introduced PLAN-BERT, a model that utilized Bidirectional Encoder Representations from Transformer (BERT) to support consecutive basket recommendation with pre-specified future reference

items. Through their empirical analysis, the authors highlight the value of pre-specified reference items in addressing the cold start problem. Furthermore, their findings suggested that PLAN-BERT's bidirectional self-attention architecture outperforms both BiLSTM and a UserKNN baseline in effectively utilizing past sequence information. Furthermore, the authors demonstrated that incorporating user and item features offers substantial benefits, particularly for students in later years of study.

Although course recommendation has gained considerable attention, limited research has focused on human factors, such as human-understandable explanations, and different dimensions of recommendation systems, such as *novelty* and *serendipity*. This issue is particularly significant when students are faced with high-stakes and complex decisions regarding course selection. One proposed design for a serendipitous university course recommendation system is presented in [21]. The authors experimented with a variety of algorithms and found that context-based models performed best on offline validation tasks. However, these models underperformed compared to a simple bag-of-words model on student measures of serendipity. Providing more information about recommended courses, including why they are recommended and how they match students' skills and interests, could better equip the students to evaluate the utility of recommended courses and reduce the likelihood of dismissing them due to unfamiliarity. This is particularly important for serendipity-focused course recommendation systems, which aim to recommend courses that are unexpected or novel yet still relevant, and thus likely to be adopted by students.

Furthermore, to the best of my knowledge, no previous studies have attempted to incorporate job information into course recommendation and explanation.

2.3 Explainable Recommendation

Explaining the mechanisms of recommendation algorithms is a rapidly evolving area of research in explainable AI. The goal of explainable recommendation systems is to not only provide users or system designers with recommendation results but also to provide explanations for why certain items are recommended. As one of the most user-centered types

of AI systems, recommender systems face the challenge of maintaining user trust and satisfaction [96]. Recommender systems often behave like a "black box", particularly those employing deep neural networks [19]. These systems present recommendations to users without providing a rationale for selecting recommendations [97]. Concerns about the interpretability and explainability of recommendation algorithms have been raised, particularly as researchers have started to apply more advanced and complicated algorithms, such as Latent Factor Models and Deep Learning, to improve the statistical accuracy of these systems in offline settings (e.g., rating prediction tasks). By providing explanations and justifications for recommendations, transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction of recommender systems can be improved.

Some of the earliest studies on explainable recommendations focused on foundational concepts. For example, Schafer et al. [98] described how recommender systems improve E-commerce sales by suggesting items that users are already familiar with, leading to the development of item-based collaborative filtering. Herlocker et al. [99] proposed a model for explanations based on user surveys to understand their conceptual model of the recommendation process in MovieLens. Additionally, Sinha and Swearingen [100] emphasized the importance of transparency in recommender systems, rather than just focusing on the system's statistical accuracy.

In the late 2000s, as recommender systems became more complex and were applied to a wider range of fields, there was an increase in research on recommendation explanation [101, 102, 103, 104]. These studies aimed to enhance recommender systems with the ability to provide explanations, which were shown to increase user perception of transparency [105] and build trust [106]. This increased trust led to a greater acceptance and perception of the relevance of recommendations [22, 23]. The promising results of early research on explanations encouraged a large body of follow-up work, which has been summarized in [107, 108, 109].

Modern explainable recommendation systems aim to develop models that not only generate high-quality recommendations but also provide intuitive explanations. These explanations can be categorized based on a variety of factors, such as display styles, reasoning models, paradigms, sources of knowledge, and explanatory goals. In particular, there are

several important explanatory goals, including *Transparency*, *Scrutability*, *Trust*, *Persuasiveness*, *Effectiveness*, *Efficiency*, and *Satisfaction* [110, 107, 109]. Recognizing the broad diversity of explanation styles has led to research comparing different types of explanations for the same recommendations [111, 112, 113].

The original research on explaining recommendations primarily focused on model-intrinsic approaches, which aim to develop interpretable models that have transparent decision-making mechanisms, thus providing natural explanations for the model decisions. This approach aims to explain *how* models suggest items and has been studied extensively in the literature [33, 99, 114, 115, 116]. However, a growing sub-stream of research has shifted towards model-agnostic approaches, also known as post-hoc explanations or justifications. Post-hoc explanations focus on developing an explanation model to generate explanations after a decision has been made, rather than revealing the process of recommendation. Furthermore, some justification models have been developed to address the challenges of black-box models, whose internal behavior is difficult to interpret, by specifying *why* the system provides certain recommendations. This approach has been studied in various contexts, such as in trust-based recommender systems [106] and tag-based explanations [105]. Post-hoc explanations have become increasingly important with the rise of deep recommender systems based on neural network mechanisms, which are notoriously difficult to explain [117, 118, 119].

To address this issue, Ni et al. [117] proposed a pipeline to identify justification candidates and built aspect-based user personas and item profiles from massive corpora of reviews. They also proposed two generation models to improve generation quality and diversity. The experiments conducted on two real-world datasets show that the models are capable of generating convincing and diverse justifications. Shmaryahu et al. [120] introduced post-hoc explanations for complex model recommendations using simple methods, such as simple collaborative filtering and content-based approaches. The study showed that users accept these explanations and react positively to simple and concise explanations, even if they do not fully explain the mechanism leading to the generated recommendations. A framework for generating post-hoc natural language justifications for recommendations was proposed in [121]. The authors used user reviews to build an explanation that was independent of the underlying recommendation model. Three different implementations of the framework were

presented, each with increasing complexity: the first used NLP and sentiment analysis to identify relevant aspects discussed in the reviews; the second introduced automatic aspect extraction and text summarization; and the last implementation generated context-aware justifications by learning a lexicon for each contextual setting. The framework was validated through three user studies, which showed that it made the recommendation process more transparent, engaging, and trustworthy for users. Mauro et al. [122] proposed a novel justification approach that uses service models to extract experience data from reviews concerning all the stages of interaction with items and organize the justification of recommendations around those stages. In a user study, the approach was compared with state-of-the-art baselines, and participants evaluated the service-based justification models higher for perceived user awareness support, interface adequacy, and satisfaction. The study also suggested the investigation of personalization strategies to suit diverse interaction needs.

In recent years, the field of recommendation explanation has garnered increasing interest; however, the majority of the research conducted has been within the context of e-commerce, with only a limited focus on the education domain. Zhou et al. [123] discussed a study on using hierarchical reinforcement learning induced pedagogical policies combined with human-authored explanations to improve student-system interaction in terms of their engagement and autonomy. The study found that reinforcement learning decisions can be paired with explanations to achieve better outcomes than using either one alone. Barria-Pineda et al. [124] discussed the use of explainable educational recommendations in a personalized practice system for Introductory Java Programming course. The study examined how students use recommendations and explanations and assessed their impact on the educational process. Two types of explanations were presented to justify the recommendation of the next learning activity. The authors evaluated the effectiveness of these explainable recommendations in a semester-long classroom study, analyzing their impact on various aspects of student behavior and performance. Takami et al. [125] introduced a simple explanation generator using Bayesian Knowledge Tracing (BKT) models to generate explanations for a quiz recommender system. The explanation generator classified recommended quizzes into distinct feature types based on the values of the BKT model parameters. Subsequently, it generated explanation texts corresponding to each category, which were manually created by a human

teacher. To assess the impact of explanations on the recommender system, the authors carried out a user study involving summer vacation assignments given to high school students. Comparing the click counts of recommended quizzes with and without explanations, the study found that the number of clicks was significantly higher for quizzes with explanations.

One of the pioneering studies conducted by Ma et al. [74] discussed explanations in course recommendation systems. The authors aimed to investigate the factors influencing students' course selection in universities with the intention of gaining insight into student perceptions, attitudes, and needs. This understanding could facilitate the employment of data-driven methodologies for recommending courses and explaining recommendations in the context of university environments. Upon analyzing the survey data, they identified that course interestingness, career objectives, and academic performance were among the most significant factors in course selection. However, the study lacked a comprehensive evaluation of the impact of explanation in course recommendations. One study by Yu et al. [38] explored ways of familiarizing students with recommendations by providing explanations designed with varying levels of personalization in a serendipitous course recommender system. They used unigrams to represent knowledge components in the descriptions of anchor courses, target courses, and taken courses. However, keyphrases may better communicate the underlying semantics, as unigrams may not have sufficient ability to convey the meaning encapsulated in course descriptions. It may not be easy for students to interpret the meaning of single words, especially technical, without context. Keyphrases have been successfully used to explain recommendations [31, 32, 33] and have shown promise in improving user comprehension over unigrams [34, 35].

3.0 PRELIMINARY WORK

In this chapter, I report my preliminary investigations into various methods for representing and articulating courses in higher education applying natural language processing and deep learning methods. I assess the value of large-scale institutional data sources, such as course enrollments for a collaborative approach and course catalog descriptions for multiple content-based strategies. These course representation models are evaluated using a dataset of course-to-course articulation from a 4-year University of California and a 2-year California Community College. The latter part of the chapter explores an early investigation into the role of explanation in serendipity-enhanced course recommendation systems. We employ unigrams as skill components and offer three types of explanations for the suggested courses. A user study at the University of California, Berkeley, is conducted to gauge the efficacy of the explanation on the course recommendation. These studies were conducted under the mentorship and guidance of Dr. Zach Pardos and in collaboration with his students.

3.1 Data-Assistive Course-to-Course Articulation Using Machine Translation

Higher education at scale, such as in the California public post-secondary system, has promoted upward socioeconomic mobility by supporting student transfer from 2-year community colleges to 4-year degree granting universities. Among the barriers to transfer is earning enough credit at 2-year institutions that qualify for the transfer credit required by 4-year degree programs. Defining which course at one institution will count as credit for an equivalent course at another institution is called course articulation, and it is an intractable task when attempting to manually articulate every set of courses at every institution with one another. In collaboration with researchers at the University of Berkeley, California, we study a methodology towards making tractable this process of defining and maintaining articulations by leveraging the information contained within historic enrollment patterns and course catalog descriptions. We provide a proof-of-concept analysis using data from a 4-year and

2-year institution to predict articulation pairs between them, produced from machine translation models and validated by a set of 65 institutionally pre-established course-to-course articulations.

As this study lays the groundwork for subsequent chapters on course representation and recommendation, it is essential to explain the comprehensive methods and insights derived from this research.

3.1.1 Introduction

Course articulation has been the bridge that connects programs from different levels of higher education to one another, forming pathways to achievement focused on equity of access. Across the United States, there is evidence that these pathways have been underperforming. Around 45% of the 20 million students entering higher education in the United States begin their post-secondary experience at 2-year public institutions [126]. A 2010 US Department of Education survey of 19,000 "Beginning Postsecondary Students" (BPS) found that 81.4% of community college students had aspirations of transferring to earn a 4-year degree [127]. Data on 852,439 public community college students, collected by the National Student Clearinghouse (NSC); however, found that only 13% had earned a 4-year degree in six years after beginning at a community college [128]. The picture looks better for those in the study who successfully transferred, with 42% of these students having completed their 4-year degree. Course articulation, or a lack of it, is not the primary culprit for these low outcomes; however, it is likely not an insignificant source either. Evidence of this is an analysis of the BPS data from the US Government Accounting Office's (GAO) in which it is estimated that 42% of credit earned at the community college level is lost upon transfer to a 4-year institution [129]. Much of this loss is due to switching majors or earning an excess of general credit before declaring a major, though it is estimated that a portion is due to lack of articulation and that a 20% increase in 4-year degree attainment, among transfers, can be expected if those articulations existed [130]. The impact of insufficient articulation on student rates of successful transfer has not been quantified, but a recent spate of state and national efforts to define additional pathways suggest that it has been an im-

portant factor [131, 132, 133]. These observations serve as mounting evidence that providing more comprehensive articulations can help improve transfer success through greater credit mobility.

Articulations at the degree level are often created by state mandate, with courses being developed at the 2-year and 4-year institutions in unison, or one modeled after the other, and with collaboration between faculty at both. Outside of these degree level articulations are those made on a course-by-course basis. In this case, there is a significant bottleneck of human resources committed to processing and validating requests for articulation. Each campus typically has a designated articulation officer, or chief instructional officer [134]. This person is responsible for receiving articulation requests, choosing which to consider, and then beginning the process of validation by conferring with the instructor of record at the other institution by way of its respective articulation officer. A diagram of the articulation process in the California public post-secondary system [135] is depicted in Figure 3. If only considering articulation of courses from one of the 115 2-year California Community Colleges (CC) to courses at the nine 4-year University of California (UC) campuses, there are 63M pairs (1,000*7,000) to consider, assuming[1] a catalog of 1,000 courses at the CC and 7,000 at each UC. This number can be reduced if assuming there always exists a clear department-to-department mapping between institutions, which is not always the case, and that courses are only considered for articulation within the mapped-to department. In this case, the lower bound number of course pairs to consider is 35,000 (20*35*50) assuming an average of 20 courses per department at the CC, 35 courses per department at the UC, and 50 departments at the CC articulating only to a single respective department at the UC. This number increases significantly when considering and maintaining articulation to the 23 institutions in California's State University System and articulation to the other 114 community colleges, necessary for lateral (CC-to-CC) transfer. The intractability of effectively curating an articulation database with a manual process increases exponentially when considering articulation to out-of-state or private institutions. The GAO estimates that 94% of credits are lost when transferring from a public to private institution [129].

[1]These numbers are assumed based on counts extrapolated from our dataset consisting of enrollments from one UC and one CC

In this paper, we posit that institutional big data have been an underutilized source that can be leveraged towards combating the bleak combinatorics of course-to-course articulation. We investigate the utility of these sources using the two datasets of course enrollments and course descriptions, one from a 4-year University of California campus (referred to as UC1) and one from a 2-year California Community College (referred to as CC1). Recent work has found that analysis of enrollment sequences using word2vec approaches can embed courses into a space of semantic structure [19] similar to the space words are embedded into based on their word contexts in a corpus [37]. We build on this finding to test if a translation can be learned between the course spaces of two different institutions, just as it has been learned between the word spaces of two different languages [136].

3.1.2 Datasets

3.1.2.1 UC1 dataset

Our UC1 dataset consists of 7,487 courses in 179 departments taken between 2008 and 2017 at the Berkeley campus of the University of California. We inherit pre-trained vectors for each course from the authors of prior work [19]. These continuous valued vectors are 229 dimensions in length trained from 4.8 million enrollments from 164,196 students using a skip-gram model and tested against a validation set of within-institution course credit restrictions (i.e., equivalencies) curated by the university. Details of this training is explained later in the models section of this paper. Also found in this dataset are the plain-text catalog description of each course. The average length of a UC1 course description is 325 words and there are 489 descriptions with fewer than 10 words.

3.1.2.2 CC1 dataset

Our CC1 dataset consists of 1,000 courses in 53 departments taken between 2013 and 2018 at Laney Community College of Oakland, located six miles south of UC Berkeley. This is a novel dataset for which no prior models had been trained. The average length of CC1 course descriptions is 27 and there are 62 descriptions which have less than 10 words.

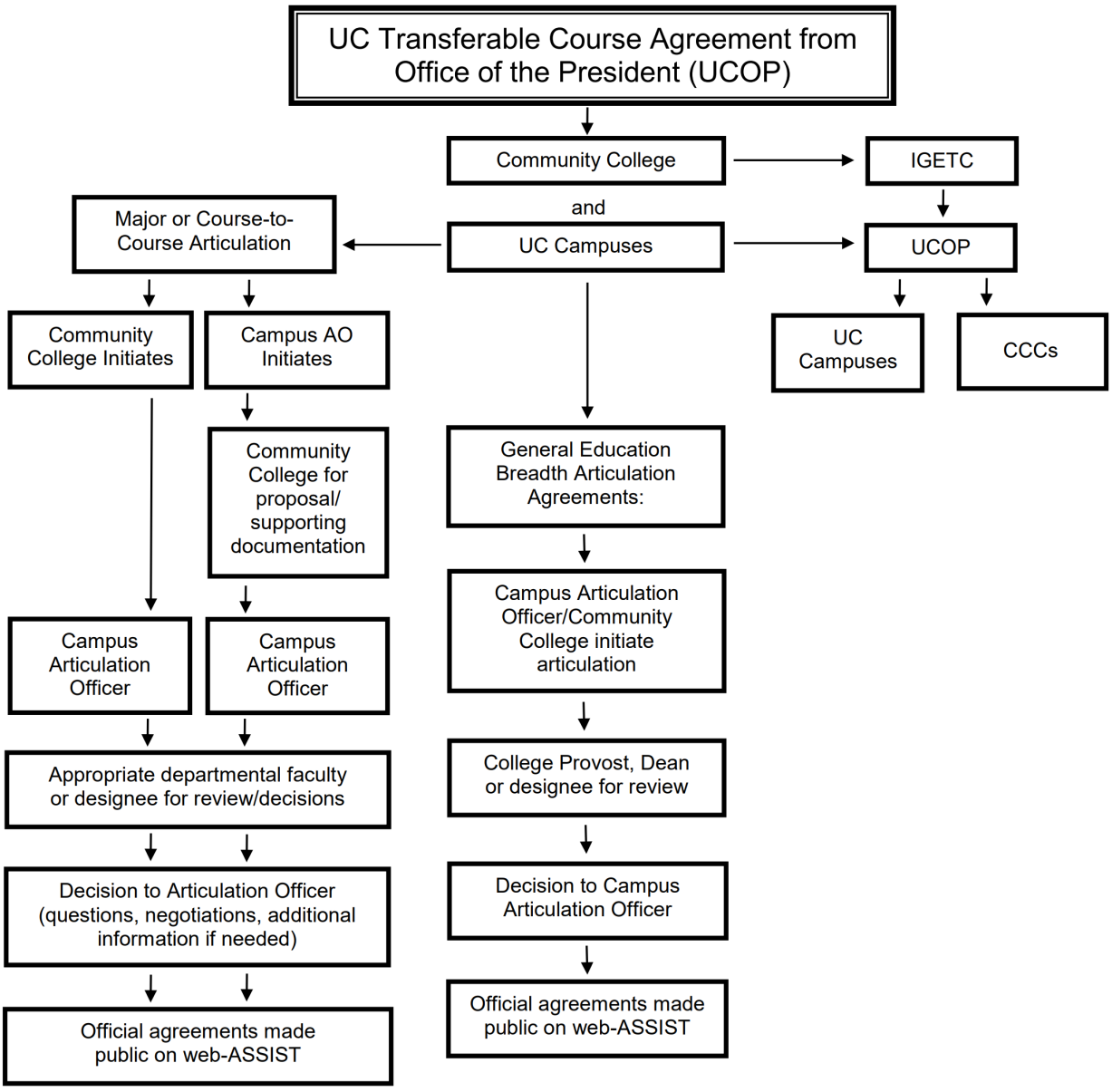


Figure 3: Diagram of the process for course articulation in the University of California system, sourced directly from the California Articulation Policies and Procedures Handbook. The process for course-to-course articulation can be seen by following the left side of the flow diagram.

Table 1: Course articulation samples from assist.org. Multiple CC1 courses denote that both must be taken to count towards the UC1 course credit.

UC1's course	*CC1's course(s)*
AFRICAM5B	AFRAM_31
ASAMST20A	ASAME_45A; ASAME_45B
ASAMST20C	NO COURSE ARTICULATED

Additionally, this dataset contains 298,174 enrollments, and their semester and year, made by 58,716 students.

3.1.2.3 Validation set

We use the existing set of course articulations between UC1 and CC1 to evaluate the predicted articulations of our models. These articulation pairs were screen scraped and manually enumerated from assist.org[2], the official information system for looking-up articulations within the California public post-secondary system. The system lists the articulations that exist between the two institutions with respect to each major offered at CC1. The total number of articulation pairs extracted was 65. Given our goal of proposing new potential articulations, we also curated a list of major satisfying courses at UC1 for which there were no respective articulated courses at CC1. There were 184 such UC1 courses. Table 1 shows samples from this course articulation dataset.

3.1.3 Models

In this section, I present several models for course representation from which to predict the similarity between courses at the two institutions. I will also describe the application of machine translation as a linear transformation from a source course vector space (i.e., UC1)

[2]These articulations were kept current up until 2017. A new system, with updated articulations, is expected within the year.

to a target course vector space (i.e., CC1). This technique is applied to our course2vec based course representations.

3.1.3.1 Collaborative-based model (course2vec)

We use an adaptation of word2vec applied to course enrollment sequences as described in prior work [19, 137]. The data are prepared by enumerating course enrollment sequences per student with the enrollment sequence consisting of course ID tokens (e.g., ECON_141) sequenced in the order in which the student took the courses. Courses taken in the same semester are serialized by randomizing their within-semester order. A skip-gram is then applied to these sequences exactly as it would be applied to sequences (or sentences) of words in a language context to produce continuous vectors for each course. Prior work has found that these vectors, learned from enrollment sequences, encode information about the topic of the course, as well as latent attributes such as its mathematical rigor and the most common major of students taking the course [137]. For the UC1 dataset, these vectors were pre-trained and inherited from the authors of that prior work. For CC1, we train course vectors, sweeping the hyper-parameters of vector size and window size and perform model selection based on the leave-one-out predictive performance on our articulation validation set, described in detail in a later section. There is a threat of overfit in this approach; however, we consider it to be minor given course2vec is an unsupervised process and a limited number of hyper-parameter combinations are used with which to generate candidates for model selection.

3.1.3.2 Content-based models

Course catalog description is the source of similarity data used by this class of models. We consider three different course representations utilizing these data; simple bag-of-words, TF-IDF, and an average of the respective word vectors of words in the description using a pre-trained word embedding.

a. BOW with term-frequency. In our simple bag-of-words (BOW) model, each course is represented as a vector of the length of the total unique words in all courses across both

UC1 and CC1. The values in a course's vector are *zeros* unless the word of the corresponding position in the vector has occurred in the description, in which case the frequency of this word in the description is used. Similarity between courses can be calculated using cosine similarity of their respective BOW. We applied a few filters to course descriptions before constructing the BOW of courses for both institutions. First, we filtered out non-words (e.g., course numbers) from the descriptions. Second, we removed the top 100 most frequent words (e.g., course, student, credit) from all descriptions. After filtering, we were left with a vocabulary of 14,316 across all descriptions.

b. TF-IDF. The simple BOW model assigns word frequencies as weights to words. However, if a word appears frequently in most of the courses, it will not help to differentiate between courses, nor help in identifying which are truly similar. We consider a *TF-IDF* (term frequency-inverse document frequency) representation to address this issue, assigning a weight to a particular term t in a course description d. As a result, instead of representing a course description as a vector of *word frequencies*, each dimension of the vector is a real-valued *TF-IDF* weight ($w_{t,d}$), calculated as following:

$$w_{t,d} = (1 + log(tf_{t,d})) \times log_{10}\frac{N}{df_t} \tag{1}$$

in which,

- $tf_{t,d}$: frequency of term t in course description d
- df_t: number of course descriptions in which term t appears
- N: number of course descriptions in the collection

c. Word2vec (DescVec). There is a large disparity in the length of descriptions between UC1 (325 words) and CC1 (27 words). Anticipating that this may introduce noise into the process of finding similar courses based on description, we decided to attempt to ameliorate this issue by representing both institution's course descriptions using a pre-trained public word embedding provided by the seminal word2vec work [37, 138]. This embedding was trained on 100 billion words from Google News with a vector size of 300 dimensions. To represent course vectors using this word embedding, we average over all the vectors of the words appearing in the course descriptions after applying the same pre-processing as

described in the above BOW sections. This approach, which we refer to as *DescVec*, is anticipated to have the added benefit of still finding similarity between courses if they use different, but synonymous words.

3.1.3.3 Model combination

The content-based and collaborative-based course representations are produced from entirely different sources of information about courses and are likely to pose their own benefits and deficits. The content-based models represent the content of the course as described by the instructor; however, the description can become out of date and a course can be described in an overly brief or generic way. In comparison, the course2vec models use student enrollment behaviors to inform the representation of a course. Because of this, they may contain important information about the course known by students (e.g., which are the courses with a reputation for being easy) but not expressed by the instructor in the description. Conversely, course2vec representations may suffer from noise in the case of courses with low enrollment (a minimum enrollment of 15 was set in the model) and also will suffer from courses that have recently changed their content considerably from historic offerings. In order to allow these two models to contribute their complementary benefits, we add a model to our evaluations which is a combination of the DescVec model and the course2vec model. The combining process is as follows:

1. *Source course vector concatenation.* Firstly, the source course vectors are transformed to the target course embedding space through a machine translating process detailed in the next section. We then concatenate the translated source course vectors with their respective DescVec course vectors (see the upper concatenation in Figure. 4).
2. *Target course vector concatenation.* No translation is needed. We only concatenate the target course vector with its respective DescVec course vector (see the lower concatenation in Figure. 4).

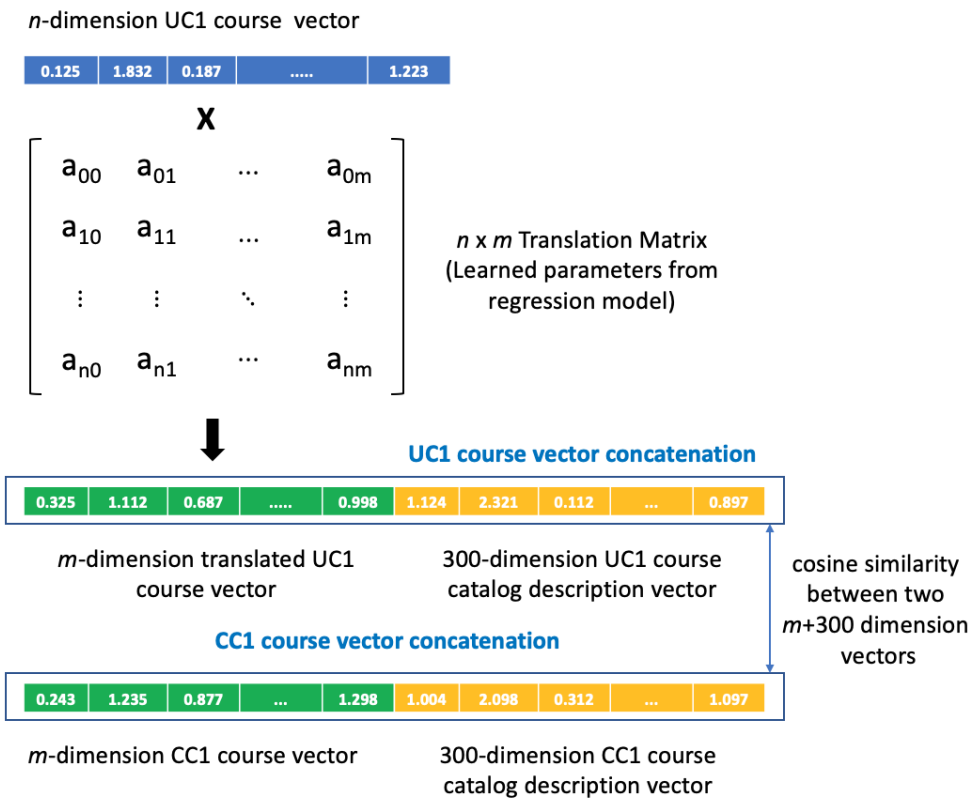


Figure 4: Process of translating a UC1 *course2vec* vector to the CC1 space and concatenating it with its *DescVec* vector for matching to a concatenated CC1 course vector via cosine similarity.

3.1.3.4 Machine Translation

For the course2vec model, UC1 course vector set and CC1 course vector set are learned separately; thus, they do not share the same coordinate frame of reference and their embeddings are subsequently different. Moreover, the dimensions of the two vector spaces are not the same. We can not directly calculate the similarity between two course vectors coming from two different spaces. However, a linear translation between skip-gram embeddings can be learned as demonstrated by Mikolov at el. [136] that showed that the same concepts (e.g., *animals*) in different languages have similar relative geometric arrangements in their embeddings. By applying the linear translation of scaling and rotation, a reasonable mapping between the two language spaces could be found based on a small set of preexisting word translation pairs. This is the key idea behind the parallel we draw to course embedding translation where we base the learning of this translation of two institution embedding spaces on a small set of preexisting course articulation pairs.

Regression-based translation. Since we anticipate that the same courses in different institutions are likely to have similar geometric arrangements in their respective institution's embeddings, the transformation from one vector embedding space to another can be expected to be linear. We perform a general linear regression with the input vector $\mathbf{s} \in R^n$ and the output vector $\mathbf{t} \in R^m$, in which n and m are the sizes of the dimensions of the source vector space and target vector space, respectively. The goal of our model is to minimize the differences between the *translated source course vectors* and *target course vectors* in the N articulation pairs. The optimization problem is described as follows:

$$min_{trans} \sum_{i=1}^{N} dist(trans(s_i), t_i) \tag{2}$$

The function ***trans*** is used to translate a course vector from the source embedding space to the target embedding space using the optimized weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$ (also called translation matrix $\boldsymbol{M}$) obtained from the regression model. The ***dist*** function is the *loss* function in the regression model. We use *cosine_proximity* and *mse* loss functions to train our models, discussed more in section "Cosine vs Euclidean". Stochastic gradient descent is used as the optimizer to fit the model to our data. After translating a course vectors from

the source embedding space to the target embedding space, the translated course vector now has the same number of dimensions as all the target course vectors, allowing it to be compared with target course vectors using metrics such as cosine similarity and Euclidean distance.

3.1.3.5 Articulation Prediction

The goal of our methodology is that, given a course c in one institute, we would like to predict an ordered list of courses in another institute that are most similar to c. With the course representations and machine translated vectors in-hand, we can compute the similarity or distance between two courses from different institutions. The course articulation process is described as following:

1. Represent all courses by one of the course representation models.
2. Translate the source course vector s through the machine translation process if the vector was produced by the course2vec or combined model. Otherwise, use the original representation of the source course vector (i.e., content-based models).
3. Compute the cosine similarity (Equation 3) or Euclidean distance (Equation 4) between the source course vector (or the translated source course vector) and all the course vectors in the target institute.
4. Rank the target institute courses based on their similarity or distance scores, and choose the top k (e.g., 10) courses for articulation recommendation.

3.1.4 Evaluation

In this section, I discuss how we validate our models, which includes choosing the hyper parameters for CC1's course2vec model, choosing between cosine similarity and Euclidean to find similar courses, and considering the difference in performance of our articulation predictions if we limit predictions to a similar department at the target institution.

Since our validation set only contained 65 labelled articulation pairs, we use a leave-one-out cross-validation. We use the metric of recall @ k to evaluate prediction performance. This means that, for each of the 65 pairs, given the UC1 course in the pair, we obtain the

top k ranked CC1 courses from the articulation prediction process explained in the previous section. The recall is calculated based on the percentage of correct CC1 courses that fall within the top k. This metric was chosen because of the anticipated scenario where we generate an articulation report to the articulation officer of CC1. This report will not show just one suggested CC1 course per unarticulated UC1 course, but rather a list of suggestions. The k in recall @ k represents the length of this hypothetical list and the recall metric represents the percentage of the 65 lists of length k that included the true articulation(s) in them.

3.1.4.1 Parameter search

The two most crucial hyper-parameters of the skip gram model are vector size v and window size w. Modification of the vector size is a way to tune the granularity with which regularities are produced in the feature space. Different languages and dataset sizes will require different vector size settings to achieve the same granularity. It is desirable for the granularity of both course vector sets to be at the same level for the feature mapping to be effective. We therefore conduct a minimal hyper parameter search of the CC1 course2vec model. We start with the pre-trained course vectors from UC1. Then, we sweep a small range of vector sizes and window sizes for CC1's course2vec by optimizing the leave-one-out recall performance described in the above section. We chose recall @ 5 as the k used for optimization as this was an ad-hoc estimate for a reasonable length list of courses for an articulation officer to consider. This process of hyper-parameter tuning went as follow:

1. For each of the 27 parameter sets, we run course2vec 20 times to learn different CC1 course vectors based on different random model initializations
2. Obtain the average recall @ 5 from leave-one-out cross validation described above for each CC1 course vector set
3. Average over the recalls @ 5 of all the 20 course vector sets for particular parameter sets
4. Select the parameter set with the best average recall performance

The result from Figure 5 shows that, within the 27 parameter sets, the vector size 20 and window size 5 achieved the best perform w.r.t recall @ 5. Therefore, we chose these values

to train the final CC1 course vector set.

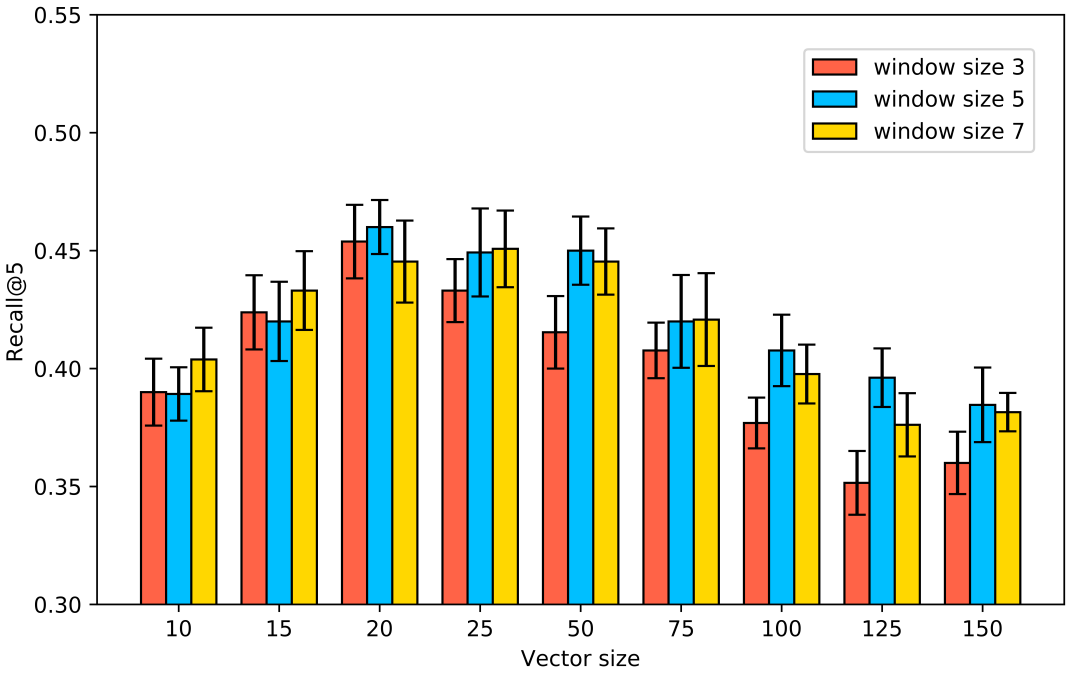


Figure 5: Recall performance @ 5 with different sets of course2vec vector sizes and window sizes for training CC1 vectors. The error bars represent 95% confidence intervals, obtained by running each model 20 times.

3.1.4.2 Cosine vs Euclidean

Given vector-space course representations, if vectors come from the same original space, it is effective to directly calculate their distances using cosine or Euclidean distances. On the other hand, vectors coming from different spaces need to be mapped to the same space by the proposed method explained in Section 3.1.3.4. After the transformation, we can calculate their distances.

- *Cosine similarity:* measure the similarity between two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ by computing the cosine of the angle between them.

$$cosine\_similarity(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}.\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}} \tag{3}$$

- *Euclidean distance:* measure the straight-line distance between two points in Euclidean space.

$$Euclidean\_distance(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Depending on which distance metric is used for evaluation, matching to an appropriate loss function used to optimize the problem defined in Equation 2 may be called for. If cosine similarity is used to evaluate, we can use cosine_proximity as the loss function, and mean squared error (mse) as the loss function in the case of Euclidean as the evaluation metric.

3.1.4.3 Department filtering

It is intuitive to think that course articulation pairs should be in equivalent departments across colleges (e.g., a course offered by Mathematics department at UC1 should be articulated to a course offered by Mathematics department at CC1). We therefore also compare the performance of the best model to its department filtering version. However, among the 65 articulation pairs, there are 2 pairs that come from departments that were not mapped to one another in the department mapping conducted by the authors. These were STAT2 to MATH13 and NUSCTX10 (Nutritional Sciences and Toxicology) to BIOL28 (Biology). In order to have a fair comparison for *with* and *without* department filtering data, we excluded these two pairs, leaving us with 63 articulation pairs for that evaluation.

3.1.5 Results

In this section, I present the articulation prediction performance of the different models and present visualizations of the course embeddings for intuition. For the course2vec model and combined model, as we obtained from our development experiments, the performances of models trained with *mse* loss function were worse than the ones trained with *cosine* loss

function. Therefore, I only report the results for the models trained with *cosine* loss functions (see Figure 6).

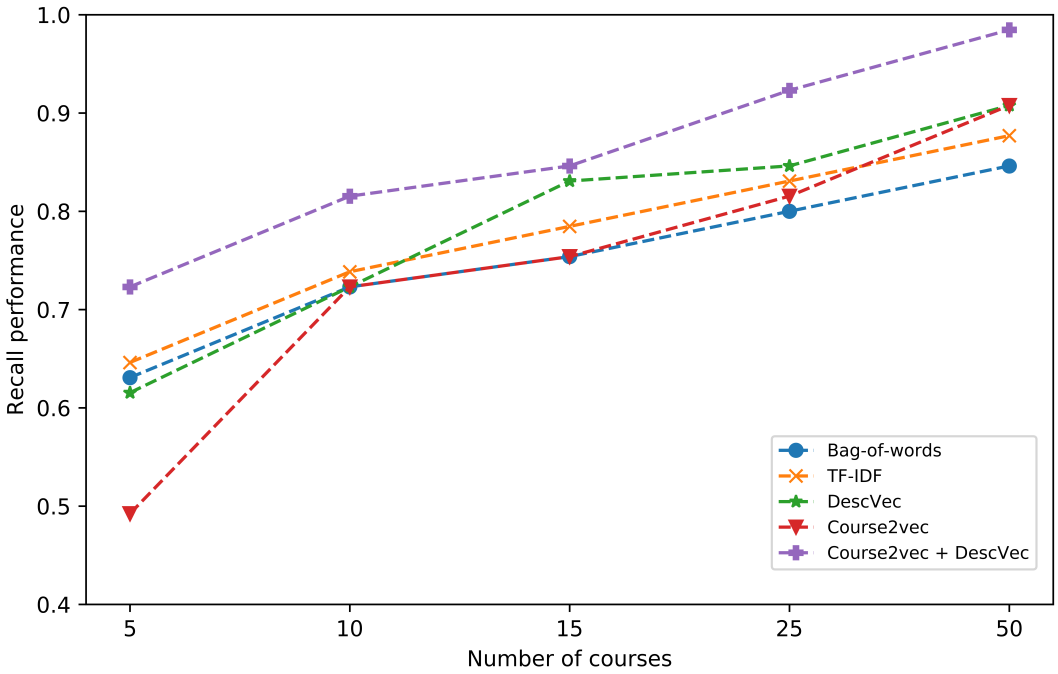


Figure 6: Recall comparison of the different models trained with *cosine_proximity* loss function @ k.

In addition to the recall performances @ k, I also report the rank of the true articulated course in our prediction results. The median and mean rank across the 65 articulation predictions is reported, as well as the standard deviation of ranks (see Table 2).

Observations:

- Although slim, the BOW model with TF-IDF shows consistent improvement over the term-frequency BOW model across values of k.
- Among the content-based models, the DescVec model performs best, overall.
- The course2vec model performs substantially worse than the content-based models on recall @ 5 but then matches their performance for all other values of k. It also can be observed from Table 2 that, while having a higher median rank, the course2vec model's

Table 2: Course articulation ranking validation from the different course representations.

Course Representation	*Median Rank*	*Mean Rank*	*Std of Rank*
Bag of words	3.0	59.12	173.28
TF-IDF	3.0	57.01	177.65
DescVec	3.0	21.06	57.94
course2vec	6.0	17.74	33.65
course2vec+DescVec	**2.0**	**7.94**	**15.73**

mean and std are lower than the content-based models, suggesting that it has fewer poor performing outliers.

- The combined model (course2vec + DescVec), which leveraged the strength of both the content-based and collaborative-based models, shows the best performance across all values of k and among all the rank metrics.

Figure 7 shows the difference in performance between the combined model with and without department filtering. The performance of the model with department filtering brings recall @ 5 up above 80%. An interpretation of this result is that if the model were to produce a set of five CC1 course articulation suggestions for each one of ten chosen UC1 courses, eight of those sets of ten suggestions can be expected to contain an appropriate articulation course.

Visual inspection of course2vec models. We visually inspect the CC1 and UC1 course2vec models to investigate if similar geometric regularities can be seen as they have been in visualization of language translation models [136]. We use PCA to reduce the course2vec vectors to 2-dimensions, then zoom into the the Computer Science departments of each visualization to compare the relative positions of courses with articulations between UC1 and CC1. As we can see from Figure 8, the 2D course vectors obtained from the skip-gram models show a similar, but not perfectly so, geometric arrangements of articulated courses. Computer Science was picked because courses within that department performed

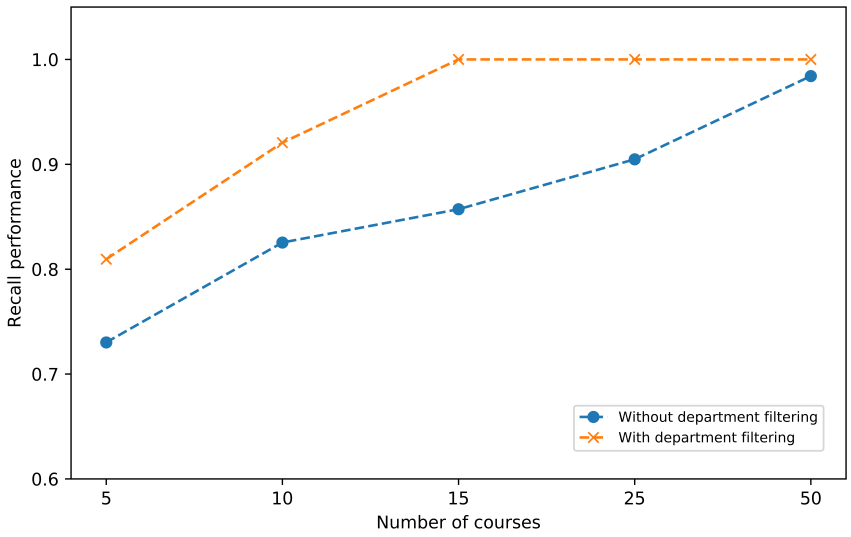


Figure 7: Recall comparison of the *CourseVec + DescVec* model with and without department filtering.

well on the articulation task and produced a PCA visualization that underscores why the course2vec representations learned from course enrollments alone can be effective.

Moreover, we also inspect the course representations for all courses at each institution, this time using t-Stochastic Neighborhood Embedding [139] to reduce the course2vec+DescVec representations to 2-dimensions. The t-SNE algorithm is chosen in this case because it is generally better at retaining global embedding structure than PCA. This visualization (Figure 9) reveals a few regularities in the relative department-level positions of the two institutions as well as a tight grouping of courses by department, also observed in [137]. The deparments of *Chemistry*, *Engineering*, *Civ Eng*, *Architecture*, *African American Studies*, and *American Studies* can be seen as appearing in clockwise order in both institutions, underscoring why the representations learned from course enrollments by course2vec alone, rival the information contained in course descriptions.

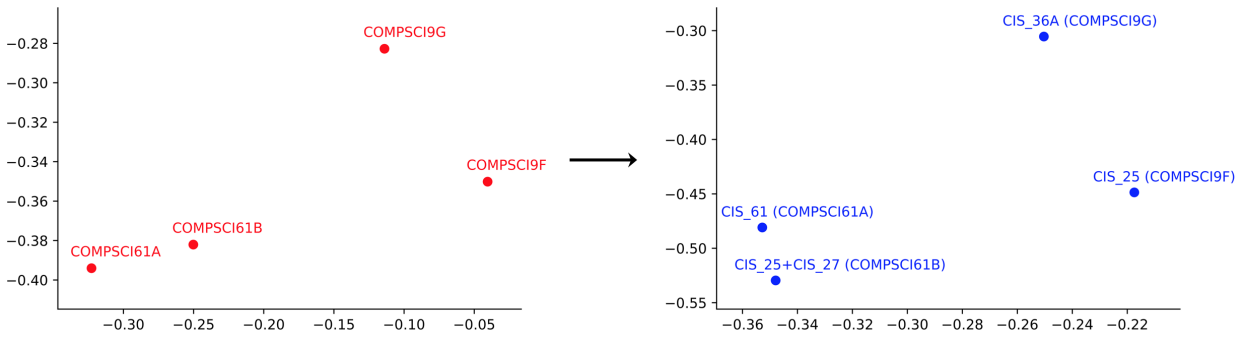


Figure 8: Distributed vector representations of Computer Science courses in UC1 and CC1. The four course vectors are reduced to two dimensions using PCA in each of the institutions. UC1 includes *Structure and Interpretation of Computer Programs* (COMPSCI61A), *Data Structures* (COMPSCI61B), *C++ for Programmers* (COMPSCI9F) and *JAVA for Programmers* (COMPSCI9G). CC1 includes *Structure and Interpretation of Computer Programs* (CIS_61), the combination of *Object Oriented Programming Using C++* (CIS_25) and *Data Structures and Algorithms* (CIS_27), *Object Oriented Programming Using C++* (CIS_25) and *Java Programming Language I* (CIS_36A).

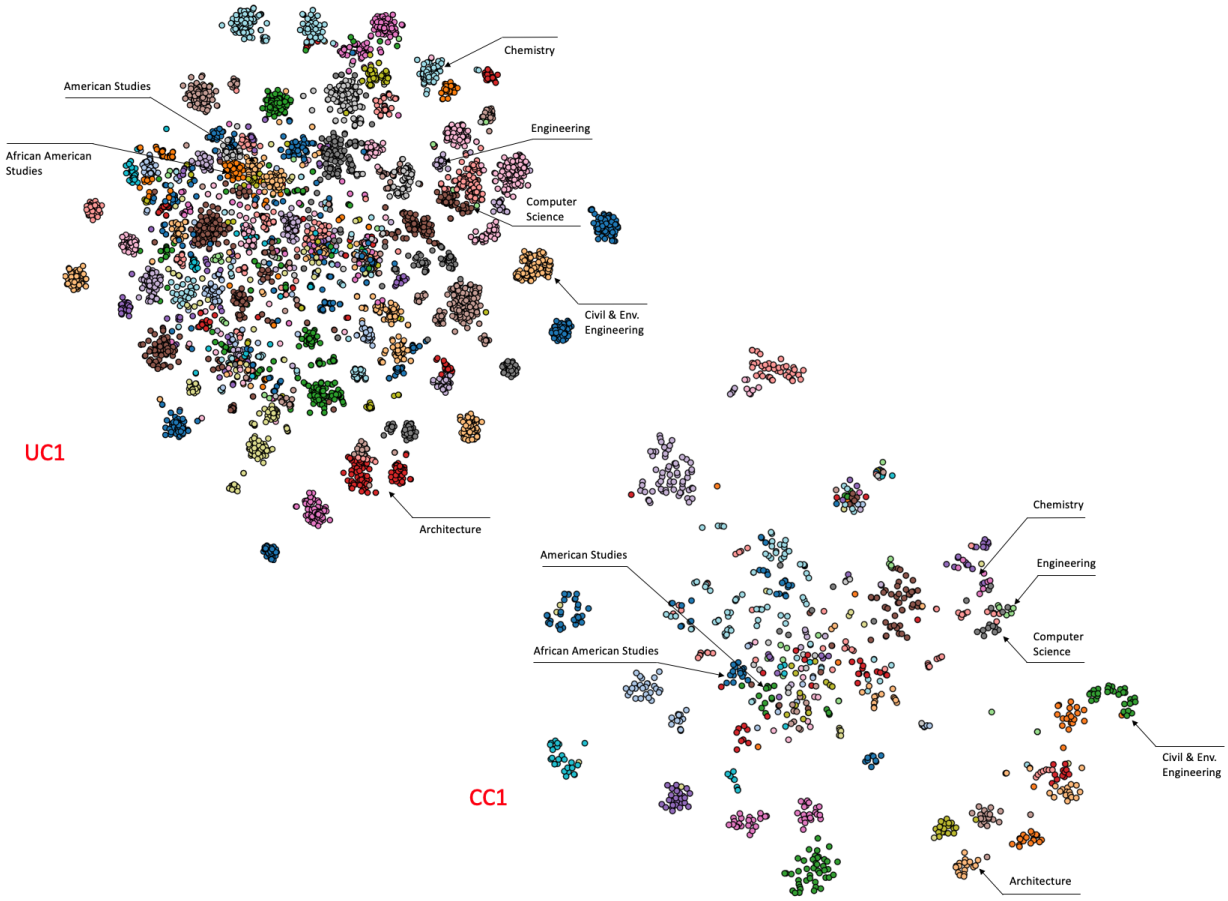


Figure 9: t-SNE scatter plots of courses obtained from the *course2vec+DescVec* models. The color of the points represent the departments and the text annotations represent the names of the departments which have sufficient courses and direct mappings between UC1 and CC1.

3.1.6 Discussion

We found that a simple word2vec approach to articulation, *DescVec*, performed equal to or better than the experimental course2vec machine translation model. However, the experimental model was shown to contain novel useful information in addition to what was found in the course description based *DescVec* model. This was made evident by the performance of the concatenation of the course2vec model vectors with the *DescVec* vectors which performed meaningfully better than any other model in our recall @ k metric for all values of k. It also performed between 30% and 50% better than the second best model in median, mean, and std. rank metrics and was therefore used to produce the articulation report.

The primary barrier to adoption of this methodology is sharing of enrollment data. It is a challenge to successfully approach institutions with a request to share this type of data. In order for this endeavor to be successful, it may take the support of existing centralized data repositories, such as the assist.org system, operated by the UC Office of the President (UCOP), or national data collectors such as the National Student Clearinghouse or Department of Ed. A secondary barrier to adoption is the degree to which this data-assistive method is accepted into the socio-technical system of course articulation. If the method is seen as a threat to articulation officers' jobs, as AI is increasingly seen as to many jobs, it will be difficult to integrate with the articulation officer as the point of contact. Direct-to-student suggestion of articulation candidates for them to petition may be more fruitful in this case.

While our study focused on these methods used to identify new articulations, out of date articulations may be just as important to identify. Transfer students receiving credit for courses that do not well enough prepare them for the material that will be encountered upon transfer are also harmful to student success. The methods described could just as well be used to identify the existing articulations with the lowest articulation scores, for re-consideration.

3.1.7 Limitations and Future Work

A limiting factor to the potential success of a report generated for UC1 to CC1 articulation is that CC1 is what is known as a "common feeder school" to UC1, meaning that it is a top source of transfer students for UC1. This means that articulations between the two institutions may be near saturation levels. A limitation of the course2vec approach is that a course must have an enrollment history in order to receive an embedding. This would rule out courses which are being offered for the first time as candidates for articulation to or from. In this case, a content-based model would need to be defaulted to. The primary limitation of the machine translation method is that it relies on existing articulations to learn the translation, ruling out the method for application to institution pairs for which no articulations exist, which are by definition the most in need of articulation. Again, the content-based methods could be applied in these cases and promising unsupervised language translation methods [140] may become candidates for overcoming a lack of existing articulations to train on. Faculty currently consider factors such as the difficulty of the source course and its syllabus when deciding to accept a proposed articulation. Future enhancements to the content-based models could include parsed syllabus information and data from the LMS. Were it available, test questions and their grading or even graded student answers might further provide a means for automatically scoring the match in difficulty between courses at different institutions. These data are available in MOOC datasets, and thus the emerging articulation context of MOOC micro-credentials[3] and their mapping to accredited degree programs is already halfway to a tenable context for this approach.

3.2 Orienting Students to Course Recommendations Using Three Types of Explanation

An emerging challenge in course recommender systems is explaining to students why they have been recommended particular courses. In the context of a university, it can be valuable

[3]https://www.edx.org/micromasters

for a recommender system to introduce students to courses they may have not otherwise taken but which are still relevant to them. This is a collaboration with the CAHL lab at the University of California, Berkeley. I will summarize the methods and study results in the next sections, more details can be found in [38].

3.2.1 Methods

Our course recommendation scenario is based on showing courses relevant to a "favorite" course chosen by a student from their course history. We evaluated two different course recommendation models in this study: one using BOW course representations, and the other using a concatenation of BOW and multi-factor course2vec with instructor and department features (referred to as the *best analogy* model, proposed by Pardos and Jiang [21]). Both methods of representing courses are described in detail in [38], and both use cosine similarity to the *favorite* course to find recommendations. Serendipitous recommendations require user-perceived unexpectedness. In both models, we attempt to generate unexpectedness by diversifying recommendations, showing a maximum of one course per department in the results and excluding the department of the favorite course.

For producing explanations, our hypothesis is that by adding more information about a course, some of it personalized, including the reasons for its recommendation, students will become more familiar with the course, potentially increasing its success. We refer to the favorite course chosen by the student as an *anchor* course and the courses recommended by the system as *target* courses. The three proposed methods for explaining course suggestions are described (details can be found in [38]) as follows:

Inferred Keywords: A course may be recommended based on latent features not described in its official catalog description. To help students understand the recommendation, these latent features can be added as part of an explanation. A method using course2vec vectors is applied to extract keywords from enrollments [141]. Once the model is trained, a softmax probability distribution is used to find high probability words predicted by course2vec that are not in the course description. These words are treated as the course's inferred keywords.

Anchored Keywords: To help students understand recommendations, explaining the relation between the recommended target courses and the anchor course is important. This is achieved by showing overlapping keywords between them. Unigrams are extracted from both course descriptions, with the course title and inferred keywords of the target course included. After preprocessing and filtering, an intersection operation is applied to find common keywords. These intersected keywords are then sorted by probability for explanation.

$$Anchored\_Keywords = (T_t \cup\ T_d \cup\ T_i) \cap\ A_d \tag{5}$$

In which, T_t is the keywords in the target course title, T_d is the keywords in the target course description, T_i is the inferred keywords of the target course, and A_d is the keywords in the anchor course description.

Taken Keywords: The final explanation approach relates target course recommendations to keywords from courses the student has taken in the past. By retrieving the student's enrollment history and producing keywords for those courses. The taken course keywords and target course keywords are preprocessed, intersected, and sorted by probability for explanation.

$$Taken\_Keywords = (T_t \cup\ T_d \cup\ T_i) \cap\ B_d \tag{6}$$

Where T_t, T_d, and T_i are the same as the previous equation. And B_d is the keywords in the description of the taken courses.

3.2.2 User Study

3.2.2.1 Study Design

The goals of the study are to understand and improve the *unexpectedness*, *successfulness*, *serendipity*, and *novelty* of serendipitous course recommender systems in higher education using different recommendation models and explanation strategies proposed in the previous sections. To evaluate the effects of these factors, we designed a user study to collect students' ratings of recommendations along various measures. Once the study began, each student was randomly assigned to one of the ten experimental conditions (see Table 3).

Participants were asked to select a favorite course they had taken before as input for the recommender. After choosing the favorite course, the system displayed five sorted courses from different departments. The display information of the recommended courses included the course catalog description, and, depending on the condition the participant was assigned to, the system would show them *inferred* keywords of the recommended courses, keywords in common with the selected favorite course (*anchored* keywords), or keywords in common with the courses they took in previous semesters (*taken* keywords), or all of them together. After examining the suggested courses based on their favorite course, students were asked to rate the recommendations based on their agreement with the following statements (used in [21]) on a 5-point Likert scale, from 1 (Strongly Disagree) to 5 (Strongly Agree):

1. *This course was unexpected.*
2. *I am interested in taking this course.*
3. *I did not know about this course before.*

These questionnaires help to measure different dimensions of recommendations. We consider the first statement as *unexpectedness*, and the second statement as *successfulness*, where students express their interest in taking the suggested course. *Novelty* is measured by the third statement. Shani and Gunawardana [142] defines *serendipity* as the combination of *unexpectedness* and *successfulness*. In our case, we use the mean of *unexpectedness* and *successfulness* as our measure of *serendipity*.

3.2.2.2 Study Results

We received a total of 329 ratings from 67 students consisting of freshmen (50), sophomores (10), juniors (1), and seniors (6). The participants were fully engaged in the study and spent a significant amount of time examining the recommendations and explanations.

The rating results of the ten conditions are shown in Table 3. Each row is the average rating results of all students assigned to a specific condition. For example, for the BOW model without any explanation (*1A*), 3.571 and 2.771 are the average values of 35 students' course ratings on the first two questions (i.e., *unexpectedness* and *successfulness*. The *serendipity* rating is 3.171, average ratings of *unexpectedness* and *successfulness*. Finally,

the average rating of *novelty* is 3.714 for condition *1A*. Among all ten conditions, the *4A* (BOW + showing keyword overlap with courses taken before) achieved the highest ratings for *unexpectedness* (3.700) and *novelty* (3.900). Condition *3B* scored the highest rating of 3.046 for *successfulness*, while *serendipity* received the best rating of 3.300 from condition *4B*.

Table 3: Average student ratings of individual course recommendations from the user study broken out by model used to generate course recommendations, method used to generate the explanations, and rating construct.

Condition	Model		Explaining Keywords			Average Student Ratings				
	BOW	Analogy	Inferred	Anchored	Taken	Unexpectedness	Successfulness	Serendipity	Novelty	Ratings
1A	●	○	○	○	○	3.571	2.771	3.171	3.714	35
2A	●	○	●	○	○	3.096	2.741	2.919	3.354	31
3A	●	○	○	●	○	3.114	2.771	2.943	3.800	35
4A	●	○	○	○	●	**3.700**	2.666	3.183	**3.900**	30
5A	●	○	●	●	●	3.025	2.975	3.000	3.575	40
1B	○	●	○	○	○	3.200	2.320	2.760	3.080	25
2B	○	●	●	○	○	3.244	2.311	2.688	2.644	40
3B	○	●	○	●	○	2.767	**3.046**	2.907	3.279	43
4B	○	●	○	○	●	3.650	2.950	**3.300**	3.450	20
5B	○	●	●	●	●	3.500	2.766	3.133	3.533	30

As we can see from Fig. 10 (A), *unexpectedness*, *successfulness* and *novelty* outcomes received average ratings from 2.75 and above (out of 5), where *Novelty* was the highest (3.400). This suggests that our design for recommendations and explanations has reasonable effects on these outcomes. There was no significant difference between BOW and Analogy models' results except for on the *Novelty* measure (see Table 4). The BOW model had a higher impact on the novelty of recommendations in this study. It has much higher mean and median ratings compared to those from the Analogy model (Fig. 10 (B)). In addition, we ran an analysis to test differences between without-explanation and with-explanation (where all the three types of explanations are considered) on the outcomes and found no significant differences, which points out that different types of explanations affect users in different ways.

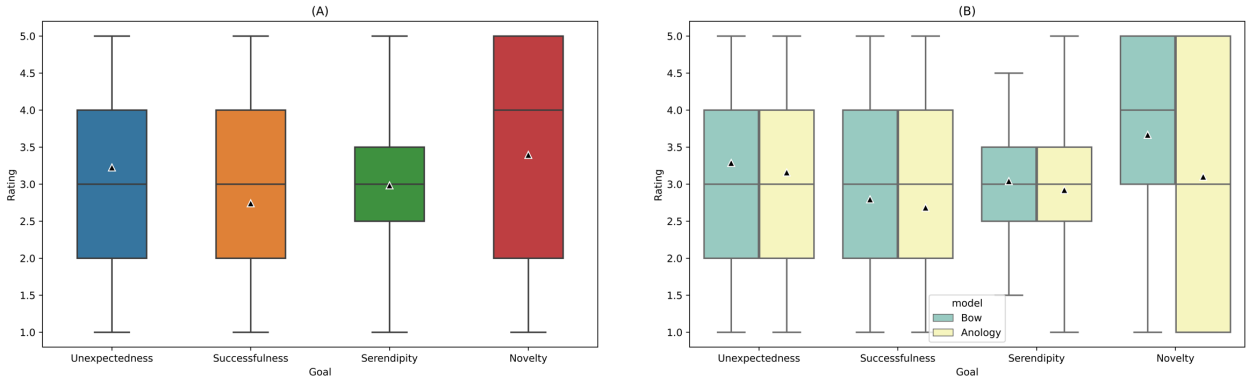


Figure 10: Student ratings for the four outcomes are presented as box plots. Middle lines represent the median ratings, while triangles represent the mean ratings. (A) presents the overall rating from all the students, and (B) presents the ratings separated by recommendation model types (BOW and Analogy).

Table 4: The impacts of the recommendation models and explanation strategies according to OLS regression.

	Unexpectedness		**Successfulness**	
Variable	Coefficient	P-value	Coefficient	P-value
Analogy Model	-0.0759	0.595	-0.1039	0.431
Inferred Keyword	-0.1982	0.180	-0.1293	0.327
Anchored Keyword	-0.3501	0.013 (*)	0.2873	0.037 (*)
Taken Keyword	0.4161	0.006 (**)	0.1182	0.442
	Serendipity		**Novelty**	
Variable	Coefficient	P-value	Coefficient	P-value
Analogy Model	-0.0899	0.292	-0.5240	0.001 (**)
Inferred Keyword	-0.1637	0.063	-0.4454	0.011 (*)
Anchored Keyword	-0.0314	0.722	0.2275	0.161
Taken Keyword	0.2671	0.005 (**)	0.3701	0.039 (*)

* p-value<0.05, ** p-value<0.01

To better understand the effects of the recommendation models and different explanation methods (independent variables) on the measures; *unexpectedness*, *successfulness*, *serendipity* and *novelty* of recommended items, we ran robust Ordinary Least Squares (OLS) regressions for each of the measures (dependent variables). The number of observations (N) was 329. For each observation, the binary independent variables (i.e., *model*, *inferred keywords*, *anchored keywords* and *taken keywords*) got value *0* or *1* depending on what condition they belonged to; solid black circles (in Table 3) indicate which features were present. The results in Table 4 explain that:

- The Analogy model surprisingly had a negative effect on the measures when compared to the base (BOW) model in the regression, which means that the BOW model achieved better results in this study even though most of them are not statistically significant. We also see the differences between the mean and median ratings of the two models in Figure 10 (B). For the *novelty* measure, the Analogy model had the largest magnitude impact (among the variables) with a coefficient of -0.524 and statistical significance with p-value of 0.001.
- Inferred Keyword method also negatively affected the measures. The difference was statistically significant for *novelty* with the coefficient -0.4454 and with the p-value 0.011. It also had a moderately negative impact on *serendipity* with the p-value 0.063. The explanations from this method were worse than those from Anchored Keyword and Taken Keyphrase methods in helping students perceive the novelty of recommended items.
- Anchored Keyword method had a significantly positive impact on *successfulness* with a coefficient of 0.2873 and p-value of 0.037 but a negative impact on *unexpectedness* with a coefficient of -0.3501 and p-value 0.013. This result agrees with the known trade-off between the successfulness and unexpectedness factors of recommendations. The Anchored keywords helped to explain the relationships (i.e., shared topics) between a student' favorite course and a recommended course. When shown more connections between a favorite course and a recommended one, students were less surprised about the recommendation.
- Interestingly, the Taken Keyword method statistically significantly positively affected three out of the four measures, including the main serendipity measure. The coefficients

were 0.4161, 0.2671, and 0.3701 with p-values of 0.008, 0.004 and 0.037 for *unexpectedness*, *serendipity* and *novelty*, respectively. Showing taken keywords from the courses students already took improved the unexpectedness of recommendations. The Taken Keyword strategy is the only one that had a significant impact on *serendipity*, which comes from the positive impacts of the method on *unexpectedness* and *successfulness*, two measures difficult to simultaneously increase.

3.3 Discussion and Conclusion

We consider this work as the first attempt to understand and improve unexpectedness, successfulness, serendipity, and novelty in higher education course recommender systems using explanations. Three novel explanation strategies were introduced, including two personalized ones using students' course history. These explanations were integrated into course recommendations to help students understand the recommendations and improve their perceptions of serendipity.

We conducted a user study with 67 students, where students were assigned to a random explanation condition and asked to rate recommendations. We found that explaining recommendations using keywords from courses the student had taken increased all measures and significantly improved serendipity. The only strategy that significantly improved successfulness was based on the keywords in the student's chosen favorite course, although it also decreased unexpectedness. The largest negative effect was observed in the analogy model on the novelty measure. Further work is needed to understand the aspects of the models responsible for differences in perceptions.

The study suggests that appealing to students' prior knowledge can be effective in generating personalized explanations. The magnitude of the results could be further improved by using keyphrases instead of unigrams for the explanations since unigrams may not have sufficient ability to communicate meaning encapsulated in course descriptions. It may not be easy for students to interpret the meaning of single words, especially technical, without context. The complete keyphrases could better communicate the underlying semantics and

have been already used successfully to explain recommendations [31, 32, 33] and have shown promise in improving user comprehension [143, 144] over unigrams [34, 35].

4.0 AUTOMATIC CONCEPT EXTRACTION FOR COURSE DESCRIPTION WITH DEEP LEARNING

Knowledge and skills have consistently been recognized as critical components in many educational AI systems over the past few decades. Not only do skills assist recommender systems in making informed decisions, but displaying skills is also one of the most intuitive ways to explain the content of documents. The absence of a standardized knowledge base describing more granular concepts and skills in higher education and the labor market underscores the urgent need to construct a knowledge base capable of standardizing and enhancing our understanding of how educational foundations influence future careers and the skills of individuals. As a step toward the goal of automated educational ontology construction, this chapter proposes a deep learning methodology to extract fine-grained concepts (referred to as 'skills' discussed in Chapter 1, Section 1.1) presented directly in educational documents, specifically course catalog descriptions, without the need for manually labeled data for training. This enables the investigation of the impact of skill-based explanations on course recommendations in higher education in the subsequent chapters.

4.1 Introduction

Transfer learning is developed to address the challenge of using representations that are first pre-trained on large quantities of unannotated data and then further adapted to guide other downstream tasks. A recent trend in transfer learning is to utilize self-supervised learning on extensive general datasets with deep neural networks to obtain a general-purpose pre-trained model that captures the intrinsic structure of the data. Deep learning not only achieves great performance but also has the advantages of transfer learning and learning feature representations from scratch (e.g., word embeddings). This pre-trained model can be fine-tuned to a specific task with a particular dataset, which has been demonstrated to be highly effective in natural language processing (NLP) [145, 146] and computer vision [147].

This approach is also effective for fine-tuning deep learning models with limited or weak labels on downstream tasks [148].

Deep neural network models for sequence tagging with Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN) and Transformers which brought some breakthroughs in NLP systems with pre-trained distributed vector representations for characters and words, and neural language models. These deep neural models and architectures are able to successfully perform sequence tagging tasks based on inferred features and semantic and syntactic information of languages from pre-trained embeddings and language models. They are robust and easy to adapt to different sequence labeling tasks, avoiding the burden of manually engineering features for specific domains and applications. These models have been successfully applied to NER tasks and have been recently shared with keyphrase extraction tasks [56, 149].

To approach concept extraction from course descriptions as a sequence labeling task with deep learning, it can be formally stated the same as a NER task. Let $d = \{x_1, x_2, ..., x_n\}$ be an input sequence of n words, where x_t present the t^{th} token. The task is to infer their hidden class labels $Y = \{k_B, k_I, k_O\}$, where k_B means x_t is the beginning of a keyphrase, k_I means x_t is the inside of a keyphrase, and k_O means x_t is not a part to a keyphrase.

Domain adaptation is a machine learning technique that aims to improve the performance of a model trained on a source domain, when the distribution of the target domain is different, yet related. In general, domain adaptation utilizes labeled data from one or more source domains to perform similar tasks in a target domain [150, 151]. The effectiveness of domain adaptation depends on the degree of relatedness between the source and target domains. In deep domain adaptation, the goal is to make use of the highly transferable features learned by DNNs to improve the adaptation of the model to the target domain. This involves fine-tuning the model on the target domain, while preserving the transferable representations learned by the DNN on the source domain. This allows the model to generalize better to the target domain and improve its performance.

The fine-tuning process in deep domain adaptation typically involves adapting the parameters of the model's later layers while keeping the earlier layers fixed. This is because the lower layers of DNNs usually capture more transferable features that are relevant across dif-

ferent domains, while higher layers capture domain-specific features. Therefore, fine-tuning only the later layers allows the model to adapt to the target domain while maintaining the transferable representations learned from the source domain. Depending on the data availability from the target domain, domain adaptation can be applied with supervised, semi-supervised, and unsupervised approaches [151].

In this study, I train concept extraction models for different types of documents such as Wikipedia articles, abstracts of scientific articles, and textbook sections. Then, these models are directly adapted to extract concepts (representing 'skills' discussed in Chapter 1, Section 1.1) in another type of document that is course description. The assumption is that the source distributions of the input "domains" are similar enough to the distribution of the target "domain" so that we can utilize the models trained on the source documents directly to the target course description without necessitating specific domain adaption techniques.

In this chapter, I will first present the two most popular deep neural architectures and how they can be used for concept tagging tasks. I then describe in detail the data, implementation and model performances on a test set and expert judgement. Finally, I conclude the chapter with a discussion.

4.2 Deep Neural Architectures for Concept Extraction

4.2.1 Bi-LSTM-CRF

Bidirectional LSTM-CRF (Bi-LSTM-CRF) models, one of the latest techniques for sequence tagging, are first presented in [56]. Different modified versions of the models have been proposed in recent years, models with static word embeddings [152], models with contextual embeddings [153], models with additional character embeddings [153, 154, 155], models adding language model (LM) embeddings [156], and models with task-aware neural language model [155]. Some of these models have been adapted to keyphrase extraction mainly for scientific articles; [57] uses the Bi-LSTM-CRF model with fixed word embeddings and character embeddings, and [58] uses the Bi-LSTM-CRF model with contextual embeddings.

These studies show the benefits and improvement of keyphrase extraction compared to state-of-the-art unsupervised and supervised models.

The standard Bi-LSTM-CRF model for keyphrase extraction consists of three main components [155] (see Fig. 11).

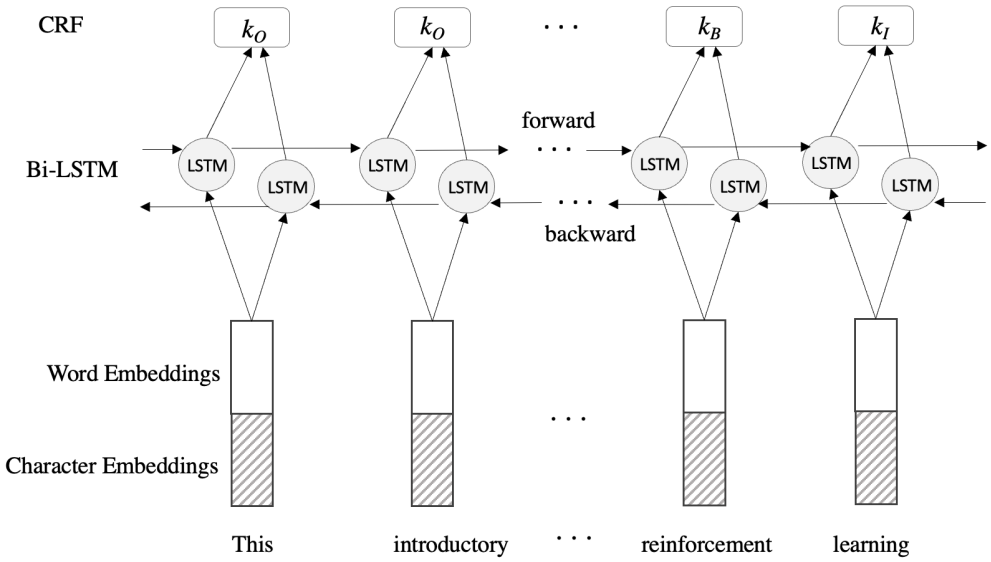


Figure 11: Bi-LSTM-CRF architecture adaption for concept extraction.

Embedding Layer: word and character-level embeddings are trained purely on unannotated sequence data from a text corpus. While word embeddings capture syntactic and semantic regularities in language, character embeddings provide additional information about the underlying style and structure of word; for example, it can mimic Shakespeare's writing and generate sentences of similar styles, or even master the grammar of programming languages (e.g., XML, LATEX, and C) and generate syntactically correct codes. Both the embeddings help to improve many NLP tasks including sequence tagging. The one-hot vector of an input word w_t is mapped to a fixed-size dense vector (the concatenation of character and word embeddings) x_t in this layer.

- *Word embeddings*: different pre-trained word embeddings could be applied in this layer: fixed word embeddings such as Glove, FastText, Word2Vec, and contextual embeddings such as ELMo, BERT.

- *Character embeddings*: in addition to regular word embeddings, representations of words can be constructed using representations of the characters they are composed. This could be helpful, especially for languages in which the spellings of words are sensitive. Character embeddings can be trained with Bi-LSTM [153, 155] or CNN [154]

Bi-LSTM Layer: LSTMs [157] are recurrent neural networks that deal with vanishing and exploding gradient problems with the use of gated architectures. Bidirectional LSTMs (Bi-LSTM) are a generalization of LSTMs that capture long-distance dependencies between words in both directions. For the keyphrase extraction task, I have access to both past and future input features for a given time, I can thus utilize a bidirectional LSTM network (Fig. 11). In doing so, I can efficiently make use of past features (via forward states) and future features (via backward states) for a specific time frame, which is more robust and efficient compared to manually engineered features that need to define a specific context window in traditional classification (e.g., SVM) or sequence tagging (e.g., CRF).

The concatenation of character and word embeddings x_t is the input for this layer, thus d is represented as a sequence of vectors $x = \{x_1, x_2, ..., x_n\}$. The corresponding class labels are $y = \{y_1, y_2, ..., y_n\}$, where $y_i \in Y$. A Bi-LSTM is used to encode sequential relations between the word representations. A LSTM unit consists of four main components: input gate (i_t), forget gate (f_t), memory cell (c_t), and output gate (o_t), which are defined as below:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t &= f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{7}$$

In which, σ denotes the sigmoid function, $tanh$ is the hyperbolic tangent function, and $\odot$ is an element-wise dot product. W and b are model parameters that are estimated during training, and h_t is the hidden state.

The input vector x_t goes through LSTM units in both directions, creating two hidden state vectors: $(\overrightarrow{h_t})$ and $(\overleftarrow{h_t})$ capturing in formation from words before and after x_t, respectively. The concatenation of these two vectors represents the semantics and dependencies of x_t in the context of the input text.

$$\overleftrightarrow{h_t} = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{8}$$

An affine transformation to map the output from the Bi-LSTM to the class space in the CRF layer: $f_t = W_a \overleftrightarrow{h_t}$, where W_a is a matrix of size $|Y| \times 2l$ and $l = |\overrightarrow{h_t}|$. The score outputs from the Bi-LSTM $f = \{f_1, f_2, ..., f_n\}$ serve as input to a CRF layer.

CRF Networks: There are two different ways to make use of neighbor tag information in predicting current tags. The first is to predict a distribution of tags for each time step and then use beam-like decoding to find optimal tag sequences, maximum entropy classifier [158] and Maximum entropy Markov models (MEMMs) [159]. The second approach is to focus on sentence level instead of individual positions, thus leading to Conditional Random Fields (CRF) models [160] (Fig. 11). CRFs are discriminative probabilistic models which has been successfully used in many sequence labeling tasks. [56] combined CRF with deep learning models, which has been shown to improve the performance of many sequence labeling tasks. In a CRF layer, the score of an output label sequence is:

$$s(f, y) = \sum_{t}^{n} \tau_{y_{t-1}, y_t} + f_{t, y_t} \tag{9}$$

τ is a transition matrix where $\tau_{i,j}$ represents the transition score from class y_{t-1} to y_t. The likelihood for a labeling sequence is generated by exponentiating the scores and normalizing over all possible output label sequences.

During inference, CRFs use the Viterbi algorithm to efficiently find the optimal sequence of labels.

4.2.2 BERT

Another popular deep learning model for keyphrase extraction is the Bidirectional Encoder Representations from Transformers (BERT) model. BERT [73], a pre-trained deep neural network model, uses multi-layer bidirectional transformers to pre-train representations for language models from large unlabeled text. Fine-tuning the BERT model has been shown to achieve state-of-the-art performance on a variety of natural language processing tasks such as language understanding, text classification, question answering and NER. Its ability to capture the contextual meaning of words and phrases makes it particularly effective for keyphrase extraction [149, 161].

BERT, a language representation model, utilizes a Transformer architecture [162] that uses stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. Specifically, the encoder reads the input text, which is a sequence of tokens, and the decoder produces a prediction for the task at hand. The BERT's attention mechanism enables the model to learn contextual relationships between words in the text. Since BERT focuses on generating a language representation model, it utilizes only the encoder part. In this process, the input tokens are projected into embedding vectors and are subsequently processed through deep neural network architectures. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings.

- *Token embeddings*: WordPiece embeddings with a 30,000 token vocabulary. A [CLS] token is added to the input word tokens at the beginning of the first sentence and a [SEP] token is used to differentiate the sentences.
- *Segment embeddings*: A marker indicating Sentence A or Sentence B is added to each token. This enables the encoder to differentiate between sentences.
- *Positional embeddings*: To signify the position of each token within the sentence, a positional embedding is incorporated.

Fundamentally, the Transformer utilizes a layer that maps sequences to sequences, resulting in an output sequence of vectors that correspond on a one-to-one basis with the input tokens at the same index. BERT does not aim to predict the subsequent word within the

sentence like traditional language models. The pre-training process for BERT leverages the following two strategies.

- *Masked Language Model*: deep bidirectional model is strictly more powerful than either a left-to-right model or the shallow concatenation of a left-to-right and a right-to-left model. Unfortunately, standard conditional language models can only be trained left-to-right or right-to-left, since bidirectional conditioning would allow each word to indirectly "see itself". The idea is simply to mask some percentage of the input tokens at random, and then predict those masked tokens, based on the context provided by the other non-masked words in the sequence. Out of the random 15% of all tokens in each sequence selected for masking in the pre-trained BERT models: 80% of the tokens are actually replaced with the [MASK] token; 10% of the time tokens are replaced with a random token; 10% of the time tokens are left unchanged. This setting helps to mitigate the issue of a mismatch between pre-training and fine-tuning, because the [MASK] token does not appear during the fine-tuning process.
- *Next Sentence Prediction*: according to [73], many crucial downstream natural language processing tasks, such as Question Answering (QA) and Natural Language Inference (NLI), require a deep understanding of the relationship between two given sentences. However, this relationship is not explicitly captured by traditional language modeling techniques. To address this issue, the training process of BERT includes a next sentence prediction task, which helps the model learn to comprehend the relationship between two given sentences. During training, BERT is fed with pairs of input sentences and learns to predict whether the second sentence is the next sentence in the original text. BERT separates the two input sentences using a special [SEP] token as mentioned earlier. Specifically, the model is trained on pairs of input sentences at a time, where 50% of the time the second sentence comes after the first one (labeled as *IsNext*) and 50% of the time it is a random sentence from the full corpus (labeled as *NotNext*). BERT is then required to predict whether the second sentence is random or not, with the assumption that the random sentence will be disconnected from the first sentence. Although this task is simple, it is highly beneficial for both QA and NLI tasks.

BERT is pre-trained on a combination of the BooksCorpus dataset, which contains approximately 800 million words, and the English Wikipedia, which contains approximately 2.5 billion words. The pre-training process of BERT combines the Masked Language Model and the Next Sentence Prediction tasks to minimize the combined loss function of the two strategies, helping BERT to capture rich contextual representations of words and sentences that can be fine-tuned for specific downstream tasks.

Fine-tuning BERT for downstream tasks is a straightforward process due to the self-attention mechanism in the Transformer architecture. The self-attention mechanism allows BERT to model various downstream tasks involving single text or text pairs by replacing the appropriate inputs and outputs. For token-level tasks, such as sequence tagging and question answering, the token representations obtained from the output layer are fed into a neural network layer, while for classification tasks such as entailment or sentiment analysis, the token representation of [CLS] is utilized.

The use of transfer learning in deep language models such as BERT, which enables the learning of feature representations similar to computer vision, allows for highly effective fine-tuning of downstream tasks. To perform concept extraction on course descriptions, I adopt the BERT architecture. The token representations obtained from BERT's output layer are used as inputs for a token classification task where each token is classified into one of three classes (i.e., *O*, *B-CON* and *I-CON*).

Figure 12 illustrates how the BERT architecture is fine-tuned for the concept extraction task. The token representations obtained from the BERT model are transformed into features that can be used for classification. The process of fine-tuning BERT for concept extraction is highly effective due to the ability of BERT to learn meaningful representations of the input text.

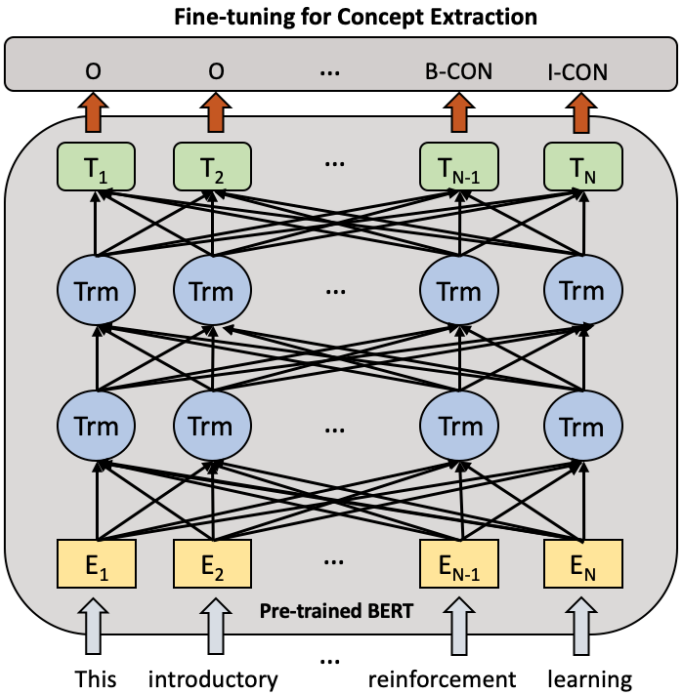


Figure 12: BERT architecture adaption for concept extraction.

4.3 Experiments and Results

4.3.1 Training Datasets

To overcome the problem of manually labeling data for course descriptions, I apply the notion of domain adaption to train the concept extraction models on source datasets close to the target course description, with expert-annotated labels or weak labels. Specifically, several existing labeled datasets in the academic and educational domains will be used to train models to extract concepts for course descriptions.

Introduction to Information Retrieval (IIR) dataset[1]: contains a section-level concept index for the first 16 chapters of the book "Introduction to Information Retrieval" (IIR) [163]. For each section (the lowest-level unit in the Table Of Contents) of the textbook, the dataset provides a list of essential concepts mentioned in that section. The dataset

[1]https://github.com/PAWSLabUniversityOfPittsburgh/Concept-Extraction

Table 5: Statistics of the IIR dataset

Characteristic	
Number of chapters	16
Number of sections	86
Number of all concepts	3175
Number of 1-grams	1121 (35.31%)
Number of 2-grams	1565 (49.29%)
Number of 3-grams	422 (13.29%)
Number of 4-grams	58 (1.83%)
Number of 5+6-grams	9 (0.28%)
Number of all unique concepts	1543
Number of unique 1-grams	278 (18.02%)
Number of unique 2-grams	871 (56.45%)
Number of unique 3-grams	330 (21.39%)
Number of unique 4-grams	55 (3.56%)
Number of unique 5+6-grams	9 (0.58%)

is annotated by three experts. Before the start of the process, the annotators received training and passed a test that focused on the understanding of the task, the "codebook" of annotation rules, and the annotation interface. Every week, three experts focused on completing annotations for one chapter (i.e., all sections that belonged to the chapter). After finishing an annotation session, they discussed the cases in which their annotations disagreed, made the final decision for the concept list, and, if necessary, added new "codebook" rules to help increase the agreement in the future. Throughout this process, the inter-annotator proportion agreement among the three annotators before discussion had gradually increased from 0.25 to 0.68 at week 3 and 0.9 at the end of the whole annotation process. The statistics of the dataset are shown in Table 5.

KP20K[2]: one of the largest datasets in scientific keyphrase studies developed by Rui at el. [62]. It contains the titles, abstracts, and keyphrases of 554,133 scientific articles in the Computer Science domain from various online digital libraries, including ACM Digital Library, ScienceDirect, Wiley, and Web of Science etc.

[2]https://github.com/memray/seq2seq-keyphrase

Wikipedia: the largest dataset in this study. It contains Wikipedia articles including page topics which can be used to filter data to train domain-specific concept extractors. For instance, I select 636,917 articles in, e.g., Computer Science, Information Science, Artificial Intelligence, Information Retrieval and Machine Learning to train models to extract concepts in Computer and Information-related course descriptions. The *weak* concept labels for each Wikipedia article are cultivated using page categories, bold texts, and phrases linked to another article [164, 165].

4.3.2 Implementation Details

Various models are trained with Bi-LSTM-CRF and BERT architectures using three different training datasets separately and two training settings (i.e., *uncased* texts and *cased* texts). In total, there are six main models trained for each architecture, plus a combined model that combines the outputs of the six models together:

- *uncased-iir*: model trained with *IIR* dataset in the *uncased* setting
- *cased-iir*: model trained with *IIR* dataset in the *cased* setting
- *uncased-kp20k*: model trained with *KP20K* dataset in the *uncased* setting
- *cased-kp20k*: model trained with *KP20K* dataset in the *cased* setting
- *uncased-wiki*: model trained with *Wikipedia* dataset in the *uncased* setting
- *cased-wiki*: model trained with *Wikipedia* dataset in the *cased* setting
- *combined-all*: model combines the output of each of the above models together

One effective technique for improving the performance of machine learning models is to use an ensemble of multiple models rather than a single model to make predictions. Each model in the ensemble is referred to as a *base learner*. Ensemble is the algorithm of choice for many winning teams in machine learning competitions. There are three approaches to creating an ensemble: bagging, boosting, and stacking. Stacking involves training base learners on the training data and then creating a *meta-learner* that combines the outputs of the base learners to produce the final predictions. The *combined-all* model in this study is particularly a stacking ensemble model consisting of six base learners. To generate concept predictions from a given text (i.e., a course description), the final prediction is computed

as the combination of all the outputs from the base learners. If two concepts, which are outputs from the base models, are adjacent or overlapping, they are concatenated into a new concept. All the trained models are available on this Github repository.[3]

The implementation of Bi-LSTM-CRF models is based on the version[4] presented in [57] at the sentence level. The character embeddings of 30 dimensions are obtained by training additional Bi-LSTM networks along with the main model. For the word embeddings, the Glove pre-trained word embeddings of 100 dimensions[5] are used. A 300-dimensional hidden layer of LSTM units is used for both the character-level embedding model and the main model. The models are trained using mini-batch stochastic gradient descent with momentum. The batch size, learning rate and decay ratio are set to 10, 0.015 and 0.05, respectively. The dropout strategy is also applied to avoid over-fitting and gradient clipping of 5.0 to increase the model's stability.

For BERT models, a distilled version (DistilBERT) is used. It is smaller, faster, cheaper and lighter, yet achieves competitive performances compared to the original architecture [166].

4.4 Model Performances

To evaluate and compare the performance among the models, I created a test set including 50 randomly selected course descriptions. I manually annotate concept labels for each of the course descriptions. The trained models are evaluated using standard keyphrase extraction metrics on this set of 50 course descriptions.

Table 6 shows the performance of two different models, BERT and BI-LSTM-CRF, for concept extraction. The performance is measured in terms of precision, recall, and F1 score for three different training datasets (i.e., IIR, KP20K and Wikipedia) and two training settings (i.e., *case* and *uncased*). The *combined-all* row represents the performance of the stacking ensemble models which combines the outputs of the six based models for each of

[3]https://github.com/HungChau/course-concept-extraction
[4]https://github.com/LiyuanLucasLiu/LM-LSTM-CRF
[5]https://nlp.stanford.edu/projects/glove/

Table 6: Model performance summary of BERT and BI-LSTM-CRF on a task of concept extraction for course descriptions.

	BERT			**BI-LSTM-CRF**		
	precision	recall	f1	precision	reall	f1
uncased-iir	**0.738**	0.269	0.394	0.773	0.256	0.385
cased-iir	0.726	0.281	0.405	**0.844**	0.208	0.334
uncased-kp20k	0.515	0.195	0.283	0.714	0.111	0.193
cased-kp20k	0.537	0.217	0.309	0.570	0.148	0.236
uncased-wiki	0.629	0.259	0.367	0.797	0.285	0.420
cased-wiki	0.608	0.248	0.352	0.809	0.313	0.452
combined-all	0.733	**0.556**	**0.633**	0.799	**0.503**	**0.617**
	BERT + BI-LSTM-CRF					
	precision		recall		f1	
deep-concept-extractor	**0.758**		**0.625**		**0.685**	

the deep architectures separately. The last row of the table shows the performance of all BERT and BI-LSTM-CRF models combined on a test set. The best performing model for each dataset and metric (or overall) is highlighted in bold.

As can be seen from Table 6, in terms of F1 score, BERT demonstrates superior performance on IIR and KP20K datasets, while BI-LSTM-CRF performs better on the Wikipedia dataset. The stacking ensemble model exhibits the best performance in terms of F1 score, regardless of the architecture employed, with the BERT ensemble model surpassing the BI-LSTM-CRF ensemble model. The best performance across all three metrics, namely precision (0.758), recall (0.625), and F1 score (0.685), is achieved by combining the BERT and BI-LSTM-CRF models.

Here is an example of concepts extracted from an actual Algorithm course at Pitt. The yellow-highlighted phrases represent the concepts extracted by the *combined* model.

"*This course emphasizes the study of the basic data structures of computer science (stacks, queues, trees, lists) and their implementations using the java language included in this study are programming techniques which use recursion, reference variables, and dynamic memory allocation. Students in this course are also introduced to various searching and sorting methods and also expected to develop an intuitive understanding of*

Table 7: Expert evaluation dataset statistics

Number of course descriptions	50
Average number of words per description	72.0
Average number of extracted concepts per description	12.64
Number of extracted concepts	632
Number of unique extracted concepts	519

the complexity of these algorithms..."

4.5 Expert Evaluation

In addition to the offline evaluation on a test set, I also conduct an expert evaluation to ensure the quality of extracted concepts for course recommendation applications. I hire two PhD students who specialize in Computing and Information Science. To begin the evaluation, I randomly sample 50 course descriptions in SCI. For each description, I apply the trained model to extract concepts in the text. The two experts are then provided with the course descriptions and a list of extracted concepts for each course. The experts are asked to rate each extracted concept as either good or not good for the corresponding course. Table 7 summarizes the statistics of the evaluation dataset.

Table 8 presents the results of the expert evaluation conducted to measure the performance of the *deep-concept-extractor* model. The table reports four metrics: macro accuracy, micro accuracy, proportional agreement, and Kappa agreement. Macro accuracy measures the accuracy of the experts' evaluations at a high level for all the course descriptions together, while micro accuracy measures the average accuracy at an individual course level. The results show that both experts have high levels of agreement with the concepts extracted by the model, with macro accuracy ranging from 90.51% to 90.98%, and micro accuracy ranging from 89.54% to 90.43%. Furthermore, the proportional agreement between the two experts was 92.88%, indicating that they agreed on their assessments most of the time. Finally, the

Table 8: The result of the expert evaluation.

Metric	**Expert 1**	**Expert 2**	**Both Experts**
Macro accuracy	90.98%	90.51%	87.18%
Micro accuracy	90.43%	89.54%	86.09%
Proportional agreement	–	–	92.88%
Kappa agreement	–	–	0.57

table reports the Kappa agreement between the two experts, which measures the level of agreement between them beyond what would be expected by chance. The Kappa agreement between both experts was 0.57, indicating a good level of agreement. Overall, the results of the expert evaluation illustrate that the two experts frequently agree with the outputs of the concept extraction model. Consequently, the combined BERT and BI-LSTM-CRF models will be employed to extract concepts from descriptions for explainable course recommendation systems in the next chapters.

4.6 Summary and Discussion

In this chapter, I applied several machine learning techniques to build a concept extraction model for course descriptions without manually labeled data for training. Transfer learning has proven to be a valuable approach in addressing the challenge of using representations that are first pre-trained on large quantities of unannotated data and then adapted to guide other downstream tasks, which has been demonstrated to be highly effective even with limited or weak labels. Deep neural network models for sequence tagging with various architectures, such as CNNs, RNNs, and Transformers, have brought breakthroughs in NLP systems. I approached concept extraction from course descriptions as a sequence labeling task with the state-of-the-art deep learning architectures, BERT and BI-LSTM-CRF. I trained the concept extraction models on several public datasets, and then stacked them together as an ensemble model for concept extraction to improve model effectiveness.

The final model is directly adapted to extract concepts in another type of document, such as course descriptions, without requiring specific domain adaptation techniques. It's worth noting that this straightforward domain adaptation approach works better if the source and target distributions of the input "domains" are close to each other. However, employing specific strategies such as supervised or semi-supervised domain adaptation could lead to enhanced performance.

I conducted an evaluation and comparison of BERT and BI-LSTM-CRF models, for concept extraction using standard keyphrase extraction metrics on a set of 50 course descriptions. The performance of the models was measured in terms of precision, recall, and F1 score for three different training datasets and two training settings. The stacking ensemble model showed the best performance in terms of F1 score, regardless of the architecture employed, with the BERT ensemble model surpassing the BI-LSTM-CRF ensemble model. The best performance across all three metrics was achieved by combining the BERT and BI-LSTM-CRF models.

In addition to the offline evaluation, an expert evaluation was conducted to ensure the quality of extracted concepts for course recommendation applications. The results showed that both experts had high levels of agreement with the concepts extracted by the model. The proportional agreement between the two experts was 92.88%, indicating that they agreed on their assessments most of the time. Finally, the Kappa agreement between both experts was 0.57, indicating a good level of agreement.

Overall, the results of the study demonstrate that the combined BERT and BI-LSTM-CRF models can be effectively employed to extract concepts from descriptions for explainable course recommendation systems. The expert evaluation further validates the quality of the extracted concepts, indicating the potential for practical applications of the models in the field of education.

5.0 SKILL-BASED EXPLANATIONS FOR SERENDIPITOUS COURSE RECOMMENDATION

The efficacy of course recommender systems is increasingly dependent on their ability to clarify to students the reasons behind specific course suggestions. Within a university setting, it can be valuable for these systems to expose students to courses they might not have considered, yet are still relevant to them. A challenge arises when students find it difficult to judge the relevance of unfamiliar courses. Our preliminary findings, detailed in Section 3.2, indicate the potential benefits of providing course recommendations with explanations; especially utilizing students' prior knowledge could be an effective way to craft explanations. However, unigram skills failed to provide effective explanations. This chapter suggests that enriching course recommendations with skill-based justifications can improve the system's value. The deep concept extractor developed in Chapter 4 enables me to extract concepts (representing 'skills' discussed in Chapter 1, Section 1.1) to pioneer skill-based explanations in a deep learning-based serendipitous course recommendation system for higher education. By delivering in-depth information about a course alongside its rationale for recommendation, students are better positioned to assess its relevance. This could potentially decrease the tendency towards neutral opinions and reduce the likelihood of students overlooking a course due to unfamiliarity. To validate this hypothesis, I conducted a user study using the AskOski system, in collaboration with the CAHL lab led by Dr. Zach Pardos at the University of California, Berkeley, and our collaborator, Run Yu.

5.1 Introduction

In recent years, the realm of course recommendation has seen a surge in interest, with deep learning emerging as a prominent player. One notable example of this trend is the AskOski system at the University of California, Berkeley. This innovative system harnesses historical enrollment data and employs a collaborative approach, strengthened by deep learn-

ing, to offer course recommendations that are tailored to the unique interests of individual students across the campus. Furthermore, it integrates with the campus degree audit system to offer personalized course suggestions that address the unfulfilled graduation requirements of the students.

The serendipitous course recommendation system aims to recommend courses that are unexpected or novel yet still relevant. The underlying hypothesis is that students are more likely to accept such recommendations. Yet, this task presents significant challenges, particularly within a university setting. In this context, relevant but unexpected courses may belong to departments outside of a student's primary field of study, and their course descriptions may employ unfamiliar terminology, potentially making them less likely to be adopted by students. Previous research, as demonstrated in our study [38], underscores the effectiveness of catering to students' prior knowledge by providing personalized explanations. However, it is posited that the efficacy of these explanations can be further enhanced by using keyphrases instead of unigrams. Unigrams, single words devoid of context, may struggle to convey the full meaning encapsulated in course descriptions, particularly in cases where technical terminology is utilized. Keyphrases, on the other hand, have shown promise in improving user comprehension over unigrams, as they are better equipped to communicate the underlying semantics [34, 35, 143, 144].

In this context, I hypothesize that augmenting course recommendations with skill-based explanations could substantially enhance measures within higher education course recommender systems. Specifically, I propose that by furnishing students with comprehensive information about a course, including how it aligns with their prior knowledge and the novel knowledge it offers, students will be better equipped to evaluate its relevance, be more confident in making decisions and be less likely to dismiss it based on unfamiliarity. In collaboration with the CAHL lab at the University of California, Berkeley, we conduct an online user study at the same institution, leveraging the AskOski system powered by PLAN-BERT, an adaptation of BERT4Rec, which is a state-of-the-art deep neural network model for top-N recommendation [39].

In this chapter, I will begin by providing a concise overview of PLAN-BERT, the deep learning-based course recommendation engine integrated into the AskOski system, which

has been developed by the CAHL lab. Subsequently, I will describe the explanation method, elucidate the study's experimental procedures and analyses in detail, and discuss its contribution, limitation, and future direction to conclude the chapter.

5.2 Recommendation Method

In the realm of recommendations, course recommendation for exploration could be considered as the top N item recommendation problem. Early works on this type of recommender system predominantly utilized content-based or collaborative-based approaches. Some of these models leverage course representation models (e.g., *skill*-based or course2vec), as discussed in previous sections, to generate recommendations. To consider time or order information, state-of-the-art technologies employ deep sequential models such as Recurrent Neural Networks or transformer-based models like BERT. Deep learning has gained a considerable amount of interest in numerous research areas such as natural language processing (NLP), speech recognition and computer vision. This surge in interest is not only due to their remarkable performance but also due to their inherent capability of transfer learning and crafting feature representations from scratch. Deep neural networks are also composite in the sense that multiple neural building blocks can be composed into a single big differentiable function and trained in an end-to-end manner. This is the key advantage when coping with content-based recommendations and inevitable when modeling users and items, where multi-modal data is ubiquitous. Deep learning models have successfully been applied to top N recommendation problem and shown to be superior over the traditional approaches. Consequently, to curate a list of potential courses to recommend to students based on their course enrollment history, we use BERT4Rec [39], a leading model for top-N recommendations. PLAN-BERT [95] (adapted from BERT) has recently been studied and evaluated for multi-semester course recommendation. Its bidirectional self-attention design is more effective at utilizing past sequence information compared to both BiLSTM and a UserKNN baseline. Additionally, PLAN-BERT incorporates student characteristics (such as major, division, and department) and course attributes (such as subject and department) to enhance

personalization and enhance the quality of recommendations. The online study demonstrated that PLAN-BERT has practical potential to assist students as they navigate the complexities of higher education. Consequently, we use a modified version of PLAN-BERT (refer to Figure 13) to generate a list of course suggestions for a specific semester.

Training Process: in order to efficiently train the BERT architecture with sequential data, we will apply the Masked Language Model, a masking technique in NLP. We will employ percentage sampling to pre-train PLAN-BERT to learn the contextual embeddings of courses by predicting the original IDs of the masked courses based only on their left and right contexts. The input format for BERT to learn course embeddings, for example, is:

$$[c_1, c_2, c_3, c_4, c_5, c_6, c_7] \xrightarrow[mask]{randomly} [c_1, [mask]_1, c_3, c_4, c_5, [mask]_2, c_7]$$

At the main training stage, we will fine-tune PLAN-BERT on the next item prediction task. The courses in the latest historical semester will be masked for prediction (see Figure 13.A). In this way, the trained model can learn the relationship between past and future courses.

Diversifying Recommendations: Among the institutional values of a large multidisciplinary university is to expose students to not solely complementary knowledge but also different viewpoints expressed through courses. It helps students expand their learning and collaboration and experience various intellectual schools of thought across the university. The goal of our course recommendation is to help students improve awareness of course options and explore courses they may find interesting but which have been relatively unexplored by those with similar course selections to them in the past. To achieve this goal and counteract the filter bubble issue of collaborative filtering based recommendation models, we diversify course suggestions by allowing only one result per department [21, 38], produced by PLAN-BERT.

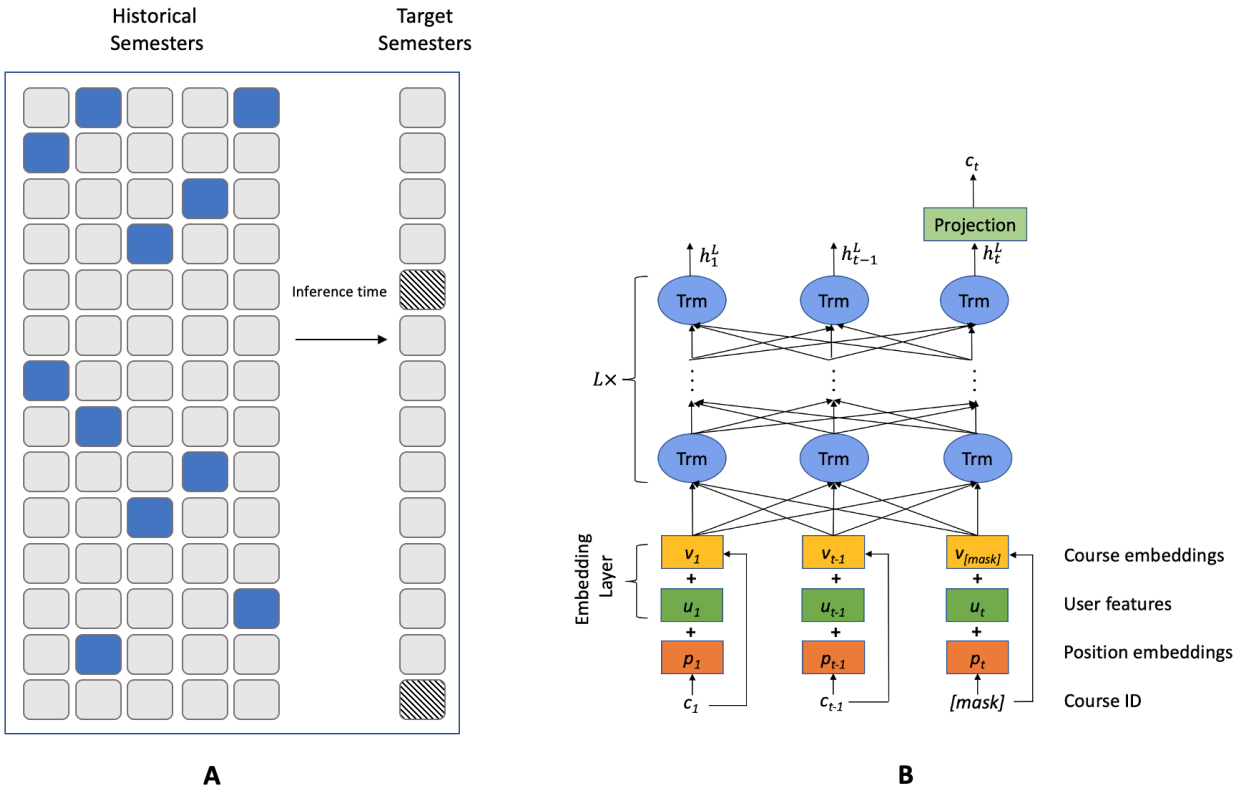


Figure 13: A) Student course enrollment history for training PLAN-BERT: before the inference time is the input and courses after the inference time are masked as the prediction targets; each column represents a semester in the student enrollment history; blue cells represents enrolled courses; and striped cells are enrolled courses in the latest historical semester and masked for prediction. B) BERT architecture for next course prediction task using student course enrollment histories and major information. The position embeddings can be encoded as relative semesters elapsed since the student began.

5.3 Explanation Method

One of the primary objectives of this study is to investigate how providing explanations or justifications impacts and enhances user responses to the recommendations generated by the method described in the previous section. Preliminary results from a prior study [38] indicated that presenting students with prior knowledge (i.e., skills shared between

the suggested course and the courses the student has taken) was an effective method for generating personalized explanations that led to increased average ratings across all outcome measures. Furthermore, presenting multi-gram skills instead of unigrams in explanations could potentially further enhance the recommendation outcomes. Unigrams may not have sufficient capacity to convey the nuanced meanings encapsulated in course descriptions. It can be challenging for students to interpret the meaning of individual words, especially technical terms. Therefore, we will utilize multi-gram concepts (representing 'skills' discussed in Chapter 1, Section 1.1), extracted by the trained model presented in Chapter 4, to provide these explanations.

The methodology. Our approach to providing explanations consists of two key aspects: (1) establishing connections between the target course recommendation and skills from courses the student has previously taken, and (2) unveiling novel skills that are taught in the target course. Consequently, the explanation for a recommended course will consist of two separate lists of skills, offering the student both familiar knowledge they have already acquired and new knowledge they have yet to encounter. The two lists of skills are defined as follows:

$$Learned\_Skills = S_t \cap \ (\bigcup_{c \in C} S_c) \tag{10}$$

$$New\_Skills = S_t - (\bigcup_{c \in C} S_c) \tag{11}$$

Where S_t is the set of extracted skills for the target recommended course's description, S_c is the set of extracted skills for a taken course's description, and C is the list of courses the student has taken in the past.

Skill ranking: To assess the relevance of an extracted skill to the target course, I compute the relationship, denoted as $r_c(s)$, between the skill and the course. First, each skill and course description are represented as embedding vectors with 768 dimensions, utilizing the *all-mpnet-base-v2* version of SBERT. Subsequently, I calculate the relationship ($0 <= r_c(s) <= 1$) using cosine similarity. This cosine score is employed to rank the list of skills, with the top N skills selected to construct the explanation.

Skill matching: two skills could be semantically equivalent but may have different word forms (e.g., *K-Means Clustering* and *K-Means Algorithm*). Exact string matching could lead to missing overlapping skills between the *target* course and a *taken* course. Therefore, I employ a *soft* matching approach, utilizing cosine similarity between two embedding vectors representing the two skills. These embedding vectors are derived from SBERT. Two skills are considered a match if their cosine similarity ($r(s_1, s_2)$) exceeds *0.85*. Note that this heuristic threshold is chosen based on experimental analysis with the SBERT's embeddings. This threshold could be relaxed depending on specific applications or adjusted for different types of embeddings.

5.4 Study Experiments

This section outlines the implementation of the proposed skill-based explanation for serendipitous course recommendation at at the University of California, Berkeley. To assess its effectiveness, we conducted a online user study involving undergraduate students, soliciting their insights on multiple dimensions. The study aims to empirically investigate the hypothesis that enhancing course recommendations with explanations can empower students to more effectively assess the relevance of suggested courses. This augmentation is anticipated to reduce the prevalence of neutral opinions and mitigate the likelihood of students disregarding recommendations due to unfamiliarity.

5.4.1 Implementation Details

PLAN-BERT underwent retraining using enrollment history data up to Fall 2022. From the initial pool of 20,282 courses, we excluded 553 courses with inadequate descriptions (i.e., fewer than 7 words), resulting in a refined collection of 20,729 courses available for recommendations across all majors. Using student enrollment data and major information as inputs, PLAN-BERT ranks these courses for each student's recommendation. The final recommendations comprise the top 5 courses from 5 distinct departments, ensuring a diverse

selection for students. The final five recommended courses are shown to participants in a random order.

To extract the skills encapsulated within these courses, I employed the concept extraction model detailed in Chapter 4. This process entailed post-processing, including the removal of generic skills such as 'homework', 'student' and 'seminar', skills containing more than 5 words, as well as the consolidation of singular and plural forms of skills. After post-processing, on average, there are 6.5 extracted skill per course.

Both individual skills and course descriptions are transformed into embedding vectors with 768 dimensions, utilizing the *all-mpnet-base-v2* version of SBERT. Subsequently, I calculated the relationships between skills and courses ($0 \leq r_c(s) \leq 1$) via cosine similarity. This cosine score is used to rank the list of skills associated with the course.

To provide insightful explanations for each course recommendation, I employ the skill matching methodology described earlier, comparing the target course with the courses the student has previously taken. This approach identifies the top 7 acquired skills and the top 7 new skills,[1] which are then presented to the student alongside the course recommendations, offering valuable insights and aiding in informed decision-making.

5.4.2 User Study

Procedures. This study is conducted online through the AskOski system at the University of California, Berkeley. Student participants are required to enroll for a minimum of two semesters to take part in the study. They are randomly assigned to one of the two *between*-subject conditions (*Explanation*). The study begins with participants logging into the AskOski system using their Berkeley credentials. Upon accessing the system, participants are presented with the study's introduction. Subsequently, they are presented with a curated list of five course recommendations. These recommendations are tailored based on the participant's past course history, major information, as well as the course history of "similar" students. Each course recommendation includes the course ID and title, the description of the course, skill-based explanation (available only for participants in the *Ex-*

[1]The actual number of skills shown to the subject may be less than 7, depending on how many skills are in the course and how many skills are matched.

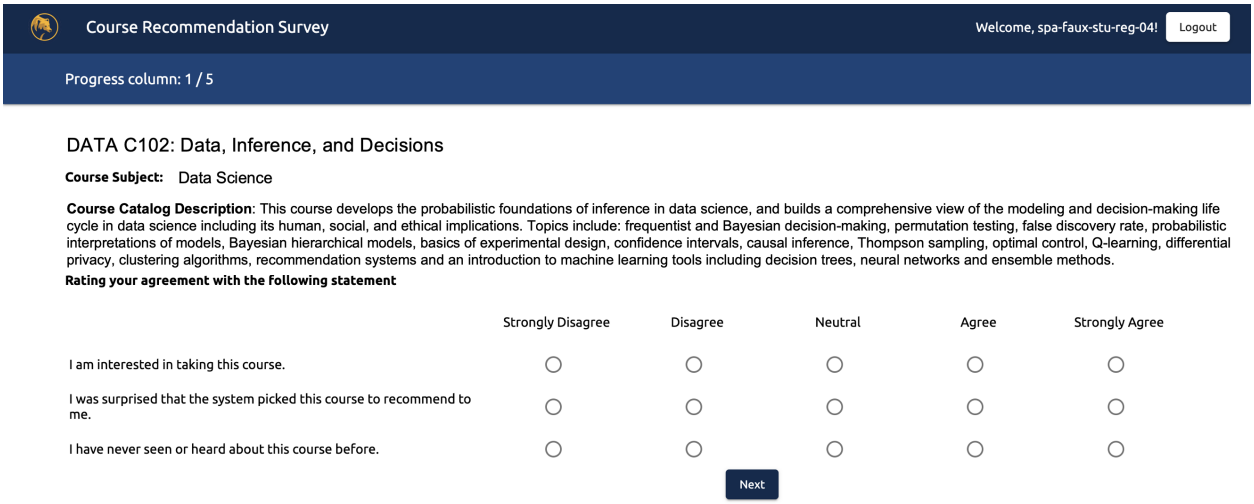


Figure 14: A demonstration of a recommended item with no explanation (group C1) through the AskOski system.

planation conditions), and a survey questionnaire consisting of multiple-choice questions (see Fig. 14 and 15). Participants are requested to thoroughly review each of the five course recommendations and respond to a series of questions regarding their preferences and feedback for each recommended course. In total, each participant is expected to evaluate and provide feedback on all five recommended courses.

Participants. A total of 53 participants were recruited for this study through a combination of social media and email advertisements. Eligibility for participation was limited to undergraduate students at the University of California, Berkeley, who had completed a minimum of two semesters. Among the recruited participants, 28 were assigned to the explanation condition, and 25 were in the no-explanation one. These participants come from a diverse array of academic backgrounds, encompassing fields such as Computer Science, Data Science, Economics, Statistics, and Media Studies, among others. The study was designed to be completed within 30 - 45 minutes. Each participant received a $20 Amazon gift card upon the successful completion of the study.

Design and Analysis. The study is designed as a *between*-subjects study to measure the effect of explanation on the serendipitous course recommendation w.r.t *success*, *unex-*

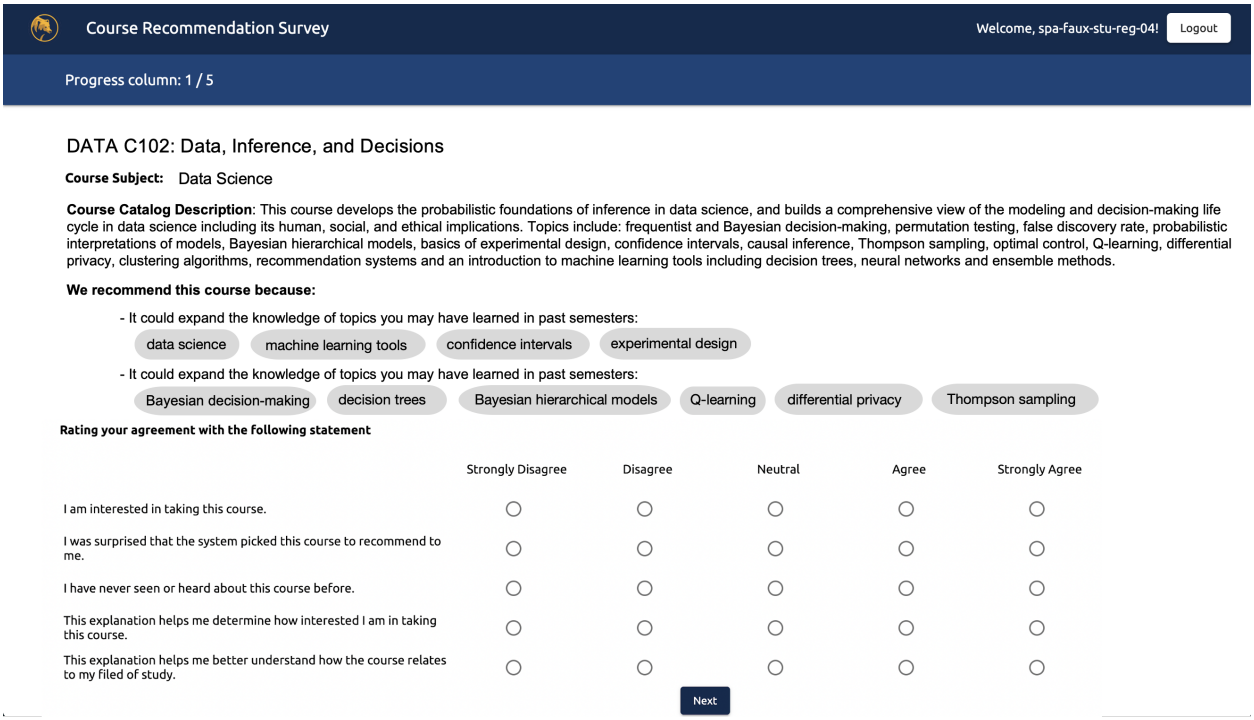


Figure 15: A demonstration of a recommended course for with skill-based explanation (group C2) through the AskOski system. The explanation shows the top 7 learned concepts as well as the top 7 novel concepts offered by the course.

pectedness and *novelty*. There are two *between*-subject conditions (*Explanation*): *No-Exp* (C1) vs. *Exp* (C2).

I collected the following measures:

- Q1. Success (Interest): Participants respond to the statement "*I am interested in taking this course.*" [167, 21, 38] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (see Fig. 14).
- Q2. Unexpectedness: Participants respond to the statement "*I was surprised that the system picked this course to recommend to me.*" [168] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (see Fig. 14).
- Q3. Novelty: Participants respond to the statement "*I have never seen or heard about this course before.*" [167] on a 5-point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (see Fig. 14).
- Q4. Explanation Effectiveness: Participants respond to the statement "*This explanation helps me determine how interested I am in taking this course.*" [105] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (see Fig. 15).
- Q5. Usefulness of Concepts: Participants respond to the statement "*The explanation helps me better understand how the course relates to my field of study.*" on a 5-point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (see Fig. 15).

In our study, each participant evaluate several recommended items. Considering repeated measurements made by the same participant as independent could potentially result in a violation of correlated errors [169, 170]. To tackle this problem, I employ Generalized Linear Mixed Models for the analyses. These models consider that the ratings are provided by the same users, treating them as random effects, and permitting the estimation of error correlations stemming from the repeated measurements.

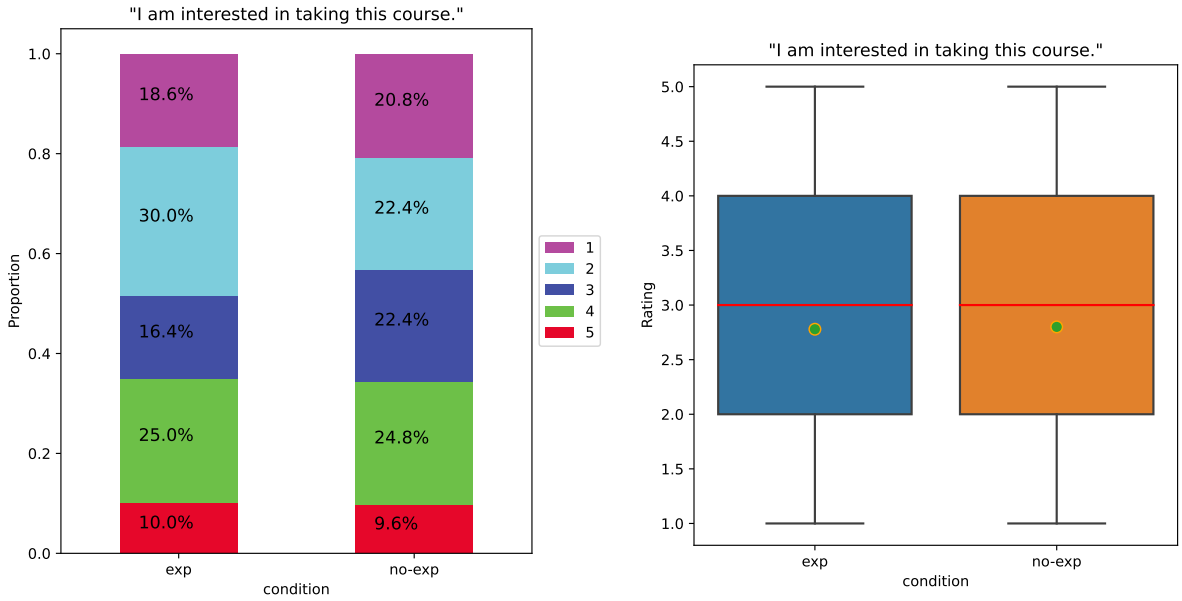
5.4.3 Results

Interestedness. When comparing the *baseline* model, representing only the intercept, with the *random intercept* model, which accounts for variations among different participants,

the findings revealed statistically significant disparities in intercepts across participants. Consequently, we incorporated *participant* random effects into the main analysis. As illustrated in Figure 16, it is generally observed that participants in the *Exp* conditions ($M = 2.78$, $N = 140$) display slightly less interest in enrolling in the recommended courses compared to those in the *No-Exp* conditions ($M = 2.8$, $N = 125$). Our statistical analysis further indicates that there is a significant variation in intercepts across participants concerning the relationship between the provision of an explanation and participants' interest in taking the course, $SD = 0.42$ (95% CI: 0.25, 0.71), $= \chi^2(1) = 5.108, p = .02$. However, the explanation itself does not appear to have a significant effect on participants' level of interest in enrolling in the course, b = -0.021, $t(51)$=0.11, $p = 0.91$.

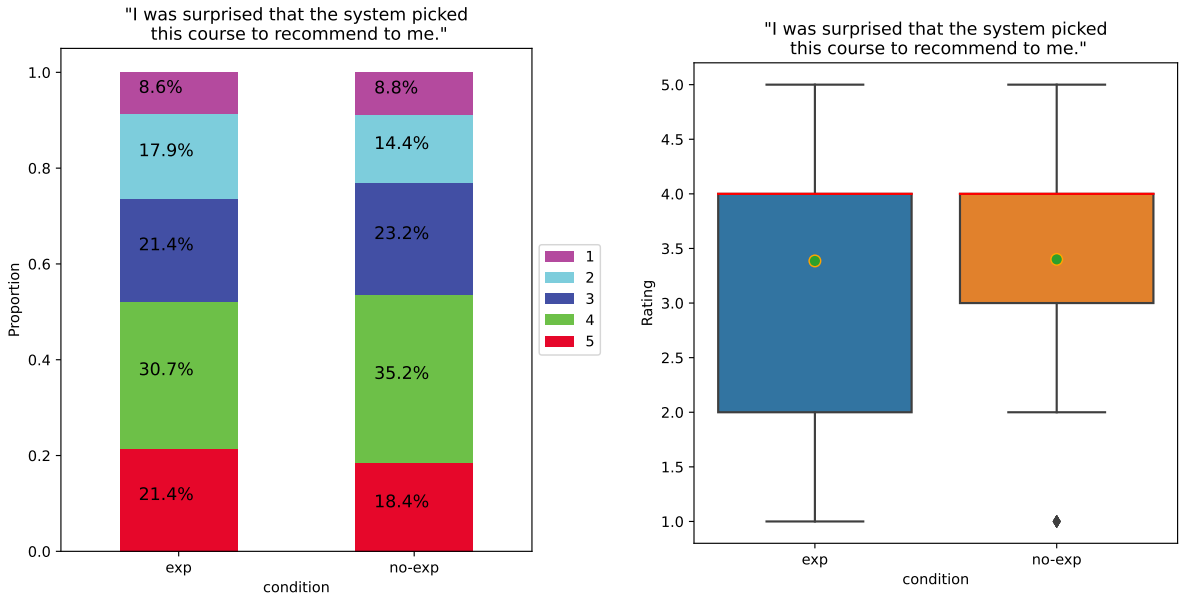
Unexpectedness. Likewise, when comparing the results of the *baseline* model (i.e., only the intercept) and the *random intercept* model (which accounts for variations among participants), the results showed statistically significant variations in intercepts across participants. As a result, the random effects of the participants are included in the primary analysis. As shown in Fig. 17, in general, participants perceive the recommendations as highly unexpected; participants in *Exp* conditions ($M = 3.39$, $N = 140$) show similar level of unexpectedness about the recommendations compared to those in *No-Exp* conditions ($M = 3.4$, $N = 125$). From the statistical analysis, the results show that the relationship between explanation and the unexpectedness of the course showed significant variance in intercepts across participants, $SD = 0.47$ (95% CI: 0.31, 0.72), $= \chi^2(1) = 9.905, p = .001$. However, the explanation has no significant effect on how surprisingly the participants perceive the course, b = -0.014, $t(51)$=0.074, $p = 0.94$.

Novelty. Similarly, there are statistically significant variations in intercepts across participants, which were included in the primary analysis as a random effect. Fig. 18 shows that participants generally perceive the recommendations as highly novel, with an average rating of 3.54 out of 5. Participants in the *Exp* conditions ($M = 3.44$, $N = 140$) perceive the recommended courses as slightly less novel compared to those in the *No-Exp* conditions ($M = 3.65$, $N = 125$). From the statistical analysis, the results show that the relationship between explanation and the novelty of the course exhibits significant variance in intercepts across participants, $SD = 0.58$ (95% CI: 0.39, 0.86), $= \chi^2(1) = 12.46, p = .004$. Having



(a) (b)

Figure 16: (a) Proportional distribution of user responses to the statement 'I am interested in taking this course.', comparing those with *Explanation* (exp) and *Without Explanation* (no-exp). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'. (b) A graph displaying the distribution of ratings in response to research question Q1 for the two conditions, with the median indicated by red lines and the average represented by green circles.



(a) (b)

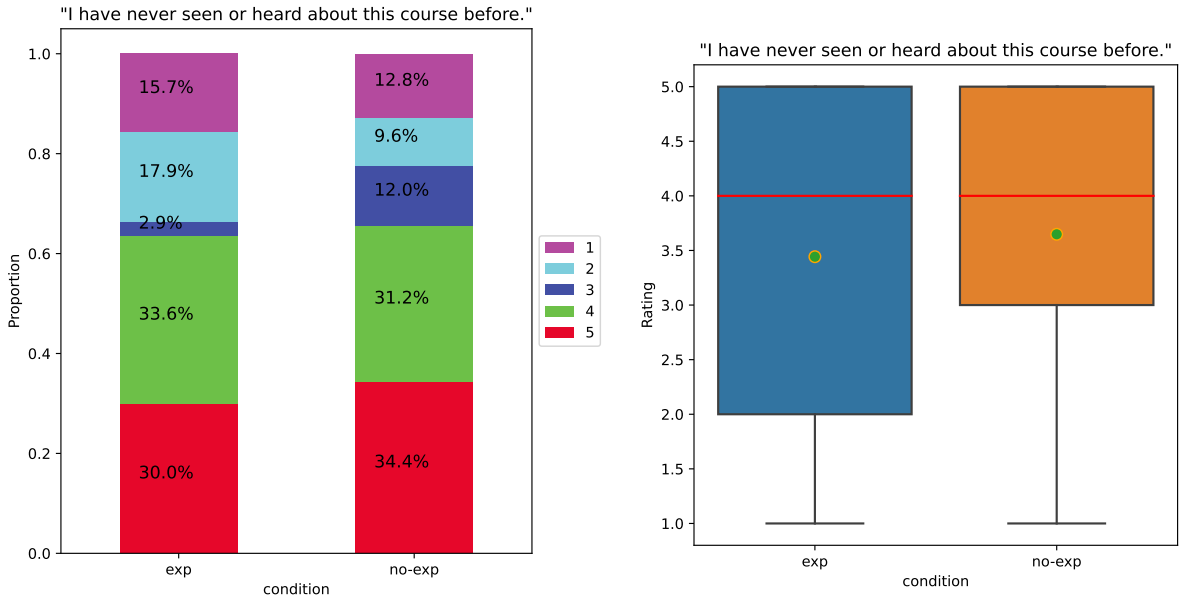
Figure 17: (a) Proportional distribution of user responses to the statement 'I was surprised that the system picked this course to recommend to me.', comparing those with *Explanation* (exp) and *Without Explanation* (no-exp). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'. (b) A graph displaying the distribution of ratings in response to question Q2 for the two conditions, with the median indicated by red lines and the average represented by green circles.

explanations slightly decreases the level of novelty perceived by participants, but it is not statistically significant, b = -0.205, $t(51)$=0.902, p = 0.371.

Serendipity. Similar to the preliminary investigation described in Section 3.2 of Chapter 3, we evaluated serendipity by computing the mean of user-perceived unexpectedness and success [142, 38]. In our primary analysis, we also factored in statistically significant variations in intercepts among participants, treating them as random effects. As depicted in Figure 19, it's evident that, on the whole, participants in the *Exp* conditions (M = 3.08, N = 140) exhibited a similar level of serendipity regarding the recommendations compared to those in the *No-Exp* conditions (M = 3.1, N = 125). Our statistical analysis showed significant variance in intercepts among participants in the relationship between explanations and the level of serendipity of the course, SD = 0.33 (95% CI: 0.24, 0.46), $= \chi^2(1) = 21.20, p < .0001$. Yet, explanations had no impact on participants' perception of the serendipity of the course, with a coefficient of b = -0.017, $t(51)$=0.15, p = 0.881.

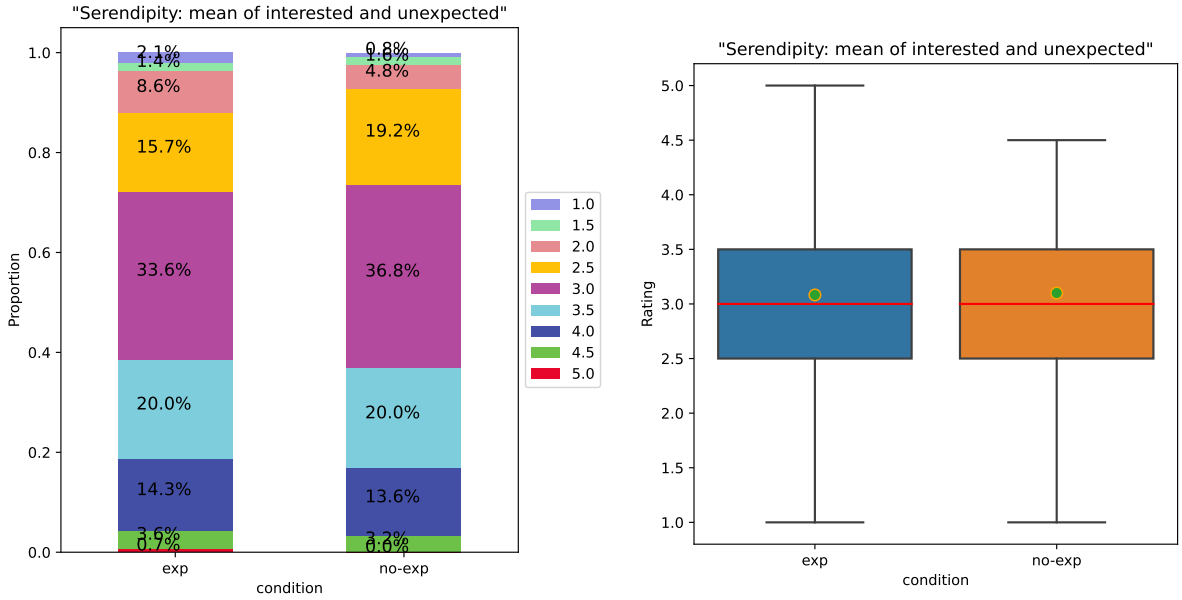
Assessing serendipity poses a formidable challenge. While existing literature suggests that it can be approximated as the mean of user-perceived unexpectedness and success, it is imperative to acknowledge that highly unexpected items often yield lower perceived relevance. Our study supports this observation, as we found a notable negative relationship of -0.39 between relevance and unexpectedness, suggesting that subjects tend to prefer courses that are less likely to provide surprises. However, it is noteworthy that items characterized by both high unexpectedness and high relevance are the ones most valued by users.

To investigate further into the impact of explanations on user perception of recommendation interest across varying levels of unexpectedness, we categorized the 5-point Likert scale ratings of unexpectedness into 'low' (comprising 'Strongly Disagree' and 'Disagree') and 'high' (comprising 'Strongly Agree' and 'Agree') unexpectedness, with neutral ratings excluded from the analysis. As depicted in Figure 20, our findings indicate that, in comparison to courses with low unexpectedness, courses with high unexpectedness elicited reduced interest from participants. Specifically, when a course presented lower levels of unexpectedness, participants in the absence of explanations demonstrated a 0.234-point higher interest in pursuing the course. While this suggests that providing additional information about a course (including its potential knowledge benefits) may prompt subjects to realize its limited



(a) (b)

Figure 18: (a) Proportional distribution of user responses to the statement 'I have never seen or heard about this course before.', comparing those with *Explanation* (exp) and *Without Explanation* (no-exp). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'. (b) A graph displaying the distribution of ratings in response to question Q3 for the two conditions, with the median indicated by red lines and the average represented by green circles.



(a) (b)

Figure 19: (a) Proportional distribution of the average ratings of questions Q1 and Q2 as a measure for serendipity. Original ratings of Q1 ad Q2: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'. (b) A graph displaying the distribution of the average ratings of questions Q1 and Q2 for the two conditions, with the median indicated by red lines and the average represented by green circles.

utility, this effect did not reach statistical significance (p-value = 0.445). Conversely, when a course exhibited high levels of unexpectedness, participants provided with explanations displayed a 0.22-point increase in their interest in taking the course. This effect also did not achieve statistical significance (p-value = 0.304). Nevertheless, it is noteworthy that explanations facilitated a more informed assessment of the course's utility and reduced the likelihood of its dismissal due to unfamiliarity.

Explanation. In accordance with the findings presented in Figure 21, it is evident that participants who were provided with explanatory information expressed a favorable disposition towards the utility of these explanations in influencing their interest in the recommendations (mean = 3.46, N = 265). Notably, a significant majority of respondents endorsed either 'Agree' or 'Strongly Agree' in response to this assertion. This implies that the provision of explanations serves as a valuable resource for participants, affording them a deeper understanding of the recommendations, which in turn facilitates informed decision-making. When being asked about how well the explanations assist participants in understanding the course's relevance to their field of study, it's clear that participants, as a whole, hold a fairly neutral position on this issue. The average rating of 3.04 out of 5 (N = 265) indicates a lack of strong agreement on whether the explanations effectively clarify how the course aligns with their academic interests.

5.4.4 A deeper analysis - Does explanation improve confidence in making decisions?

The hypothesis posits that when subjects are provided with explanations, they are better equipped to assess the utility of recommendations, feel more confident in their decision-making, and are less likely to give a neutral rating to an item. This would suggest that explanations play a pivotal role in guiding their choices by helping them discern their preferences. Additionally, as students who have not yet declared their majors (a majority in their early program) may possess less knowledge about courses and have more options to choose in comparison to those who have already declared their majors, we anticipate the presence of an interaction effect between the declaration of a major and the provision of explanations.

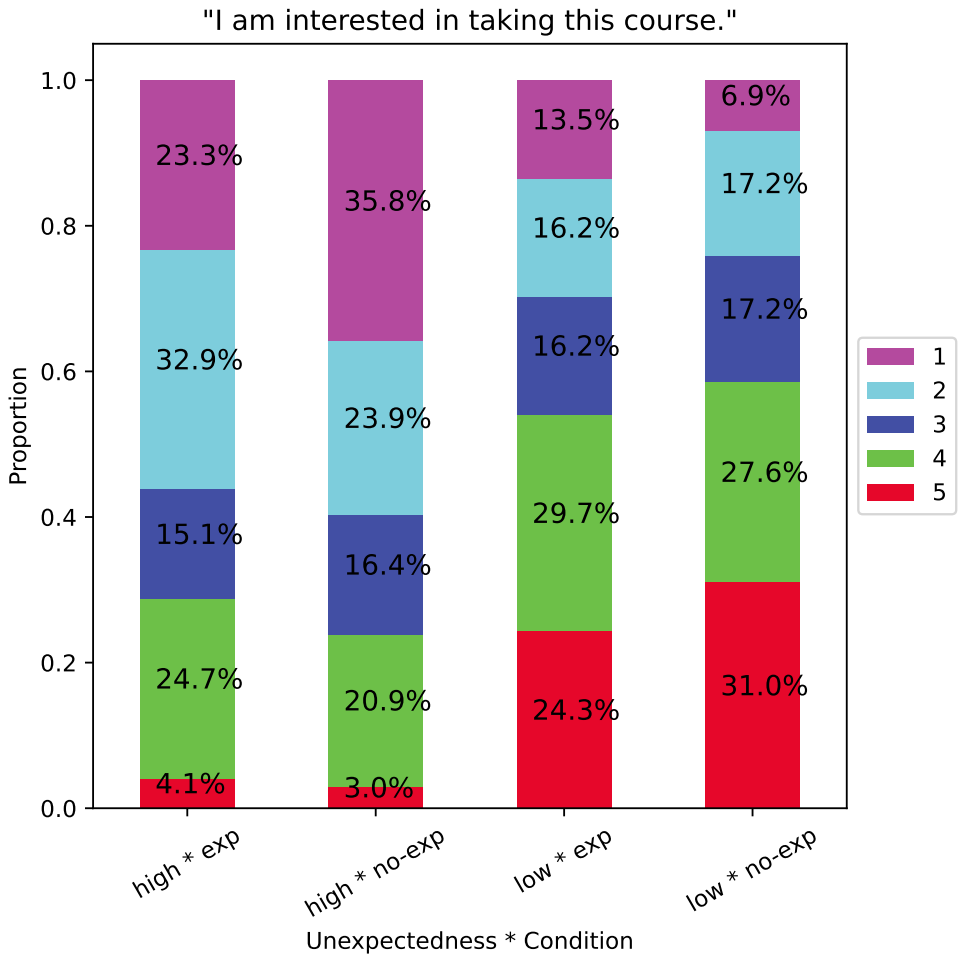


Figure 20: Proportional distribution of user responses to the statement 'I am interested in taking this course.' across different *Unexpectedness* levels and *Explanation* conditions: *High Unexpectedness* with *Explanation* (high * exp), *High Unexpectedness* without *Explanation* (high * no-exp), *Low Unexpectedness* with *Explanation* (low * exp), and *Low Unexpectedness* without *Explanation* (low * no-exp). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'.

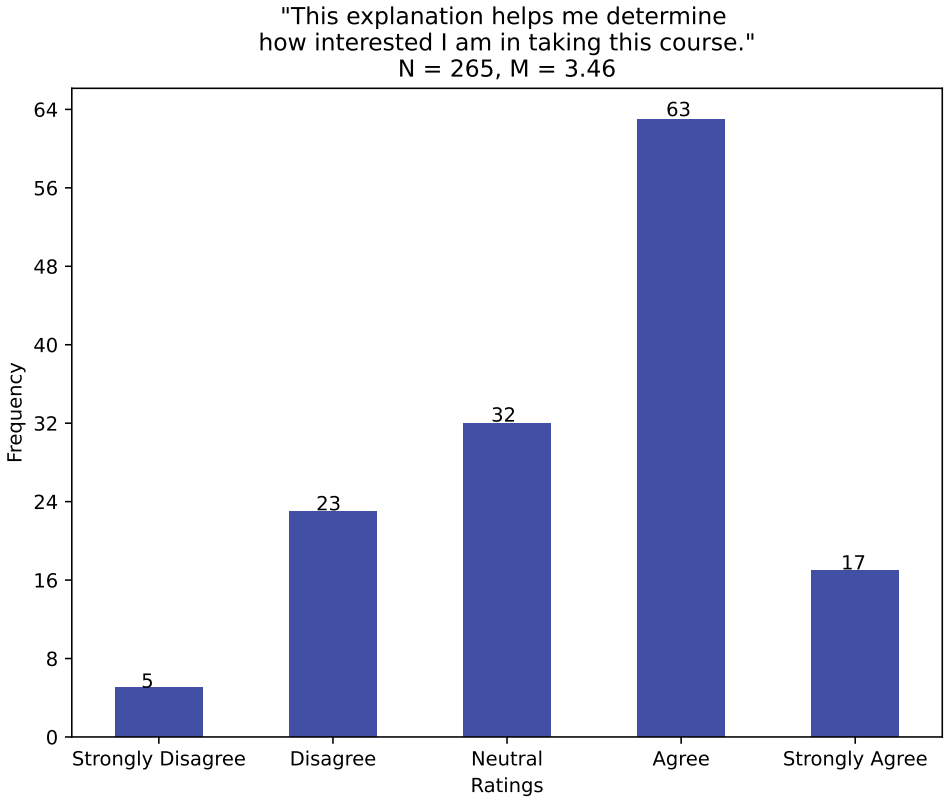


Figure 21: Frequency distribution of user responses to the statement 'This explanation helps me determine how interested I am in taking this course.'.

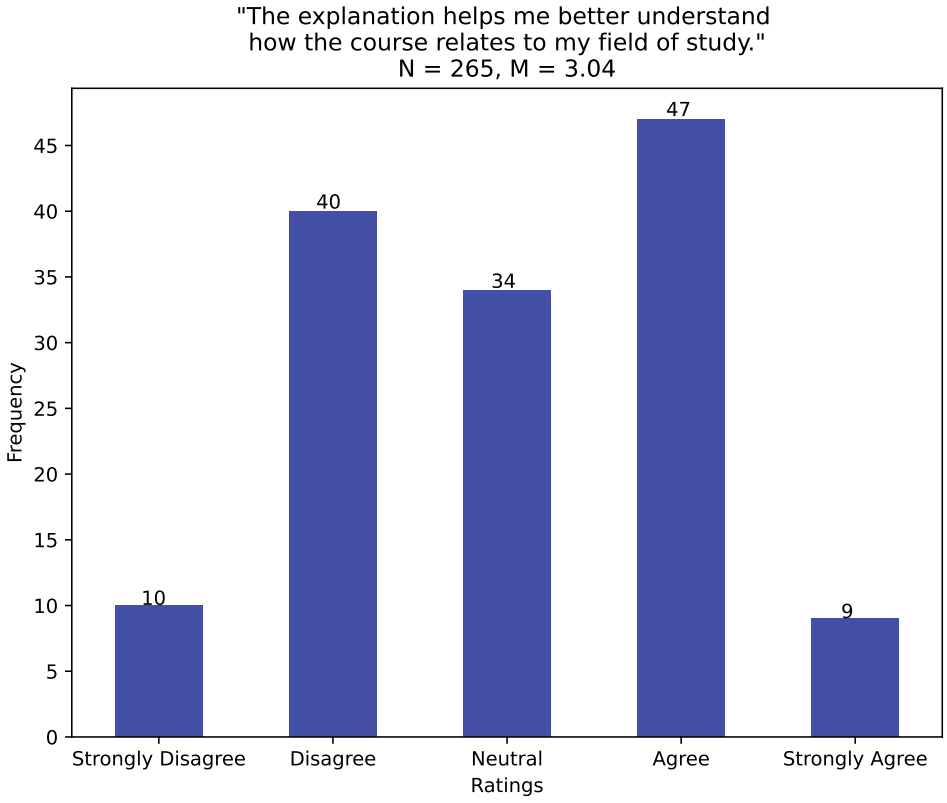


Figure 22: Frequency distribution of user responses to the statement 'The explanation helps me better understand how the course relates to my field of study.'.

Consequently, it is expected that explanations will have varying impacts on students with undeclared majors as opposed to those with declared majors. To achieve this, I first convert the original ratings (on a 5-point Likert scale) to neutral ratings (i.e., 'Yes' for a rating of 3 ('Neutral') and 'No' for all other ratings).

There are 53 participants in our study, coming from diverse academic backgrounds (refer to Appendix B, Table 13 for details). Out of this group, 15 individuals have not yet declared their majors. The distribution of 'Neutral' ratings is displayed in Figure 23, segmented into four distinct categories based on major declaration and the presence of an explanation. Our analysis reveals a compelling additive interaction effect between the declaration of a major and the provision of explanations. Specifically, subjects in the 'Exp' groups consistently exhibit a reduced tendency to provide 'Neutral' opinions. Similarly, individuals in the *declared_major* groups also display a decreased inclination toward 'Neutral' ratings. Notably, those participants who have not yet declared their majors and do not receive explanations exhibit the highest percentage of 'Neutral' ratings (36.7%), significantly higher than those who receive explanations (16.3%).

Our statistical analysis revealed a noteworthy interaction effect between the declaration of a major and the provision of explanations, yielding a p-value of 0.017. Upon further examination of the impact of explanations on participants belonging to declared and undeclared major groups, our findings showed that among participants with a declared major, the presence of explanations did not influence their neutral opinion significantly, as indicated by a p-value of 0.618. In contrast, for participants without a declared major, the absence of explanations was associated with a 0.20-point increase in their neutral opinion, and this difference was statistically significant, with a p-value of 0.0006.

We also examined the 'Neutral' ratings for questions Q1 and Q2, as Q3 was excluded due to its minimal percentage of neutral responses. Notably, participants who have yet to declare their majors and who did not receive explanations showed the highest percentages of 'Neutral' ratings: 40.0% for Q1 and 42.2% for Q2 (refer to Appendix B, Fig. 59 and 60). For Q1, the data revealed that participants without a declared major were 0.20 points more likely to choose a neutral stance in the absence of explanations. This trend approached statistical significance with a p-value of 0.06, possibly affected by a smaller sample size. For

Q2, these participants exhibited a 0.14-point increase in neutrality without explanations, although it was not significant with a p-value of 0.2. Overall, for both questions, the absence of explanations led those without a declared major to be more inclined to rate as 'Neutral.' This was statistically significant with p-values of 0.01 for Q1 and 0.003 for Q2, respectively.

5.5 Summary and Discussion

This chapter presents the design of skill-based explanations within the realm of serendipitous course recommendation systems. The system aims to provide students with comprehensive insights into courses, encompassing their alignment with existing knowledge and the acquisition of novel skills. This, in turn, empowers students to evaluate a course's relevance more effectively and increases their confidence when making choices.

To enhance user comprehension beyond the confines of mere unigrams, which is one of the limitations in my preliminary work discussed in Section 3.2, I have trained a skill extraction model presented in Chapter 4. This model effectively extracts multi-gram skills from the course catalog descriptions, thereby improving the communication of underlying semantics. In a collaborative effort with the CAHL lab at the University of California, Berkeley, we embarked on an exploration of the impact of skill-based explanations on a serendipitous course recommendation system. This exploration was conducted through an online user study at the same institution, using the capabilities of the AskOski system, powered by PLAN-BERT — an advanced deep neural network model well-known for its excellence in top-N course recommendation, enriched with a diversification strategy [21, 38].

Our study stands as one of the pioneering efforts to examine the influence of skill-based explanations on serendipitous course recommendations within higher education. While our overall findings did not show a clear impact of the explanation on the recommendations produced by PLAN-BERT, under the proposed diversification strategy, they did reveal a notable increase in participant interest for courses that exhibited high levels of unexpectedness. This boost amounted to a significant 0.22-point increment in interest when explanations were provided. It is clearly apparent that individuals who received explanatory information dis-

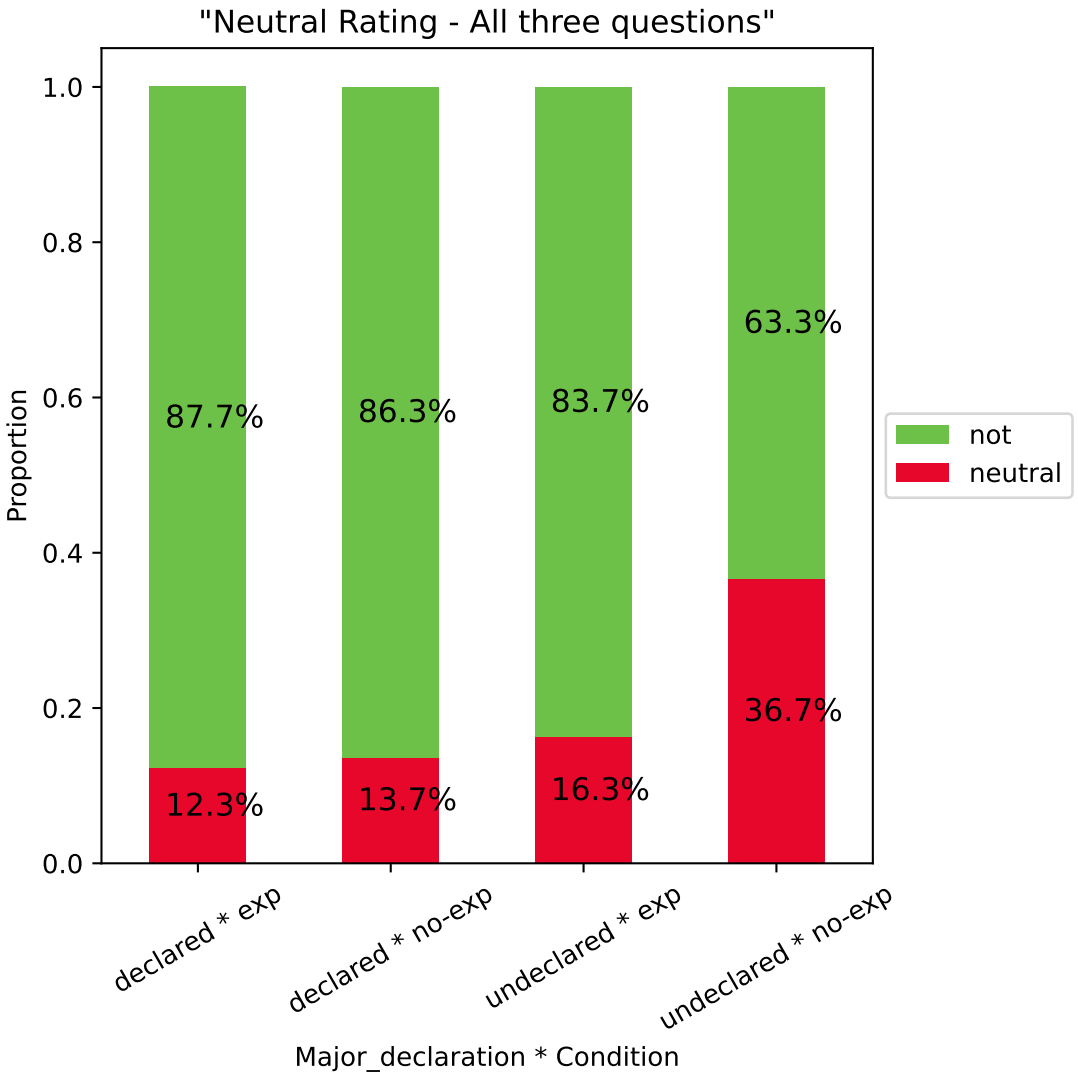


Figure 23: Distribution of 'Neutral' ratings among four groups based on the interactions between major (declared vs. undeclared) and the presence of an explanation (vs. no explanation): declared * exp (N=285), declared * no-exp (N=285), undeclared * exp (N=135), undeclared * no-exp (N=90). The 'Neutral' ratings are aggregated from the responses to the three primary research questions: Q1, Q2, and Q3. The percentage of 'Neutral' ratings is 16.22% (129 'Neutral' ratings of 795).

played a positive attitude towards the usefulness of these explanations in influencing their interest in the recommendations. Importantly, a substantial majority of respondents either 'Agreed' or 'Strongly Agreed' with this assertion.

Furthermore, our research revealed another important aspect: the significant impact of explanations in strengthening users' confidence in their decision-making process. Consequently, this reduced their inclination to provide 'neutral' opinions. A detailed statistical analysis highlighted a compelling interaction between participants' major declaration status and the presence of explanations. Specifically, among participants who did not declare a major, the absence of explanations was associated with a 0.20-point increase in their likelihood to express neutral opinions, and this difference was statistically significant.

In essence, our research underscores the pivotal role that skill-based explanations can play in elevating the user experience within serendipitous course recommendation systems. These insights have profound implications for the design and refinement of such systems in higher education contexts.

Our recommendation and explanation approach has several limitations that require attention in future research. PLAN-BERT has demonstrated its ability to effectively utilize past sequence information, and the incorporation of user and item features has yielded significant benefits. However, it is important to note that our approach to diversifying the recommendation list, applied atop PLAN-BERT's output, is intuitive yet relatively simplistic. In an effort to enhance students' awareness of course options and encourage exploration of courses they may find interesting but which have been relatively unexplored, we limited recommendations to one course per department. However, this constraint had the potential to introduce irrelevant course suggestions, as, in the context of academic course offerings, some departments naturally share stronger connections with related disciplines, while others may have fewer neighboring fields. To address this issue, future studies should consider relaxing the constraint and establishing a certain relevance threshold for course recommendations. If a course from a department fails to meet this threshold, multiple courses from the same department can then be recommended.

Another avenue for enhancing recommendation diversity and optimizing for serendipity is to frame the problem as a multi-task/multi-label optimization problem during the rec-

ommender system's training phase [171, 172]. This approach allows the recommendation engine to simultaneously optimize for both relevance and unexpectedness, striking a balance and effectively constraining them to achieve a unique, non-dominated solution. To implement this, the collection of labels for the unexpectedness of relevant courses is imperative for training the models.

Moreover, while this study did not conclusively demonstrate the impact of providing explanations on the course recommendations, it is still possible that detailed insights into why specific courses are recommended — and how they align with students' abilities and interests — could enable students to better understand the value of these suggestions. Doing so may decrease the likelihood of students overlooking courses simply because they are unfamiliar with them, thus increasing the serendipity factor of recommendations. The efficacy of such explanations might differ based on the student's academic field and progression. Exploring these factors further can enhance the recommendation. Interestingly, our study did find that explanations positively influenced interest in unexpected courses, especially among students who have not yet chosen a major. Future research might prioritize this demographic to amplify the sample size and achieve more robust conclusions.

Furthermore, this between-subjects experiment comprised only 53 participants with diverse academic backgrounds. Within this group, 15 individuals had not yet declared their majors. Consequently, we encountered challenges in effectively controlling for variations in the fields of study when examining the influence of recommendation and explanation strategies. Specifically, we were unable to assess the distinct impacts of these strategies on different student groups, such as those pursuing specific academic disciplines or those with declared majors versus those who were undeclared. To mitigate this limitation and enhance the generalizability of our findings, future research could consider expanding the sample size in between-subjects studies. This step would contribute to more robust and reliable results by allowing for a more comprehensive analysis of the effects of recommendation and explanation strategies across various academic disciplines and student categories.

Finally, in this study, we solely experimented with skills extracted from course catalog descriptions as knowledge components to represent the course for explanation. While this is a conventional approach for content-based methods and has been employed in numerous prior

studies, an alternative approach involves representing courses based on their relationship with individual skills. This approach could be more robust across various skill taxonomies, such as extracted concepts and O*NET DWAs developed by the U.S. Department of Labor to describe the U.S. workforce. Consequently, applying this course modeling approach and experimenting with different skill types to generate explanations for course recommendations represents a potential avenue for future research. In addition, the future of course recommendations isn't just about suggesting what to study but also integrating insights from the job market. By doing this, students can see the real-world applicability of their courses. The skills imparted in higher education need to resonate with market demands. Bridging this gap ensures students are not only educated but also employable.

6.0 CONNECTING HIGHER EDUCATION TO WORKPLACE ACTIVITIES AND EARNINGS

In the previous study, we leveraged student enrollment sequences to identify relevant courses, diversified by department information, and enriched them with concepts extracted from course descriptions for explanation. The future of course recommendations extends beyond academics, incorporating job market insights for students' careers. My goal is to go even further and pursue greater personalization in course recommendations that will be beneficial for students in their future careers. However, do the concepts extracted from course descriptions truly align with the skills demanded in the labor market? In this chapter, I analyze a large novel corpus of over one million syllabi from over eight hundred bachelor's degree-granting U.S. educational institutions to connect the material taught in higher education to the detailed work activities (also referred to as 'skill' discussed in Chapter 1, Section 1.1) in the U.S. economy, as reported by the U.S. Department of Labor. I propose a novel knowledge framework that incorporates the granular workplace activities into course syllabi. Through a comprehensive evaluation involving two predictive tasks, this framework has proven to be highly effective in extracting essential features for accurate predictions. This research project was undertaken in partnership with Dr. Sarah Bana and Dr. Morgan Frank. In Chapter 7, I will demonstrate how this course modeling method can be applied to build explainable course recommendation systems that focus on skills and careers, aiming to help students discover courses that align with their career goals and equip them with the necessary knowledge and skills.

6.1 Introduction

Education plays a critical role in economic growth and social progress. College degrees are generally associated with higher potential lifetime earnings, larger professional networks, and more adaptable careers [1, 3]. Higher education is a major part of US workforce de-

velopment but information on the skills and expertise taught during higher education remain absent—even as recent research highlights the critical role of skills in shaping labor trends [173, 174, 28]. However, most empirical work relies on coarse labor distinctions, such as college major and institutional information (e.g., school brands), to explain these occupational trends [175, 176, 177, 178]. While useful, these coarse educational and labor categories may hide further insights into the skills of "high-skilled" workers that contribute to positive career outcomes [179].

Many workers acquire skills through higher education that shape their careers. Studies have shown that social-cognitive skills and sensory-physical skills are correlated to high- and low-wage occupations, respectively, and that skill polarization divides workers with and without higher education [26]. Discrepancies between skills demanded, taught, and researched have been identified by applying textual matching techniques to job advertisements, course syllabi, and research publications in Computer Science [24]. These analyses of skills reveal gaps between the workforce and educational/training systems. Understanding the sources of these gaps, across all fields of study, may improve curriculum design, inform educational policy, and improve student outcomes when they enter the workforce.

In this work, I analyze the recently available Open Syllabus Project (OSP) dataset, which contains over 1.4 million course syllabi from more than 3,000 US colleges and universities from 2008 to 2017. While relatively new, this data source has proven useful for modeling higher education. For example, one study quantified the skill (mis-)alignment between academic research, industry, and educational offerings in data science and data engineering [24]. They used Burning Glass (BG) skill taxonomy and applied matching techniques to extract skills appearing in job titles and descriptions, course syllabi, and publication titles and abstracts. Another study proposed a new measure for the "education-innovation gap" using the textual similarity between course syllabi and academic journals to model the dissemination of frontier knowledge into college classrooms while relating these dynamics to students' graduation rates and incomes [180].

This work is the first attempt to connect workplace activities to higher education through course syllabi; here, I use the granular workplace activities designed and produced by the U.S. Department of Labor (i.e., O*NET Detailed Work Activity (DWA) taxonomy described

in Section Materials) to explain the underlying knowledge structures across college majors (*i.e.*, fields of study (FOS)) and among US universities. I use word embeddings to represent textual documents [181, 36], and explore different distance metrics to measure the similarity of two embedded skill vectors. Consequently, I am able to apply agglomerative hierarchical clustering techniques to the DWA-based vector representations of FOS and universities to discover their clusters. Hierarchical clustering [182] produces a nested sequence of clusters, and the hierarchy of clusters enables me to explore clusters at any level of detail without the need of identifying a specific number of topics as would be the case with K-means clustering techniques. Motivated by the principle of relatedness [183], I model the relationships between pairs of skills across academia to forecast how skills change over time. Based on the out-of-sample earnings prediction evaluation with *5-fold cross validation*, I also discover that differences in acquired skills help to explain the variance of graduates' earnings. The results offer an approach that connects college education to future careers. These insights may enable educational policy and academic programs to adapt to the skill dynamics in the labor market. For example, information systems that bridge between higher education and workforce skill data may inform updates to course design that prepare students with the necessary skills for their desired careers.

In summary, this study attempts to answer the following research questions:

- Q1. Can the granular workplace activities used by the Department of Labor to describe the US workforce also distinguish between different college majors and institutions?
- Q2. How do the DWAs taught in a curriculum or field of study evolve over time? Can the relationships between pairs of skills across all of academia help to predict skill evolution?
- Q3. Do the differences in taught skills during higher education predict graduates' earnings? Similarly, do differences in taught skills within college majors correspond to earnings differences of recent graduates?

In the next section, I describe multiple datasets that enable me to answer the aforementioned research questions. I then describe my methodology in detail, present my analysis and discuss its implications and potential weaknesses to conclude the chapter.

6.2 Materials

Open Syllabus Project Dataset[1] is one of the largest corpora of syllabi in the world. As of October of 2019, it contains over eight million syllabi, collected from 5,381 colleges and universities, including over three million syllabi taught at 3,186 US institutions. OSP's fields-of-study classifier draws heavily from the Classification of Instructional Programs (CIP) taxonomy used by the National Center for Education Statistics to determine the academic field of study (*e.g.*, *Economics, Business, Computer Science*) best associated with each syllabus. It includes 62 fields of study. Each syllabus has a unique identifier and the text assignment data including a description of its content, a list of references and recommended readings, and course requirements (such as assignments and exams). Syllabi can be directly mapped to graduation and enrollment statistics from the US Department of Education's Integrated Postsecondary Education Data System (IPEDS). Syllabi are annotated with metadata including the institution, department, and academic year associated with the course. I extract and concatenate course titles, course descriptions and learning objectives from syllabi's textual data to create "course descriptions." More details can be found in Appendix A, Section A.1. I limit the data from 2008 and 2017 (the ten most recent years in OSP), resulting in roughly 1.4 million syllabi representing college courses from 1,481 institutions. More about courses statistics per year and/or per field of study (FOS) can be found in Appendix A, Fig. 57 and 58.

O*NET Detailed Work Activity (DWA) Taxonomy[2]. O*NET is designed and produced by the U.S. Department of Labor/Employment and Training Administration. The O*NET database allows snapshots of the relationships between occupations and skills. It has 2070 DWAs (*e.g.*, *"develop methods of social or economic research."*, *"design integrated computer systems."*, *"design public or employee health programs."*) representing specific work activities performed across a small to moderate number of occupations within a job family. For example, the occupations with related activities to DWA *"design public or employee health programs."* include "Preventive Medicine Physicians", "Occupational Health

[1] https://opensyllabus.org (OSP)
[2] https://www.onetonline.org/help/online/dwa

and Safety Specialists", "Occupational Health and Safety Technicians", "Dietitians and Nutritionists", and "Dentists, General".

Integrated Postsecondary Education Data System[3] **(IPEDS)** is the core postsecondary education data collection program of the U.S. Department of Education's National Center For Education Statistics (NCES). It annually collects information from all providers of postsecondary education, including public institutions, private nonprofit institutions, and private for-profit institutions, in fundamental areas such as enrollment, program completion and graduation rates. Providing data is required for any institution that applies for or participates in any Federal financial assistance program. IPEDS also includes a wide range of information about institution and institution groups, such as Degree-granting status, Institutional category, and Carnegie classifications. The Carnegie Classification, or more formally, the Carnegie Classification of Institutions of Higher Education,[4] is a framework for categorizing all accredited, degree-granting institutions in the United States. It is designed to group colleges and universities based on their research activities.

College Scorecard[5] is a U.S. Department of Education data initiative providing transparency and consumer information related to individual institutions of higher education and individual fields of study (*e.g.*, majors) within those institutions. College Scorecard provides information about post-college earnings including median earnings of graduates working and not enrolled after completing the highest credential in their first and second years for the two graduation cohorts of years 2016 and 2017. I only use the first year earnings of graduates. I process the data for Baccalaureate colleges and universities, and create the mapping between College Scorecard CIP code and OSP CIP code (the mapping can be found in this GitHub folder[6]). As a result, I obtain 9007 earnings records for 832 institutions in 54 fields-of-study.

[3] https://nces.ed.gov/ipeds/
[4] https://carnegieclassifications.iu.edu/
[5] https://data.ed.gov/
[6] https://github.com/HungChau/OSP-connect-higher-education/tree/main/cip_code_mapping

6.3 Methods and Results

6.3.1 Modeling course syllabi with workplace skills

Are the workplace activities tracked by the US Department of Labor robust and effective in describing the knowledge in higher education? The O*NET database is produced by the US Bureau of Labor Statistics and details the labor market trends of workplace skills and activities by occupation. Specifically, detailed work activities (DWAs) are elements in the O*NET database that provide information about occupations' labor requirements. This data has been used to analyze several labor market dynamics including job polarization [26, 184] and the economic resilience of cities [173, 185]. Although O*NET relates occupations to skills in the workforce, similar data is not reported for educational programs even though many high-skilled workers obtain skills in college before entering the workforce.

I bridge this gap by detecting O*NET's detailed work activities from syllabus course descriptions. Each syllabus in the OSP data contains a description of the course content, a list of references and recommended readings, and course requirements, such as assignments and exams. Given a syllabus, I extract the course's title, description, and learning objectives from the text and concatenate them to form the *course descriptions* (details are in Appendix A, Section A.1.1). I apply word embeddings [37] and document similarity techniques from natural language processing to represent each DWA and syllabus as continuous vectors distributed in the same pre-trained language embedding space. Language embedding models enable me to describe the semantic similarity between two textual documents or sentences; here, I compare syllabus course descriptions to DWAs. I choose pre-trained *fastText* word embeddings from [186], which is constructed from all Wikipedia pages in 2017, the UMBC webbase corpus, and the statmt.org news data. I choose these word embeddings because the semantic diversity of Wikipedia and news articles should capture the semantic diversity of topics taught across FOS. This model has been used in several applications [187, 188, 189], and achieves better performance than simple bag-of-words and TF-IDF [36]. I compute the *relationship* ($0 <= r_s(dwa) <= 1$) between a syllabus s and a DWA by comparing their word embedding vector representations with soft cosine measure [190] (details are in Appendix

A, Section A.1.2). As a result, syllabi are represented based on their relationships with the DWAs (called the DWA-based syllabus representation). I provide an example of the most and least prevalent DWAs detected for a political science syllabus at Harvard University in 2013 (see Figure 24A).

In addition to course descriptions, syllabi are annotated with metadata about where and when the course was taught. Metadata includes the institution, department/major/FOS, and academic year. OSP's field classifier is trained and tested on the IPEDS 2010 CIP taxonomy to determine the academic field (*i.e.*, FOS) best associated with each syllabus. This enables me to calculate the relationship between each pair of DWAs based on the co-occurrence of dwa_1 and dwa_2 in any set of course syllabi S; for example, the set of all syllabi within a given FOS, $sim_f(dwa_1, dwa_2)$ for $f \in FOS$, or across all of academia, $sim(dwa_1, dwa_2)$. I experiment with various semantic distance metrics to compute DWA relationships through syllabi including Jaccard similarity, Cosine similarity, Euclidean distance, and Manhattan distance (see Appendix A, Section A.2). I find Jaccard similarity to be the most predictive and I present those results in the main text. It is worth noting that relationships between two DWAs can be directly computed by measuring the cosine similarity of their embedding vectors. However, this approach measuring a static relationship between DWAs fails to distinguish the dynamics of how one DWA relates to another locally (i.e., within a FOS or a university) and globally (i.e., across all of academia) over time, which will be discussed in Section 6.3.3. For example, social skills and computer programming skills may be semantically different but co-taught as complementary skills across syllabi (e.g., computational social science, social network analysis, or econometrics).

The syllabus-DWA relationships ($r_s(dwa)$) also enable me to model a FOS f and a university u in terms of their relationship to each of the DWAs according to, respectively,

$$r_f(dwa) = \frac{1}{|S_f|} \sum_{s \in S_f} r_s(dwa) \quad \text{and} \quad r_u(dwa) = \frac{\sum_{f \in FOS} \sum_{s \in S_{f,u}} \alpha_{f,u} \cdot r_s(dwa)}{\sum_{f \in FOS} \alpha_{f,u} \cdot |S_{f,u}|}. \tag{12}$$

These relevance scores are a measure of how strongly the skill (i.e., *dwa*) is represented in a field or university. While $r_f(dwa)$ (the relevance score of the *dwa* to FOS f) is the average

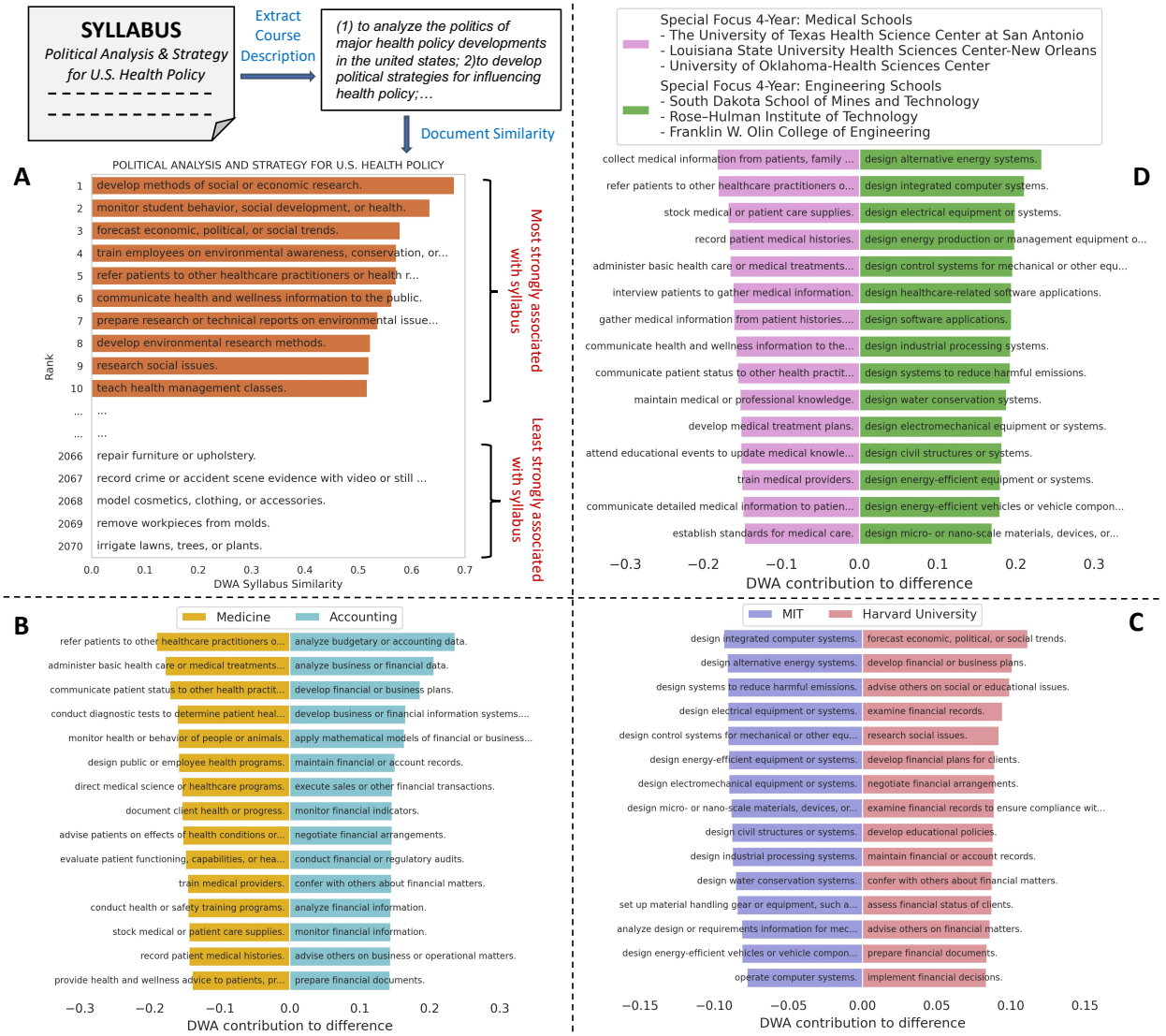


Figure 24: The work activities inferred syllabi reveal key differences among universities and fields of study. (A) An example political science syllabus from Harvard University and the activities that are most and least strongly associated with its course description. DWA-syllabus similarity scores range from *0* (not detected) to *1* (strongly detected). (B) The DWAs that most significantly distinguish Accounting syllabi from Medicine syllabi. (C) The DWAs that most strongly separate MIT syllabi from Harvard syllabi. (D) The DWAs that most strongly separate Special Focus 4-Year Medical Schools syllabi from Engineering Schools syllabi. More examples can be found in Appendix A, Figures 46, 47, 48, & 49.

over the similarity scores of that DWA across $s \in S_f$, $r_u(dwa)$ (the relevance score of the *dwa* to university u) is the mean similarity score of that DWA across syllabi weighted by the estimated graduation rates ($\alpha_{f,u}$) of the syllabus's field of study at that university. In the absence of course enrollment data, I use graduation rates for each FOS at each university to approximate the number of students who learn from each syllabus. S_f represents all of the syllabi within a given FOS f, and $S_{f,u}$ represents all of the syllabi within a given FOS at a university u.

These tools enable me to compare pairs of syllabi, FOS, or universities based on their most common DWAs. I publish the DWA similarities by different metrics, DWA scores for each FOS and for each university by year from 2008 to 2017 in a Github repository.[7] Specifically, I compare entities of the same type (*e.g.*, one FOS to another) by subtracting its DWA vector representation from the other's and rank the resulting vector in descending order. I visualize the top 15 DWAs of each entity that contribute most to the difference of the pair in Figures 24B, 24C & 24D. For example, the DWAs "refer patients to other healthcare practitioners or health resources" and "administer basic health care or medical treatments" most strongly distinguish Medicine from Accounting, while "analyze budgetary or accounting data" and "analyze business or financial data" identify Accounting from Medicine (see Fig. 24B). Similarly, I compare pairs of universities based on their taught DWAs. As an example, "design integrated computer systems" and "design alternative energy systems" most strongly distinguish Massachusetts Institute of Technology (MIT) from Harvard University, while "forecast economic, political, or social trends" and "develop financial or business plans" more strongly identify Harvard from MIT (see Fig. 24C). These results match my intuition as MIT is the world-leading engineering university and Harvard is in the top ten universities in each social science area according to U.S. News rankings. Building on this, I can group universities based on their Carnegie classification to identify the major differences in taught DWAs. I compare Medical Schools to Engineering Schools in Fig. 24D. More examples can be found in Appendix A, Figures 46, 47, 48, & 49.

[7] https://github.com/HungChau/OSP-connect-higher-education

6.3.2 Identifying Field-of-Study and university clusters

Do DWAs capture the focal knowledge offered by an academic field or a university? To further compare education among FOS, I use agglomerative hierarchical clustering on DWA-based vector representations of each FOS. Hierarchical clustering [182] produces a nested sequence of clusters like a tree (also called a dendrogram). Agglomerative clustering builds the dendrogram from the bottom level, and merges the most similar (or nearest) pair of clusters at each level to go one level up. Hierarchical clustering can take any form of distance or similarity function, and the hierarchy of clusters enables me to explore clusters at any level of detail without the need of picking a number of topics k as would be the case with K-means clustering. Pairs of FOS are similar if they are associated with similar types of work activities. For instance, *Accounting* is clustered together with *Business* and *Marketing*; *Medicine* is clustered together with *Nursing*, *Nutrition*, *Health Technician*, *Dentistry* and *Veterinary Medicine*; the STEM cluster includes *Mathematics*, *Physics*, *Astronomy*, *Biology*, *Earth Sciences*, *Atmospheric Sciences* and *Chemistry*; and the Social Science cluster includes *Social Work*, *Political Science*, *History*, *Sociology*, *Women Studies*, *Anthropology* and *Religion* (see Fig. 25).

Similarly, I compare all US universities in my data set using agglomerative hierarchical clustering performed on the *weighted* DWA-based vector representation of each institution in Figure 26. I see that similar universities are clustered together. For example, *The University of Texas Medical Branch*, *The University of Texas Health Science Center*, and *Oregon Health and Science University* are clustered together. Although our dataset contains a large number of universities, I select a subset of Ivy Plus universities and universities from various IPEDS Carnegie Classifications to visualize in Figure 26. I filter out universities that have less than 100 syllabi or were missing syllabi in any year from 2008 to 2017. Carnegie classifications are mostly recovered by the clusters (see colors in Fig. 26). Additionally, engineering schools like *California Institute of Technology*, *Massachusetts Institute of Technology*, and *Carnegie Mellon University*, are clustered together. Similarly, liberal arts schools including *Cornell University*, *Harvard University*, and *University of Pennsylvania* are clustered together.

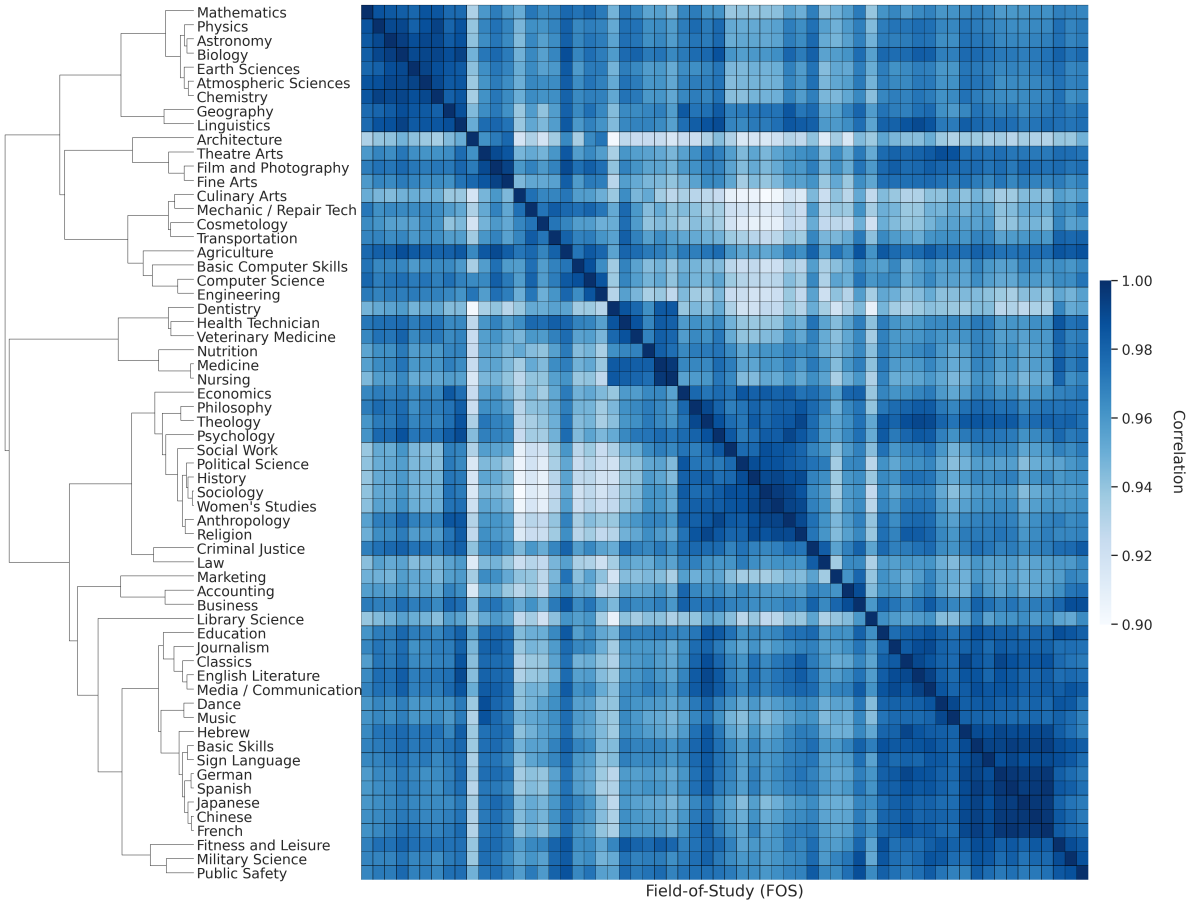


Figure 25: The similarity of FOS based on the prevalence of DWAs in syllabi from within those fields. The dendrogram and heatmap show similar FOS clustered together based on their DWA-vector representations.

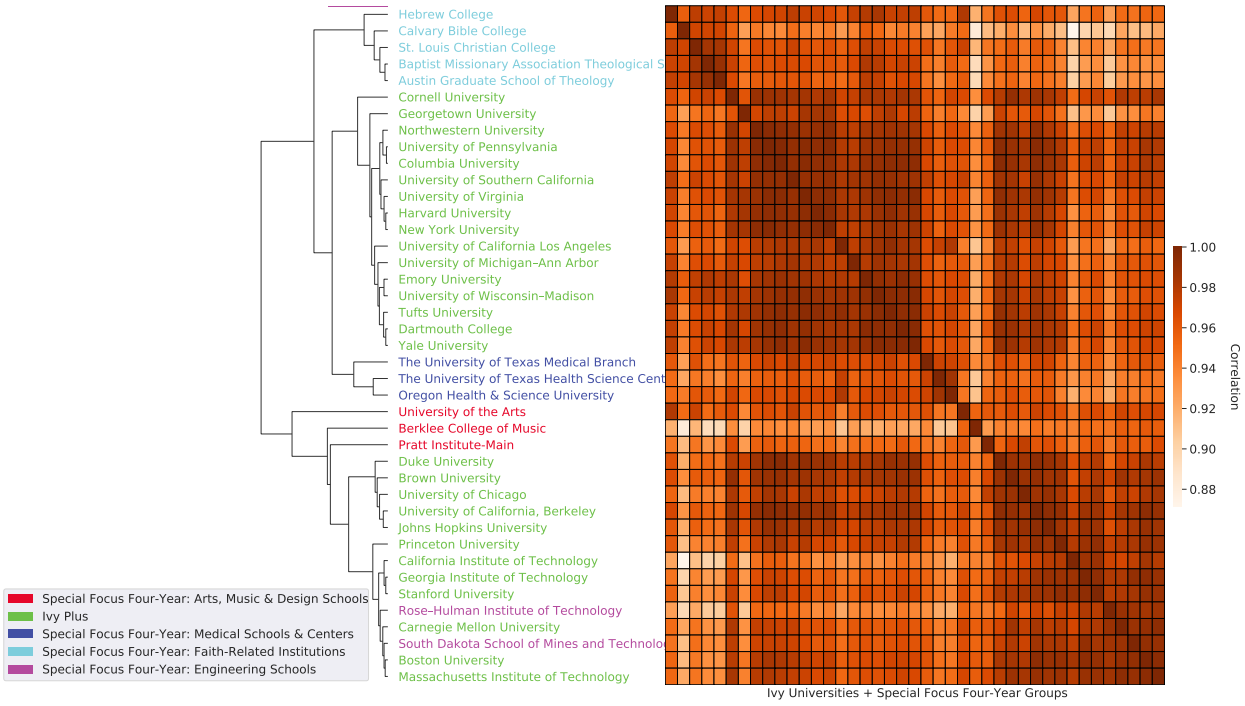


Figure 26: The similarity of universities based on the graduation-weighted prevalence of DWAs offered in their course syllabi. The dendrogram and heatmap reveal the hierarchical clustering of the Ivy Plus group and Special Focus Four-Year groups from the Carnegie Classification 2018 based on DWA vector representations.

6.3.3 Predicting the change in taught skills

How do the DWAs taught in a field of study evolve over time? In particular, which new skills or topics will emerge in a field's syllabi? Forecasting these educational trends enables proactive course design by educators and could inform educational incentives from policymakers. Here, I use the principle of relatedness [183] to hypothesize that DWAs that occur together across all of higher education are more likely to be co-taught within a given FOS in the future. If correct, then modeling the relationships between pairs of DWAs across all of academia should forecast the introduction of new topics within a FOS even if that topic has not been part of that FOS historically. As an illustrative example, although largely absent from Economics syllabi today, machine learning may become more common in Economics because Economics already teaches linear regression which is commonly taught as an example of machine learning in Computer Science courses. As a more specific example from my data, DWAs that relate to machine learning, such as "analyze website or related online data to track trends or usage" may become more prevalent in Economics syllabi moving forward (e.g., in studies of online job postings [24, 191]).

I test our hypothesis using OSP data to predict which DWAs become important in a FOS (f). I use the relevance scores ($r_f(dwa)$) calculated from the syllabi of each FOS in two different years (i.e., 2008 and 2017). I recast this problem as predicting the score difference (Δr) of a DWA between the two years:

$$\Delta r_{dwa,f} = r_f^{2017}(dwa) - r_f^{2008}(dwa) \tag{13}$$

I also perform classification analysis for predicting DWAs becoming important in future, which can be found in Appendix A, Section A.3.2. I run several ordinary least squares (OLS) regressions to predict $\Delta r_{dwa,f}$ using the relevance scores of the dwa to FOS f ($r_f(dwa)$) and various models of inter-DWA relationships (described in Section Modeling course syllabi with workplace skills). As a baseline, I first consider Model 1 using only the current relevance scores of DWA within each FOS with FOS fixed effects (denoted λ_f) according to

$$\Delta r_{dwa,f} = \beta_0 + \beta_1 r_f^{2008}(dwa) + \lambda_f. \tag{14}$$

Next, I additionally include a variable representing the co-occurrence of DWAs across syllabi within a FOS (denoted R_f) to create Model 2

$$\Delta r_{dwa,f} = \beta_0 + \beta_1 r_f^{2008}(dwa) + \underbrace{\beta_2 \left(\frac{\sum_{dwa' \in DWA} sim_f(dwa, dwa') r_f^{2008}(dwa')}{|DWA|} \right)}_{R_f} + \lambda_f \tag{15}$$

and yet another similar Model 3 using DWA pair co-occurrences across syllabi from every FOS (denoted R)

$$\Delta r_{dwa,f} = \beta_0 + \beta_1 r_f^{2008}(dwa) + \underbrace{\beta_2 \left(\frac{\sum_{dwa' \in DWA} sim(dwa, dwa') r_f^{2008}(dwa')}{|DWA|} \right)}_{R} + \lambda_f. \tag{16}$$

Model 4 includes an interaction term between DWA's relevance score within a FOS (*i.e.*, R_f) and DWA pair co-occurrences within that FOS according to

$$\Delta r_{dwa,f} = \beta_0 + \beta_1 r_f^{2008}(dwa) + \beta_2 R_f + \beta_3 \left(r_f^{2008}(dwa) * R_f \right) + \lambda_f \tag{17}$$

and, in Model 5, using DWA pair co-occurrence across all FOS

$$\Delta r_{dwa,f} = \beta_0 + \beta_1 r_f^{2008}(dwa) + \beta_2 R + \beta_3 \left(r_f^{2008}(dwa) \cdot R \right) + \lambda_f \tag{18}$$

As robustness checks, I run Models 2, 3, 4 & 5 with the two different methods and four distance metrics aforementioned in Section Modeling course syllabi with workplace skills for computing the DWA relationships. Although I could compare DWA pairs based solely on their semantic similarity using their word embedding vectors, this approach would miss DWA pairs that capture complementary topics. For example, Models 2 and 3 would be identical to Models 4 and 5, respectively. The results (see Appendix A, Section A.3.1) show that modeling DWA relationships based on their co-occurrence in syllabi with Jaccard similarity

yields the best performances across all the models involving inter-DWA relationships. I discuss these results in the chapter.

I compare model performance using root mean squared error (RMSE) with 5-fold cross-validation in Figure 27 (R-squared metric is reported in Appendix A, Figure 56A). First, including variables representing DWA relationships decreases RMSE (*i.e.*, Model 2 (R^2 = 0.231) & Model 3 (R^2 = 0.239) are statistically significantly better than Model 1 (R^2 = 0.191)). Second, measuring DWA co-occurrences across all of academia (i.e., using R) instead of only within a single FOS (i.e., using R_f) improves model predictions. Specifically, Model 3 (R^2 = 0.239) outperforms Model 2 (R^2 = 0.231) and Model 5 (R^2 = 0.244) outperforms Model 4 (R^2 = 0.231).

These results suggest that FOS educational trends within a FOS correspond to global educational trends across all of academia. In particular, this evidence supports our hypothesis that DWAs tend to be co-taught more within a given FOS if they are bundled together across all of higher education (*e.g.*, Computer Science may increasingly teach "analyze green technology design requirements" since it is commonly taught with "identify information technology project resource requirements" in other FOS including Engineering). Although Model 4 does not outperform Model 2, including the interactions between current DWA relevance scores and the average of the proximity of *global* DWA relationships does yield a significant improvement (*i.e.*, Model 5 outperforms Model 3). In conclusion, the best performing model is Model 5 which leverages the information about the current score of the DWA, their relationships with other DWAs across academia, and the interaction of these two variables. Model 5 improves 3.3 percent (27.5 percent) in terms of RMSE (R-squared) over Model 1, which only uses the *2008* DWA relevance scores. Therefore, I train Model 5 using the entire data, and use it to predict the relevance scores of DWAs in a FOS nine years later. Table 9 shows some examples of DWAs that became important within a FOS —in terms of ranking DWAs—in nine years. The full list of DWAs that are predicted to increase their ranks by at least five units and ranked in the top 50 in 9 years can be found in the aforementioned Github repository.

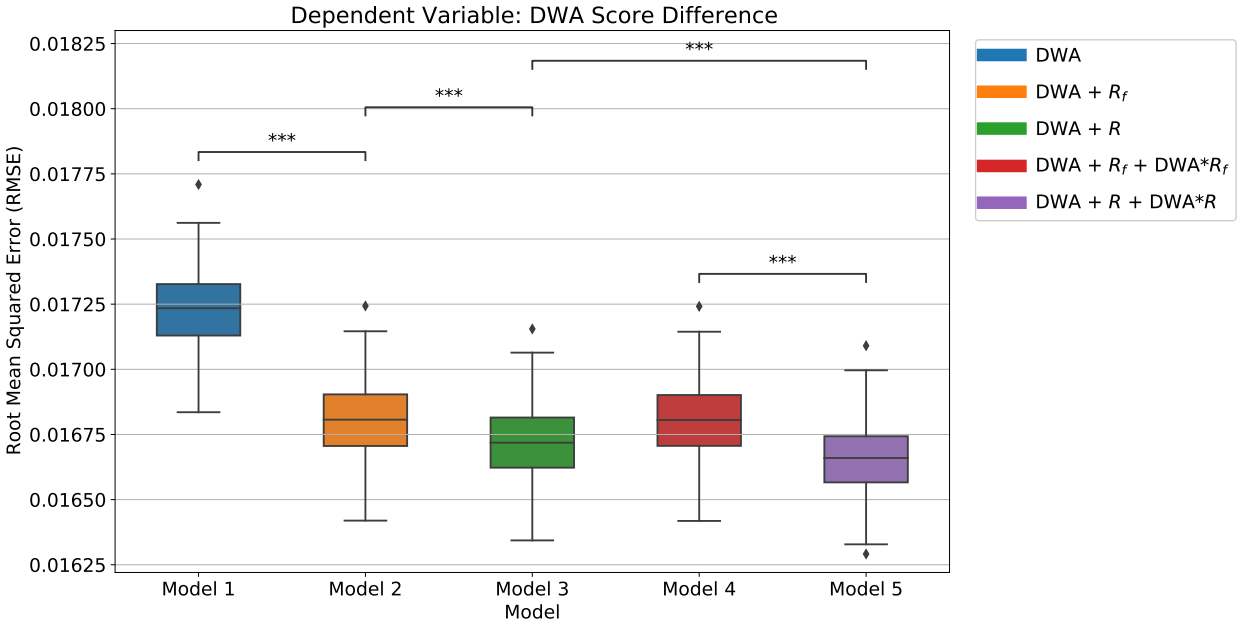


Figure 27: **Workplace activities detected from syllabi predicting teaching dynamics within a field of study**. I perform 5-fold cross-validation and repeat 40 times (i.e., 200 trials in total) for each model and measure RMSE by the resulting model applied to the test set. Asterisks indicate the statistically significant difference between two models' performances with Bonferroni correction. Predicting the importance of DWAs changing in nine years (2008 vs. 2017). As a baseline, model 1 only considers the current DWA score and FOS fixed effects. The other models consider the relationships between DWAs, and how they interact with each other to predict how they may change in future.

Table 9: Examples of DWAs that are predicted to increase their ranks in 9 years in particular fields. I only select DWAs that are ranked in the top 50 in future. The full list of predicted DWAs can be found in the same GitHub folder.

Field-of-Study	Detailed Work Activity	Rank (2017)	Rank (2026)
Computer Science	analyze green technology design requirements.	40	33
	apply information technology to solve business or other applied problems.	46	40
Economics	evaluate plans or specifications to determine technological or environmental implications.	37	27
	develop marketing plans or strategies for environmental initiatives.	58	50
Journalism	gather information about work conditions or locations.	37	24
	prepare scientific or technical reports or presentations.	48	42
Medicine	develop healthcare quality and safety procedures.	28	23
	operate laboratory equipment to analyze medical samples.	65	50
Physics	develop procedures for data entry or processing.	43	33
	develop performance metrics or standards related to information technology.	41	34

6.3.4 Predicting graduate earnings

Do detected DWAs predict the variation in graduates' earnings? Most—if not all—educational programs aim to provide students with the skills and abilities to successfully enter the workforce (*e.g.*, to gain employment and maximize earnings). Most empirical work relies on coarse labor distinctions such as college major and institutional information (e.g., school brands) to correlate to graduate earnings [176, 192, 193, 178], but none have provided insights into the skills students learn that could contribute to their future earnings. My analysis of DWAs in university course syllabi provides the first data set connecting taught skills to students' earnings after graduation. I collect earnings of graduates from the College Scorecard earnings data from the U.S. Department of Education. Though large, the OSP course syllabus data is not distributed evenly across fields-of-study and institutions. Some fields and institutions have much fewer course syllabi. Thus, to sufficiently estimate work activities taught in a FOS at a university, I limit earnings records for FOS (in an institute) that have at least 10 course syllabi; and perform the Kolmogorov-Smirnov statistical test to make sure the remaining earnings records representative for the entire population of the

field at the institute (more details on the selection process and criteria are in Appendix A, Section A.4). I build several OLS regression models to predict *average* graduate earnings across FOS (f) at a university (u) based on the relevance scores of the DWAs across fields (*DWA*) and within field (*FOS*DWA*), FOS fixed effects (*FOS*), school brands (i.e., school ranks[8] if available) fixed effects (RANK), and geography fix effects (*GEO*). Due to the limited availability of earnings data, I use groups of 10 ranks (i.e., 1-10, 10-20) for national universities and 15 ranks (i.e., 1-15, 15-30) for liberal arts colleges. For geographical features, I group universities together based on their divisions[9] (e.g., New England Division, West North Central Division). These groups are represented using indicator variables in the regression analyses.

To avoid model over-fitting, I perform 5-fold cross-validation and LASSO feature selection on the models that include DWA features. LASSO [194] is one of the most popular methods for feature selection; it minimizes the residual sum of squares subject to the sum of the absolute value of coefficients being less than a constant. This constraint tends to "regularize" large models by producing some 0 coefficients when variables are co-linear. In other words, the penalty factor determines how many features are retained; using cross-validation to choose the penalty factor helps ensure that the model will generalize well to future data samples. As a result, I find that DWAs improve predictions of graduate incomes (see Fig. 28 for *RMSE* metric and Appendix A, Figure 56B for *R-squared* metric according to 5-fold cross-validation). Including DWAs improves predictions of earnings compared to FOS fixed effects (*i.e.*, smaller RMSE). Also, $R^2 = 0.684$ of the *DWA* model is significantly better than that of *FOS* model $R^2 = 0.677$). Controlling for university rankings and geography further improves the *FOS* model (i.e., *FOS+RANK+GEO* ($R^2 = 0.757$) model is significantly better than *FOS* ($R^2 = 0.677$) model). But combining DWA variables with RANK and GEO variables and FOS fixed effects yields even further improvement (*FOS+RANK+GEO+DWA* model ($R^2 = 0.761$) is statistically significantly better than that of *FOS+RANK+GEO* model). This evidence suggests that some of the information about

[8]Historical U.S. News and World report rankings are compiled by Andy Reiter and available at https://andyreiter.com/datasets/

[9]U.S. Geographic Levels are available at https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html

graduate earnings represented in university rankings is also encoded in the DWA variables (e.g., a LASSO regression model containing DWA variables accounts for 48% of the variation in college rankings; year and FOS fixed effects account for 7.9%). Finally, the best model (*FOS+RANK+FOS*DWA*) is found when I allow DWA variables to interact with FOS fixed effects which suggests that different DWAs correspond to earnings variation in different FOS ($R^2 = 0.779$). The geographic variables also help to improve the best model's performance but are not significant ($R^2 = 0.782$).

6.3.5 Within Field-of-Study skill variation and the earnings of recent college graduates

Do differences in taught skills within college majors correspond to earnings differences of recent graduates? To study how DWAs relate to the earnings of graduates of a specific field of study, I perform separate regression analyses for each FOS with at least 100 institution-year observations. I employ LASSO feature selection for DWAs and report model performance using 40 independent trials of 5-fold cross-validation to mitigate over-fitting. The remaining DWAs are used to predict earnings. As can be seen from Figure 29, the *DWA+GEO* models perform significantly better than the baseline *GEO* models in terms of RMSE. Due to the limited earnings data within FOS to perform cross-validation, the school ranking is omitted; the baseline models only include geographic variables (*GEO*). I obtain similar performance when alternatively using the model variance explained (R^2) (see Appendix A, Figure 56C). This result again shows that the DWAs complement the FOS information by increasing the share of the earnings explained by the model and improving the model's predictions. However, *DWA+GEO* model performance varies across FOS. For example, the *DWA+GEO* model improves 27.2% RMSE over the *GEO* model for *Business* compared to a more modest improvement of 4.2% for *Psychology*. Although O*NET DWAs improve predictions in general, this varied performance across FOS could be because DWAs represent key skills and activities better in some FOS than in others. Nevertheless, our methodology shows that using granular workplace skills helps to identify important features contributing to the earnings of graduates beyond course educational and labor categories.

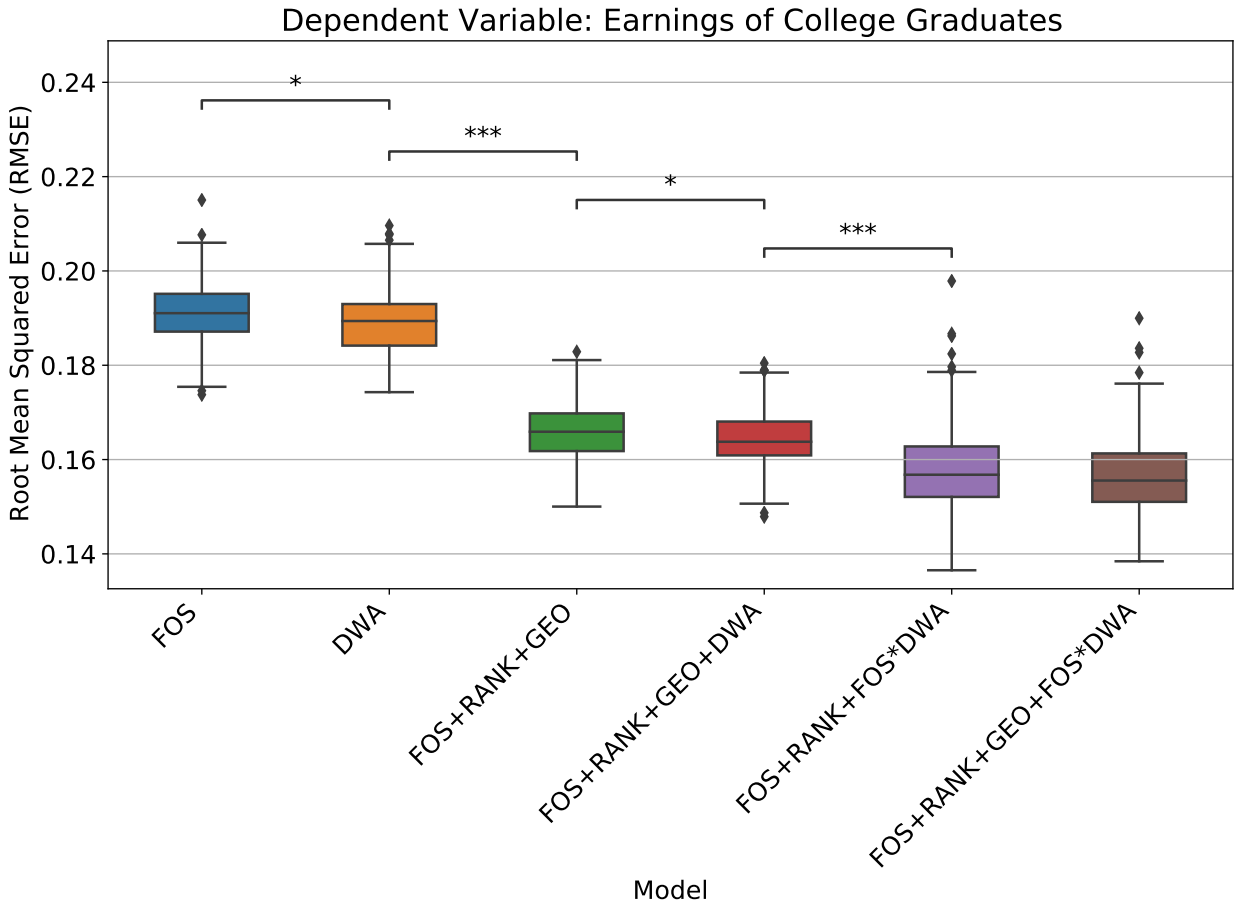


Figure 28: **Workplace activities detected from syllabi predicting median first-year earnings of college graduates across fields of study**. I perform 5-fold cross-validation and repeat 40 times (i.e., 200 trials in total) for each model and measure RMSE by the resulting model applied to the test set. Asterisks indicate the statistically significant difference between the two models' performances with Bonferroni correction. As a baseline, I consider the FOS, school ranking, and geographic fixed effects to predict earnings.

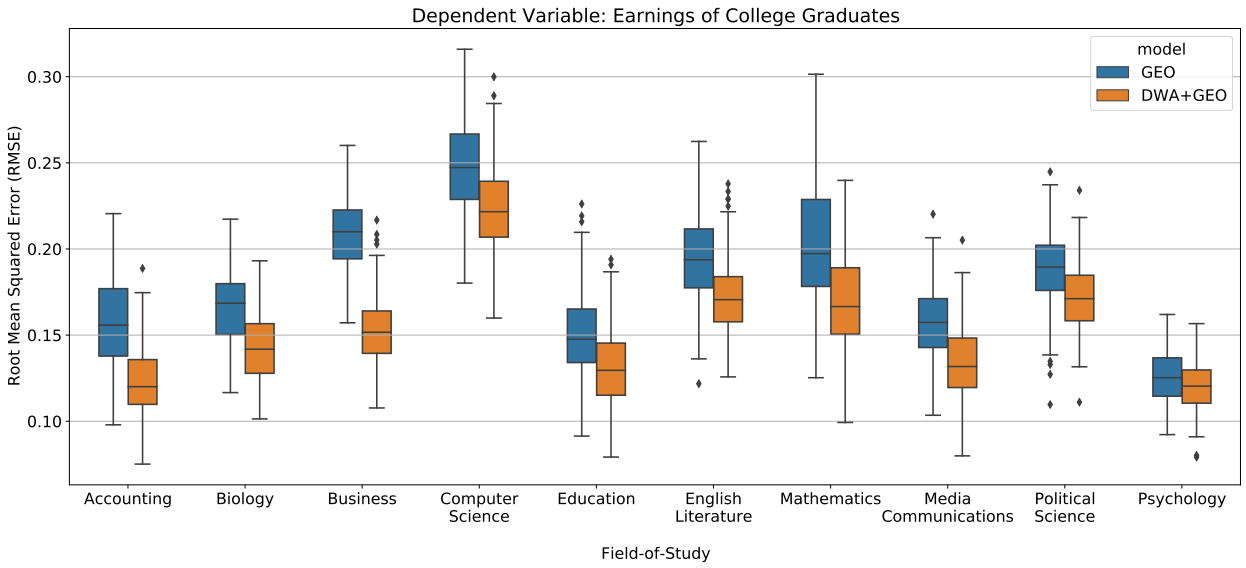


Figure 29: **Workplace activities detected from syllabi predicting median first-year earnings of college graduates within a field of study**. I perform 5-fold cross-validation and repeat 40 times (i.e., 200 trials in total) for each model and measure RMSE by the resulting model applied to the test set. The baseline *GEO* model only includes geographic variables. The performances of the *DWA+GEO* models are statistically significantly better than the *GEO* models with the *p-values* < 0.05 for all of the reported FOS (the school ranking is omitted due to the limited earnings data).

Identifying DWAs that correspond to increased earnings after graduation could inform students' course selection based on the demand for skills in the labor market. To demonstrate this, I analyze the regression of FOS *Business* as an example. After performing 5-fold cross-validation on the model determined by LASSO feature selection, there are 57 DWAs remaining. Based on our statistical regression analysis, the 57 DWA features are able to explain 69.2% of the variance of the earnings in *Business*. Among those, 10 DWAs have significant coefficients with the *p-values* below 0.05. DWAs "*complete documentation required by programs or regulations*," "*evalutate program effectiveness*," and "*advise others on career or personal development*" are positively associated with earnings while "*conduct health or safety training programs*" is negatively associated with earnings (regression coefficients estimated with $p_{value} < 0.01$ in each case). The list of DWAs that have significant coefficients for all the 10 FOS can be found in Appendix A, Table 12. The full list of all the selected DWAs including the coefficients and statistics can be found in this GitHub folder.[10]

6.4 Discussion

Knowledge, skills, and abilities shape workers' careers, and so, quantifying their sources may impact workforce development and our understanding of the labor market. Largely, higher education is a source of skill acquisition for many middle and high-skilled jobs in America. However, there is a disconnect between work and learning in the US; higher education can fail to meet the skill demands of the labor market thus creating "skill gaps" across the country. A labor market information system where work skills are shared across entities, connecting education to work, could help students know what skills they need, educators know what skills to instruct for, employers know what skills workers have, and policymakers more effectively impact workforce development. This chapter demonstrates a methodology to bridge material taught in U.S. colleges and universities with the detailed work activities (DWAs) used by the Department of Labor to describe the US workforce. This creates new opportunities to track changes in the evolution of higher education and

[10] https://github.com/HungChau/OSP-connect-higher-education/tree/main/selected_DWAs

workforce development; for example, the emergence of DWAs within the syllabi of a field of study (FOS), or major, corresponds to the co-occurrence of DWA pairs across all of academia (see Fig. 27). As an illustrative example, discussions of green technology design requirements may become more prominent in Computer Science programs because they go hand-in-hand with information technology project resource requirements, commonly taught in courses across academia. Educators, educational policy, and course recommendation systems could use these insights to design educational programs and to advise students towards the classes offering the experience that will be most valuable for their career goals. Following our example, proactive curriculum design might include green technology topics to prepare students for jobs in Computer Science.

However, it is likely not the case that every FOS will teach every skill or ability, in part, because labor market incentives for specific DWAs vary by industry, region, and employer. Thus, insights into the course topics that correspond to increased, or decreased, earnings after graduation (see Fig. 29 for example) may increase the relevance of an educational program or policy and increase students' success when they enter the workforce. For example, academic programs might grow to include new high-demand skills while decreasing emphasis on outdated topics. Such insights could inform *goal*-based learning [195] in course recommendation systems while improving explanations of recommendations. Increasingly personalized course recommendations can identify relevant topics based on students' predefined goals (*e.g.*, maximizing job earnings). For example, recommending *Business* courses that include "*complete documentation required by programs or regulations*" work activities might proactively prepare today's students to meet the growing demand for Business Analytics in the labor market.

This study has a few limitations. This study demonstrates how novel syllabus data and natural language processing (NLP) techniques can connect labor market data to higher education by predicting the change in taught skills within a FOS and linking DWAs to graduate earnings. Future work might build on my study by analyzing the causal implications of skill-level adjustments to course content. In particular, my study's approach is unable to address selection bias when students choose a university in which to enroll. However future work may study natural experiments that overcome this barrier. Potential examples

include the hiring, firing, or retirement of new faculty, the creation of a new school or department, the emergence of a large employer (*e.g.*, resulting from new tax credit), or large donations focused on specific learning outcomes. For example, future work might augment my analysis of graduate's recent earnings with other career outcome measures. Our analysis of the College Scorecard earnings data is limited to only two graduation cohorts and similar Post-Secondary Employment Outcomes data is limited to only a few institutions. Furthermore, I only consider earnings one year after graduation, which may not capture the full career trajectory [196]. However, future analysis involving workers' resumes will enable direct connections between workers' educational foundations during college and their career dynamics (*e.g.*, worker adaptability, tenure, and mobility) in addition to earnings. Similarly, job postings analysis might compare employer demands to the DWAs detected in our study thus identifying the most or least adaptive educational programs (*e.g.*, [24]). Future research along this dimension will offer new insights into the sources and sinks of the high-skilled workers that shape job polarization [26] and urbanization today [174, 185].

I have demonstrated, using mean cohort level graduate earnings, that there is already detectable variation in earnings based on skills taught in courses offered. My approach has focused on outcomes for groups of graduates (e.g., by major or university). Future work with alternative data might investigate variations in labor market outcomes for individuals. For example, students studying the same major could take different courses offered, thus learning different skills. Whether the course selection by individual students leads to different occupations and different earnings, and how much learned skills could explain individual career variation are interesting questions left to be discovered. One challenge in undertaking such research is the availability and accessibility of this type of datasets at scale due to privacy concerns. Further, my analyses focused on students with bachelor's degrees, but future work might study the skills of graduate education or the undergraduate education that lead to graduate school admission.

My study relied on simple off-the-shelf techniques in combination with novel data sources, but future work might expand my methods with more sophisticated approaches. For example, this study used pre-trained *static* word embeddings and standard document similarity techniques to detect work activities from syllabi, but more complex NLP techniques could

yield further insights. Static word embeddings are a powerful tool for capturing syntactic and semantic regularities in language, but each word is represented by a single vector regardless of context. That is, all senses of a polysemous word have to share the same representation. *Contextualized* word representations, such as Transformer-based embeddings, overcome those issues and have yielded significant improvements on many NLP tasks. Additionally, my study relies on the O*NET taxonomy used by the US Department of Labor to describe labor market trends. These granular DWAs reveal core differences between courses, fields and universities. For example, DWA relevance scores improved predictions of graduate earnings within many fields of study, but not all. This suggests that "skill" differences may impact the effectiveness of college education (in terms of earnings) but O*NET DWAs may not be the most precise taxonomy to describe the granular level of knowledge expressed in courses. This is in part because O*NET data is not designed to describe higher education, but to describe workers. There is no standard knowledge base describing more granular concepts and skills in higher education and the labor market. This highlights an urgent need for future educational research that builds a knowledge base that could standardize and advance insights into how educational foundations shape workforce development and the skills of workers. With the advances of text mining methods, one could extract skills described in course syllabi and job postings, and align those skills to connect educational content with the demands of the labor market. There are some existing job skill taxonomies to describe job postings' requirements such as BG's or LinkedIn's proprietary skill taxonomies. Börner et al. (2018) analyze course syllabi and BG's job postings focusing on areas of Data Science and Data Engineering. They use BG's skill taxonomy instead of the one used by the U.S. Bureau of Labor Statistics to analyze skill discrepancies between research, education and jobs. Modeling job postings with NLP techniques has also been shown to be useful in understanding wage premia [25]. Although my study focuses on the work side of job seeking, I acknowledge that the demand from the employer side is also important to understand the holistic picture from skill offerings in higher education to skill demands in the labor market; which could benefit many applications such as identifying potential curricular gaps or recommending courses to meet jobs' requirements.

Increasingly, researchers and policymakers use workers' skills and abilities to describe

labor market outcomes in addition to workers' educational attainment based on their occupation [28]. However, similar data and methods are only just being developed and applied to workforce development and, in particular, to higher education. This study offers an approach and a methodology to connect higher education to workplace skills thus enabling new strategies for course recommendation, curriculum design, and education policy that prepare students to meet their career goals.

In summary, I introduce a proof-of-concept method that bridges the gap between workplace skills and those imparted through higher education. This innovative approach involves the creation of modeling course syllabi that intricately align with real-world work activities. By doing so, it enables us to accurately forecast educational trends and predict the future earnings of graduates. This method holds the potential to revolutionize the way we approach education and career preparation, offering a pathway to create personalized course recommendation systems that are both explainable and tailored to individual needs. It opens doors to more informed and adaptable educational information systems, empowering students to make informed choices and institutions to better meet the evolving demands of the job market.

7.0 CAREER-ORIENTED EXPLAINABLE COURSE RECOMMENDATION

With compelling results emerging from our course representation methods detailed in Chapter 6 and the concept extraction model for course descriptions outlined in Chapter 4, the objective of this chapter is to design an explainable, personalized course recommendation system that aligns with students' career aspirations. Additionally, this study seeks to compare two distinct approaches for skill representation within course recommendation systems: concepts automatically extracted from course descriptions, as detailed in Chapter 4, and O*NET DWAs manually constructed by experts to describe work in the US labor market, as discussed in Chapter 6. This system is the first to incorporate job information and skills to enhance student achievement and future career prospects. Its goal is to guide students toward specific courses that will equip them with the necessary skills for their desired careers. The system tailors course suggestions based on students' enrollment histories and career preferences while also providing explanations for its recommendations. To assess the effectiveness of the proposed design, I will deploy the system and conduct a user study involving undergraduate students from the School of Computing and Information at the University of Pittsburgh.

7.1 Introduction

In today's dynamic and competitive job market, students face the formidable challenge of choosing the right courses that align with their career aspirations. The importance of this academic decision cannot be overstated, as it directly impacts their future employability and success. Recognizing this critical need, this research aims to create an innovative and essential tool—the explainable, personalized course recommender system. This system aims to guide students on a tailored educational journey that not only aligns with their career preferences but also equips them with the requisite knowledge and skills demanded by their

chosen profession.

Fundamentally, my research embarks on a pioneering quest to bridge the gap between college education and the job market, focusing on the pivotal concept of unified skills. By harnessing these skills as the fundamental building blocks, the course recommender system will provide students with course recommendations that are both insightful and easily understandable. The system leverages a student's course enrollment history, assuming they have already taken courses for at least one semester, in conjunction with their career preferences to curate a personalized list of courses. Each course recommendation is accompanied by a skill-based explanation of how it contributes to the acquisition of skills essential for their future career.

This study is underpinned by three primary objectives:

1. **Proposal of an Approach for Unified Skills-Based Explainable Course Recommendation**: I will introduce an innovative approach that utilizes unified skills to enable explainable course recommendations. This approach seeks to establish a direct link between the educational sphere and the job market, thereby empowering students to make informed decisions about their academic journey.
2. **Evaluation of the Utility of Job Information in Enhancing Course Recommendations**: I will assess the impact of incorporating job market information on the quality of course recommendations. By leveraging job-related data, I aim to enhance the precision and relevance of our recommendations, thus ensuring that students receive guidance that is both forward-looking and aligned with industry needs.
3. **Validation of the Value of Explanation in Enhancing User Perception**: I will investigate the efficacy of providing explanations alongside course recommendations. My research seeks to validate that these explanations not only aid users in making more informed choices but also foster acceptance of the recommendations, thereby enhancing the overall user experience.

To the best of my knowledge, this study represents a pioneering effort in the realm of career-oriented, explainable course recommendation, uniquely leveraging real-world job information. I will deploy an explainable course recommendation engine using course de-

scriptions sourced from the School of Computing and Information at the University of Pittsburgh, O*NET datasets, and job postings from Burning Glass Technology. Subsequently, I will conduct a user study with undergraduate students at the School of Computing and Information, University of Pittsburgh, to validate the importance of job information in course recommendations and test the hypothesis that explanations could improve user perception of recommendations.

In this chapter, I will first present how to apply the course modeling method described in Chapter 6 to represent courses and jobs through the unified skills; which serves as the foundation for the career-oriented explainable course recommendation system. Subsequently, I will provide comprehensive insights into the recommendation and explanation method, share details of the study experiments and analyses, and finally discuss the study's contribution, limitations and potential future directions to conclude the chapter.

7.2 Skill-based Document Representation

Similar to the way course syllabi are represented based on their relationships with the DWAs discussed in Chapter 6, I generalize it to model *documents* (i.e., course descriptions or job postings) based on their relationships with skills (i.e., Concepts or DWAs). Instead of using the pre-trained *fastText* word embeddings, I utilize Sentence BERT (SBERT) [197] for generating state-of-the-art text embeddings. SBERT is a modified version of the pre-trained BERT network that employs Siamese and triplet network architectures to generate semantically meaningful text embeddings, allowing for direct comparison through cosine similarity. I obtain the embedding vectors of 768 dimensions from the *all-mpnet-base-v2* version[1] of SBERT for each document d and each skill s, and then compute their *relationship* ($0 \leq r_d(s) \leq 1$) using cosine similarity.

Skills: Discussed in Chapter 1, Section 1.1, these are described as knowledge and expertise in higher education and the labor market in the social economic research community. It has been used in recent studies to universally describe courses, scientific articles and jobs

[1] https://www.sbert.net/docs/pretrained_models.html

[24, 26, 27]. In this work, I will use the term *skills* to represent different types of knowledge including extracted concepts (or Concepts) from texts (e.g., *support vector machine*) and O*NET DWAs (e.g., *develop methods of social and economic research*). Skills are considered atomic units of learning and labor. I will use skills to represent courses, jobs and occupations to guide recommendation and explanation.

Career Categories: a list of careers in the domain of Computing and Information is defined based on O*NET Career Taxonomy and Burning Glass Specialty. Each career is a combination of an occupation and a specialty such as a career in Web Developers occupation with PHP Develope specialty.

- *O*NET Career and Occupation*[2]: used by the Department of Labor to describe work in the US labor market, including 24 occupations related to Computer and Information Science; e.g., Web Developers, Computer Hardware Engineers, Business Intelligence Analysts.
- *Burning Glass Specialty*: each job posting in the BG dataset has information about O*NET occupations and specialties defined by BG technologies. There are 145 BG specialties linked to 24 O*NET occupations. Though it is not a one-to-one mapping, a specialty could be considered as a sub-category of an O*NET occupation; e.g., specialties Data Analyst (Finance) and Data Analyst (Healthcare) in occupation Business Intelligence Analysts, specialties .NET Developer / Engineer and Back End Developer / Engineer in occupation Web Developers.

Course Representation. To represent a course c, I use its catalog description as an input for SBERT to obtain an embedding vector. Similarly, each skill s in a fixed set of skills is represented by an embedding of the same dimension. Then I compute the relationships between each of the skills and the course $r_c(s)$. As a result, courses are represented based on their relationships with the skills (called the *skill*-based course representation); $V_c : r_s^{|S|}; r_s \equiv r_c(s) \in [0, 1]$, explaining how strongly the skill $s \in S$ associated to the course description c; S is the set of skills (i.e., Concepts or DWAs). Lastly, the vector V_c is transformed into the unit norm form, $\sum_{s \in S}(r_s) = 1$.

[2]https://www.onetonline.org/find/career?c=0&g=Go

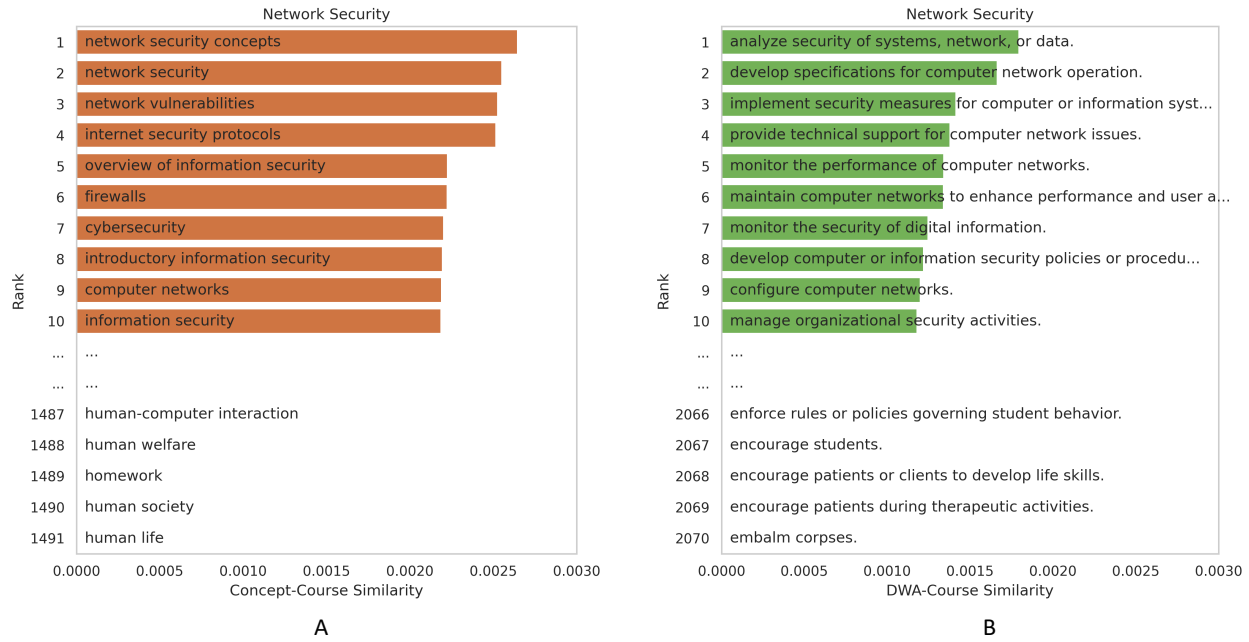


Figure 30: An example of a Network Security course at the University of Pittsburgh and the skills that are most and least strongly associated with its description. (A) Concept-inferred course representation. (B) DWA-inferred course representation.

Job Representation. Since I do not have access to actual job posting descriptions in the Burning Glass dataset, I use the BG skills associated with each job posting as an approximation of its description. I get the embedding vector for each of the BG skills associated with the job using SBERT, and then averaging those vectors to obtain a vector representation for the job. As a result, the job vector also has the same 768 dimensions as the skill vectors do. From now, I model the skill-based job representation via job-skill relationship $r_j(s)$ the same way for the course presentation described above; $V_j : r_s^{|S|}; r_s \equiv r_j(s) \in [0, 1]$, explaining how strongly the skill $s \in S$ associated to the job j; S is the same set of skills. The job vector V_j is finally normalized to the unit norm form, $\sum_{s \in S}(r_s) = 1$.

Representation Adjustment. One of the issues with the current representation approach with either type of skill is not handling ubiquitous skills across courses within a major or jobs within a career; for example, DWAs "Identifying Objects" and "Communicating with Supervisors and Peers", and Concepts "Data analytics" and "Data and analytics". Thus, I

Figure 31: An example of Machine Learning Engineer job posting in Burning Glass dataset and the skills that are most and least strongly associated with its *approximate* description. (A) Concept-inferred job representation. (B) DWA-inferred job representation.

focus on skills that are "overexpressed" in a course or job and penalize skills that are ubiquitous. First, I calculate the revealed comparative advantage (RCA) or "location quotient" [26, 198] of each skill s in a document d (i.e., a course or job) according to

$$rca_d(s) = \frac{r_d(s)/\sum_{s' \in S} r_d(s')}{\sum_{d' \in D} r_{d'}(s)/\sum_{d' \in D, s' \in S} r_{d'}(s')} \tag{19}$$

This technique has been used in a variety of applications, including analyzing the polarization of workplace skills [26], identifying the key exports of nations [198], and finding the primary sectors in urban areas [199]. Courses (or jobs) could be distinguishable from each other according to their "effective use" of skills. According to [26]; the effective use of skills is defined using $e(d, s) = 1$ if $rca(d, s) > 1$, and $e(d, s) = 0$ otherwise. Therefore, I adjust the vector representation of a course $V_c : r_s^{|S|}$, with $r_s = 0$ if $e(c, s) = 0$; and a job $V_j : r_s^{|S|}$, with $r_s = 0$ if $e(j, s) = 0$. Figure 30 shows an example of the most and least prevalent skills detected for a Network Security course at the University of Pittsburgh. Figure 31 shows an example of the most and least prevalent skills detected for a Machine Learning Engineer job in the BG dataset.

Career Representation. Similar to the way I represent a field-of-study and a university using their relationships to each of the DWAs via course syllabi (described in Chapter 6). The job-skill relationships ($r_j(s)$) help to model a career ca in terms of its relationship to each of the skills according to

$$r_{ca}(s) = \frac{1}{|J_{ca}|} \sum_{j \in J_{ca}} r_j(s) \tag{20}$$

These relevance scores are a measure of how strongly the skill s is represented in a career; $r_{ca}(s)$ (the relevance score of the skill s to the career ca) is the average over the similarity scores of that skill across job $j \in J_{ca}$ (which represents all of the jobs within a given career ca). Figure 32 shows an example of the most and least prevalent skills detected for the career in Web Developers occupation and PHP Developer specialty using the BG job data.

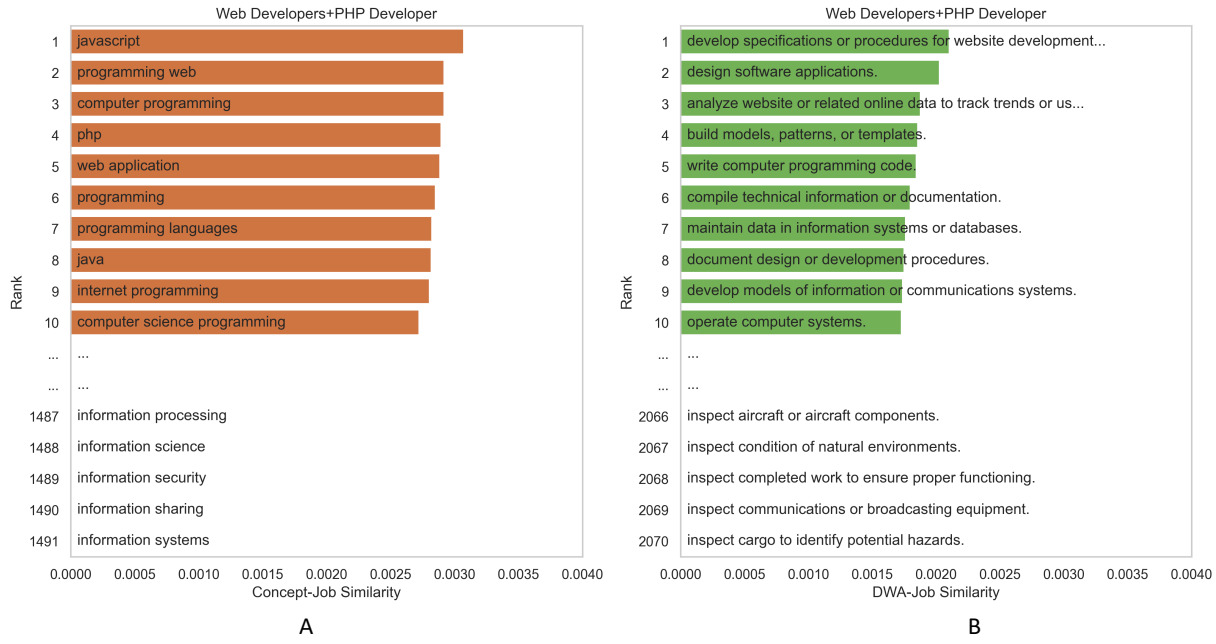


Figure 32: An example of the career in Web Developers occupation and PHP Developer specialty using the BG job data and the skills that are most and least strongly associated with its job posting descriptions. (A) Concept-inferred course representation. (B) DWA-inferred course representation.

7.3 Recommendation Method

In this study, a career pathway is defined by O*NET occupations and BG specialties described above. Students can provide their career preferences by selecting an occupation (e.g., Web Developers) and a specialty (e.g., Back End Developer / Engineer). Based on their preferences and course enrollment histories, the system is designed to recommend a list of courses the student should consider to obtain the necessary skills for their future career. The explainable career-oriented course recommendation system is designed as follows (and see Figure 33):

1. First, represent each course in the student enrollment history as a continuous vector described above: $V_c : r_s^{|S|}$.
2. Create the student skill profile as a continuous vector $V_p : r_s^{|S|}; r_s = max(r_{c \in C_p}(s))$ (C_p is the list of courses in the student enrollment history); $r_c(s)$ is how strongly skill s associated with course c; $r_c(s) = 0$ when skill s is not associated with any course in the enrollment history; S is the set of skills (i.e., Concepts or DWAs).
3. From the student's input preferred career (O*NET occupation + BD Specialty), retrieve the list of jobs that belong to the career. Represent each job as a vector V_j by averaging all the vector representations of the jobs linked to the career, V_{ca} (presented in the previous section).
4. Calculate the *required*-skill vector that contains information about skills needed for the career but the student has not learned yet or has not learned enough, $V_{re} : r_s^{|S|}; r_s = max(0, r_{ca}(s) - r_p(s))$; where $r_{ca}(s)$ and $r_p(s)$ are how strongly skill s associated with career ca and student profile p, respectively; $r_s = 0$ if career ca does not require skill s or the student has already mastered the skill.
5. Finally, generate a list of recommended courses by minimizing the objective function that is defined as the *distance* between the target course V_{tc} and the *required* skill vector V_{re} for the preferred career.

$$dist(V_{tc}, V_{re}) = \sqrt{\sum_{s}^{|S|} min(0, r_{tc}(s) - r_{re}(s))^2} \tag{21}$$

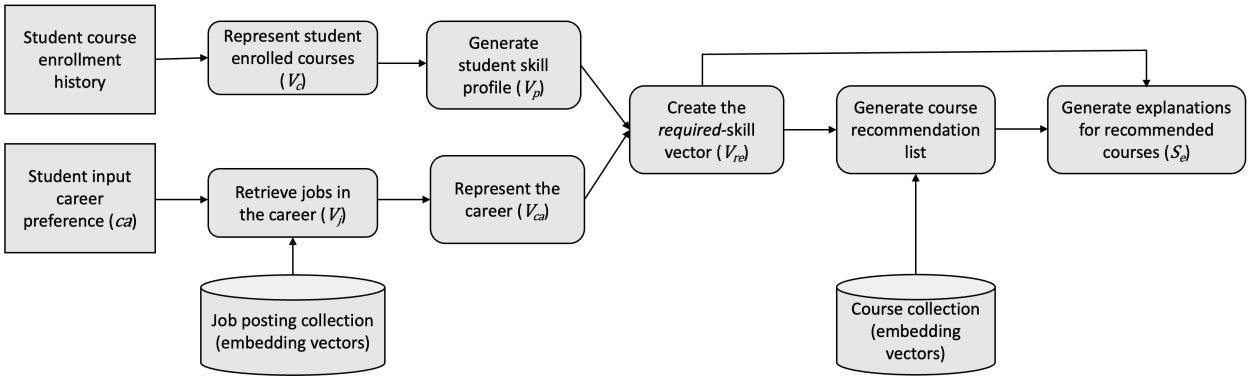


Figure 33: The design of an explainable career-oriented course recommendation engine using student course enrollment history and job posting data.

$r_{tc}(s)$ is how strongly skill s associated with target course tc; $r_{re}(s)$ is a required level of skill s the student needs to master for the career; $dist(V_{tc}, V_{re}) = 0$ means that the target course tc provides all the remaining knowledge that the student needs for their preferred career. The smaller the distance $dist(V_{tc}, V_{re})$, the closer the student is ready for their career.

7.4 Explanation Method

Besides evaluating the effectiveness of the proposed career-oriented course recommendation methods, my goal is to examine how explanation affects the way users respond to the recommendation. Thus, I will generate skill-based explanations for each of the courses in the recommendation list based on three sources of information: (1) student prior knowledge, (2) skills the target course offers, and (3) skills needed for the occupation in the preferred career.

My approach to recommendation explanation is to connect skills required for the career to target course recommendations. The explanation will include information about required

skills that the student has not learned yet or learned enough. The set of *explaining* skills (S_e) is defined as following:

$$S_e = S_{re} \cap S_{tc} \tag{22}$$

Where S_{re} is the list of skills that the student has yet to master or requires further improvement in proficiency, and S_{tc} is the list of skills offered by the target course tc. Given $s_e \in S_e$, $0 < s_e = min(r_{re}(s_e), r_{tc}(s_e)) <= 1$. The explanation displays the top 10 skills in S_e.

Diversify the set of explaining skills: concepts extracted from the course descriptions could be similar to each other or describe the same topics such as *K-Means Clustering* and *K-Means Algorithm*. To improve the explanation with extracted concepts, the diversification process is aimed to provide a diverse set of skills used for explanation by removing concepts that are too similar to those already in the current explaining skill set. Since DWAs, which are manually constructed by experts, are considered a diverse set of skills to describe work in the US labor market, this process does not apply to the explanation with DWAs. Algorithm 1 shows the process of diversifying the list of explaining skills.

7.5 Study Experiments

This section outlines the implementation of the proposed explainable course recommendation at the School of Computing and Information, University of Pittsburgh. I conducted a survey-based user study with students to gather their feedback on various aspects. The study aims to validate the importance of job information in course recommendations and test my hypothesis that explanation is helpful in improving user perception on recommendations.

7.5.1 Implementation Details

I collected all the lecture courses offered or required by the School of Computing and Information at University of Pittsburgh. The courses with insufficient description (i.e., less

Algorithm 1 Generating a diversified list of skills for explanation

```
explaining_skill_list = []
threshold = 0.85 //between 0 and 1, the higher the less diversified
for s_e in sorted(S_e):
    flag = False
    for s_i in explaining_skill_list:
        if cosine_similarity(s_e, s_i) >= threshold:
            flag = True
            break
    if flag==False:
        explaining_skill_list.append(s_e)
    if len(explaining_skill_list)==10:
        break
return explaining_skill_list
```

than 15 words) are removed from the collection, resulting in 209 courses. I use the trained concept extraction model (presented in Chapter 4) to extract all the concepts in these courses. After post-processing including removing generic concepts such as *'cs', 'sci', 'infsci'*, concepts containing more than 5 words and combining singulars and plurals of the concepts, there are 1491 concepts remaining. For the BG dataset, I collect 77,624 jobs from the O*NET occupations related to computing and information science including 'Bioinformatics Scientists', 'Biostatisticians', 'Computer Hardware Engineers', 'Computer and Information Research Scientists', 'Data Scientists', 'Robotics Engineers' and the occupations in 'Information Technology', 'Information Technology; Information Technology' O*NET Career Clusters. Jobs that have less than 5 BG associated skills and no specialty are removed from the collection. To sufficiently estimate the skills required in a career, I only select careers that have at least 20 job postings. As a result, 67,238 jobs, 24 O*NET occupations and 145 BG specialties are obtained (see Table 10).

I implemented two recommender systems, one for Concept skills and the other for DWA skills. In the Concept system, courses, jobs, and careers are represented as continuous

Table 10: The summary of data used in career-oriented course recommendation

Number of SCI lecture courses	209
Number of extracted concepts from the course descriptions	1491
Number of O*NET DWAs	2070
Number of Computing & Information-related jobs in BG	67238
Number of O*NET occupations	24
Number of BG specialties	145

vectors with 1491 dimensions, corresponding to the number of concepts. Similarly, in the DWA system, courses, jobs, and careers are represented as continuous vectors with 2070 dimensions, reflecting the number of DWAs.

7.5.2 User Study

Procedures. This study is conducted online using the Qualtrics Survey system at University of Pittsburgh. The student subjects are required to provide a list of courses they already took and passed and their preferred career (by choosing an O*NET occupation and then a BG specialty). They are randomly assigned to one of the two *between*-subject conditions (*Explanation*), and each subject participates in both the two *within*-subject conditions (*Skill*). The subjects participate in two separate sessions for two *within*-subject conditions. The sessions are held three days apart, and during each session, participants are shown a list of five course recommendations based on their past course history, career goals, and the skill condition (i.e., DWA or CON) they are assigned to. The recommendations are presented through Qualtrics, and include information about the selected occupation and specialty, course ID and title, a description of the course, skill-based explanation (only for *Explanation* conditions) and a survey questionnaire with multiple choice questions. Participants are asked to carefully review the recommendations and answer a series of questions. In total, each participant will evaluate and provide feedback on ten recommended courses.

Participants. I recruited 52 participants via email advertisements from their advisors and ad flyers at SCI buildings. Participation is limited to undergraduate students at

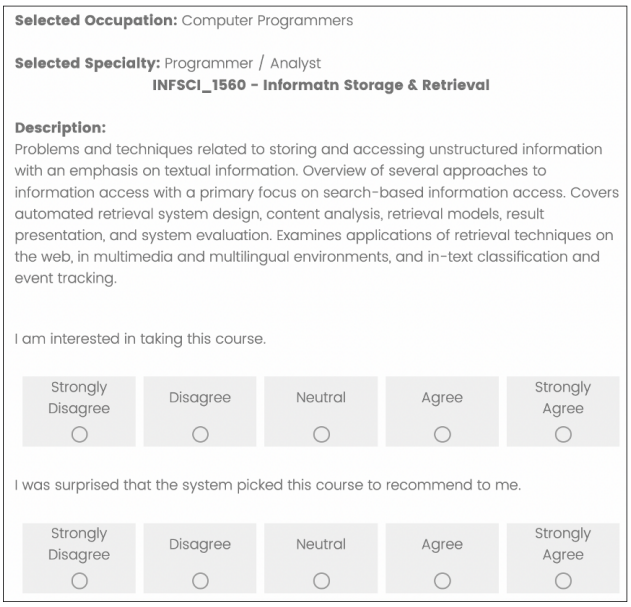


Figure 34: A recommended item for no explanation (C1 & C3) via Qualtrics.

SCI, University of Pittsburgh and having completed at least one semester. Six participants dropped after the background information collection or the first session, remaining 46 subjects for final analyses. Among the subjects, 24 are in the *Explanation* conditions and 22 are in the *Non Explanation* ones. The subjects include 19 freshmen, 14 sophomores, 10 juniors and 3 seniors. The entire study takes about 45 - 60 minutes. Each participant is paid 20 USD after completing the study.

Design and Analysis. The study is designed as a mixed *between*- and *within*-subjects study to measure: (1) The *success* and *unexpectedness* of the two recommendation methods; and (2) The effectiveness of explanation on the recommendation w.r.t *success* and *unexpectedness*. There are four groups C1, C2, C3 and C4:

- *Within*-subject conditions (*Skill*): *DWA* (C1 & C2) vs. *Concept* (C3 & C4)
- *Between*-subject conditions (*Explanation*): *NoExp* (C1 & C3) vs. *Exp* (C2 & C4)

I collected the following measures:

- All conditions:

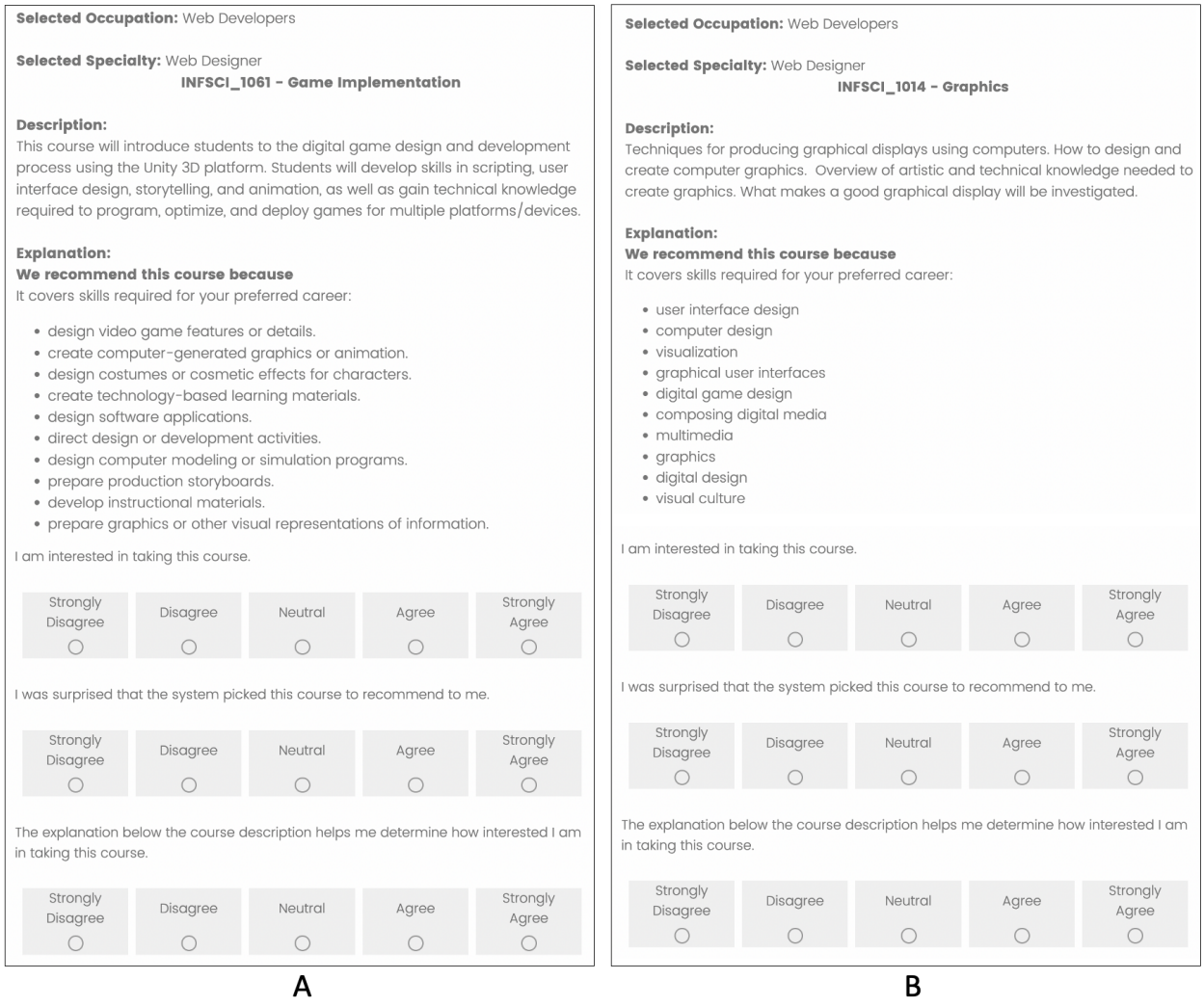


Figure 35: A recommended item for with skill-based explanation (C2 & C4) via Qualtrics. (A) The explanation shows the top 10 DWAs offered by the course and required for the student's career. (B) The explanation shows the top 10 concepts offered by the course and required for the student's career.

  - Q1. Success (Interest): Participants respond to the statement "*I am interested in taking this course.*" [167, 21, 38] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (for each course recommendation) (see Fig. 34).
  - Q2. Unexpectedness: Participants respond to the statement "*I was surprised that the system picked this course to recommend to me.*" [168] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (for each course recommendation) (see Fig. 34).
- Explanation conditions (C2 & C4):
  - Q3. Explanation: Participants respond to the statement "*The explanation below the course description helps me determine how interested I am in taking this course.*" [32] on a 5 point Likert scale from 1=Strongly Disagree to 5=Strongly Agree (for each course recommendation) (see Fig. 35).
  - Q4. Skill Quality Comparison: At the end of the study, participants respond to the two statements "*Compare the two systems, which type of skills helps me better understand how the recommended courses relate to my field of study and selected career?*" and "*Compare the two systems, which type of skills describes the content of the recommended courses better?*" on a 5 point Likert scale from 1=Much Better to 5=Much Worse.
  - Q5. Skill Quantity: in addition to the main measures, at the end of each session, participants are asked to respond to the statement "*The system presents to you a list of 10 skills to explain the recommendations. The number of skills is:*" on three options: Too Few, Good or Too Many. They can also provide a number of skills they think are sufficient for them to assess the recommendations.

In this study, each participant rates multiple items during the session. Treating repeated measurements on the same participant as independent could potentially lead to a violation of correlated errors [169, 170]. To address this issue, I use Generalized Linear Mixed Models for the primary analyses. These models take into account that the ratings come from the same users, incorporating them as random effects and allowing for the estimation of error correlations resulting from repeated measurements. For the comparison questions (i.e., Q4 and Q5), I use paired t-tests for analysis.

7.5.3 Results

Interestedness. Comparing the *baseline* model (i.e., only the intercept) and the *random intercept* model (i.e., for different participants), the results showed statistically significant variance in intercepts across participants, $\chi^2(2) = 52.61, p < .0001$. Therefore, the *participant* random effects are included in the main analysis. As depicted in Fig. 36 and 37, in general, the participants in *Concept* conditions ($M = 3.74$, $N = 230$) show a higher interest in taking the recommended courses compared to those in *DWA* conditions ($M = 3.62$, $N = 230$). Similarly, the participants in Exp conditions ($M = 3.77$, $N = 240$) show higher interest in taking the recommended courses compared to those in NoExp conditions ($M = 3.58$, $N = 220$). Those in condition C4 (*Concept + Explanation*) ($M = 3.87$, $N = 120$) show the highest interest. From the statistical analysis, the results show that the relationship between skill type and interest in taking the course showed significant variance in intercepts across participants, $SD = 0.54$ (95% CI: 0.41, 0.71), $= \chi^2(2) = 52.61, p < .0001$. The explanation has a positive effect on how interested participants are in taking the course but it is not significant, $b = 0.19$, $t(44) = 1.01$, $p = 0.32$. There is a positive impact of Concept over DWA on the *interestedness* but not significant either, $b = 0.12$, $t(45) = 1.22$, $p = 0.23$. The effect of providing explanations for both skill types is positive, but not statistically significant. Additionally, there is no significant interaction between skill type and the presence of an explanation.

It's possible that the study was underpowered. The lack of a significant effect in the explanation group could be because the effect size was medium or smaller, requiring a larger sample size than the 46 participants in the study. Additionally, the *between*-subjects condition often requires more participants. Furthermore, students in the no-explanation condition may have rated the recommendations relatively high because they agreed with them, but they may not have realized what they were missing without the explanation. If they had participated in both conditions, they could have compared the conditions, potentially leading to a more significant difference. This *within*-subjects setting could help to remove potential subject variability and allow for direct comparison of conditions; the interaction, however, may not be realistic. In addition, the participants in the *Concept* conditions showed a greater

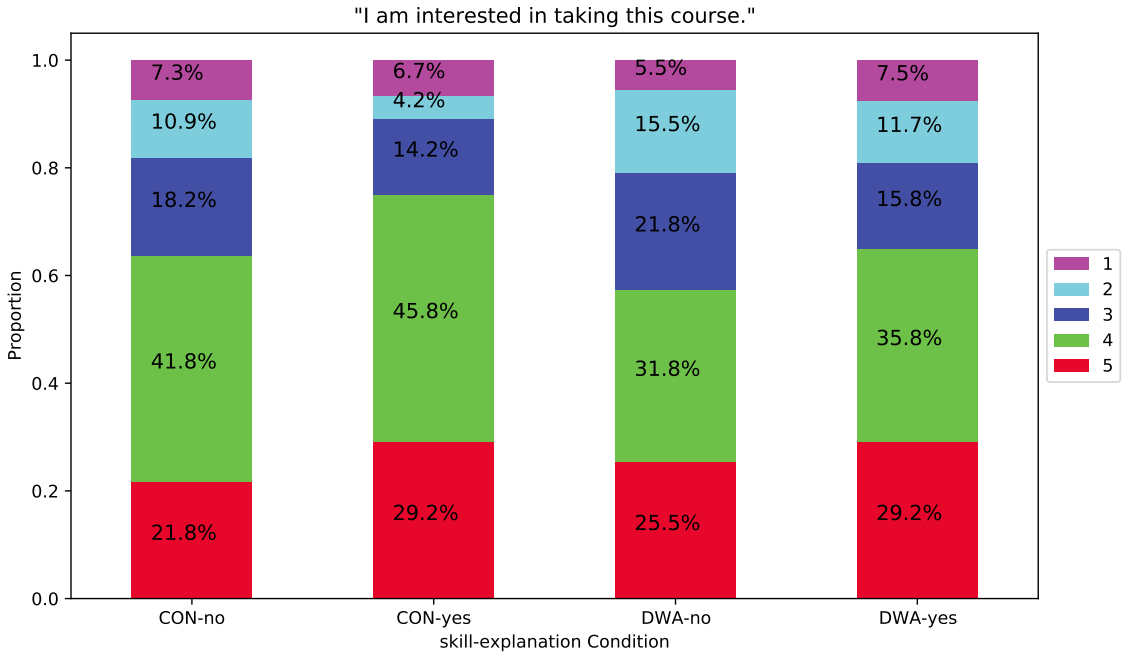


Figure 36: Proportional distribution of user responses to the statement 'I am interested in taking this course.' across different skill-explanation conditions: *Concept* system without *Explanation* (CON-no), *Concept* system with *Explanation* (CON-yes), *DWA* system without *Explanation* (DWA-no), and *DWA* system with *Explanation* (DWA-yes). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'.

interest in taking the recommended courses compared to those in the *DWA* conditions. This may be because the concepts are directly extracted from the course descriptions, resulting in a course representation that is closer to the content of the course. As a result, the recommendations are more relevant to the participants' interests and preferences, leading to greater interest in taking the courses.

Unexpectedness. Likewise, when comparing the results of the baseline model (which only includes the intercept) to the random intercept model (which accounts for variations among participants), it was found that there was a statistically significant difference in the intercepts across participants, $\chi^2(2) = 98.92, p < .0001$. As a result, the random effects for the participants are included in the primary analysis. As depicted in Fig. 38 and 39, in generally, the participants in *DWA* conditions ($M = 2.78$, $N = 230$) demonstrated higher levels of surprise about the course recommendations when compared to those in *CON* conditions ($M = 2.56$, $N = 230$). Similarly, the participants in *Exp* conditions ($M = 2.86$, $N = 240$) showed higher levels of surprise about the recommendations when compared to those in *NoExp* conditions ($M = 2.46$, $N = 220$). The highest levels of surprise were found among participants in the C2 condition (*DWA + Explanation*) ($M = 3.02$, $N = 120$). The statistical analysis reveals that there is significant variation in intercepts among participants regarding the relationship between skill type and surprise with the recommendation, $SD = 0.66$ (95% CI: 0.51, 0.85), $= \chi^2(2) = 98.92, p < .0001$. The explanation has a positive impact on participant's surprise with the recommendation, but it's not statistically significant, $b = 0.40$, $t(44) = 1.82$, $p = 0.075$. However, for the between-subjects DWA conditions (C1 and C2), the explanation has a statistically significant positive effect on the recommendation, $b = 0.49$, $t(44) = 2.24$, $p = 0.031$. Furthermore, the study finds a significant positive impact of DWAs over extracted concepts on the level of surprise, $b = 0.23$, $t(45) = 2.44$, $p = 0.018$. With the explanation, participants in the *DWA* condition reported a significantly higher level of surprise compared to those in the *Concept* condition, $b = 0.32$, $t(23) = 2.62$, $p = 0.015$.

According to the result, providing an explanation has a positive effect on the outcome, with statistical significance observed for the *DWA* skill type and near-significance across both skill types. As the recommendations from the *Concept* system demonstrate higher

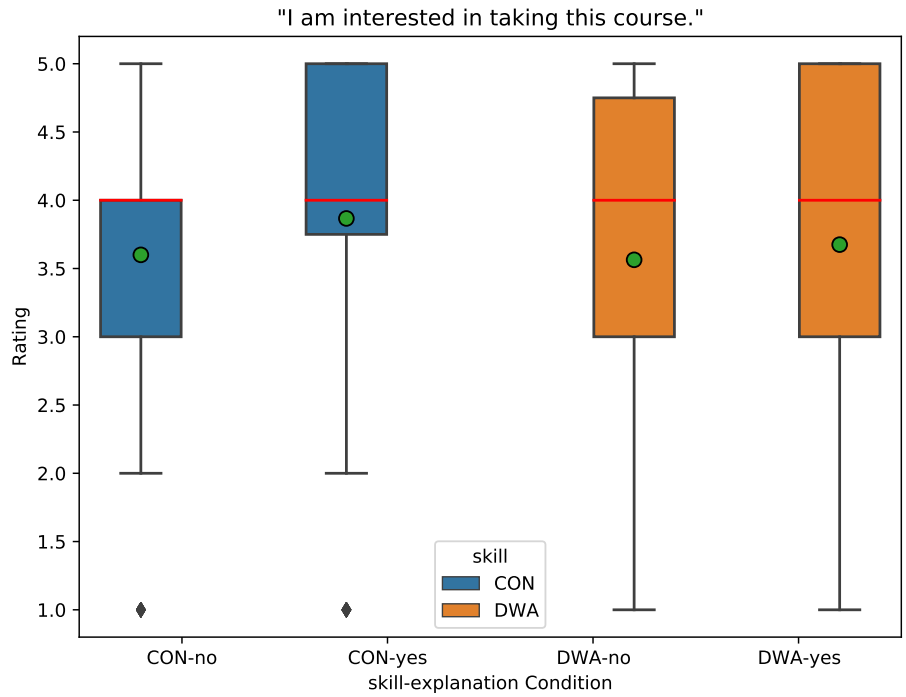


Figure 37: Graph illustrating the ratings in response to the statement 'I am interested in taking this course.' across four conditions: *Concept* system without *Explanation* (CON-no), *Concept* system with *Explanation* (CON-yes), *DWA* system without *Explanation* (DWA-no), and *DWA* system with *Explanation* (DWA-yes). The red lines indicate the median ratings, while the green circles depict the average ratings.

relevance, as previously described, they are generally less surprising to participants compared to those from DWAs, regardless of whether an explanation is given. The impact of providing an explanation is statistically significant. The knowledge captured by DWAs may not be immediately apparent or easily recognized. DWAs may be more effective in selecting unexpected courses, which could have either a positive or negative effect on the resulting recommendation.

Explanation. Like previous measures, the results of the statistical analysis showed a statistically significant difference among participants in the relationship between the skill type and the usefulness of explanation in helping users' decision on accepting the recommendation, $SD = 0.31$ (95% CI: 0.20, 0.47), $\chi^2(2) = 27.02, p < .0001$. As shown in Fig. 40, participants who received explanations from either skill type agreed that the explanation helped in determining their interest in the recommendation ($M = 4.26$, $N = 240$). There was no significant difference between *DWA* skill type ($M = 4.27$, $N = 120$) and *Concept* skill type ($M = 4.26$, $N = 120$) in terms of their effect, $b = 0.01$, $t(23) = 0.10$, $p = 0.921$. Nevertheless, the explanation is beneficial for students to gain more insight into the recommendation, which helps them make an informed decision. As a result, this could potentially lead to an increase in user trust and the acceptance rate towards recommendations.

Skill Quality Comparison. As demonstrated in Fig. 41 and 42, the majority of participants favored the *DWA*-based explanation over the *Concept*-based explanation. For the first question, 14 out of 22 participants rated the DWA system as better or much better than the *Concept* system, while only 4 participants rated the opposite. The two-tailed paired T-test results indicated that the *DWA* system is significantly better than the *Concept* one with M = -0.55, $p = 0.05$ (on a normalized rating scale of -2 to 2). Similarly, for the second question, 13 out of 22 participants rated the *DWA* system as better or much better than the *Concept* system, while only 3 participants rated the latter better. The two-tailed paired T-test showed that the *DWA* system is significantly better than the *Concept* system with M = -0.64, $p = 0.01$.

Both types of skills approximately describe the course based on their relevance. While concepts are automatically extracted from the course descriptions, some may be semantically similar but treated as distinct entities, such as 'data analytics' and 'data and analytics'. This

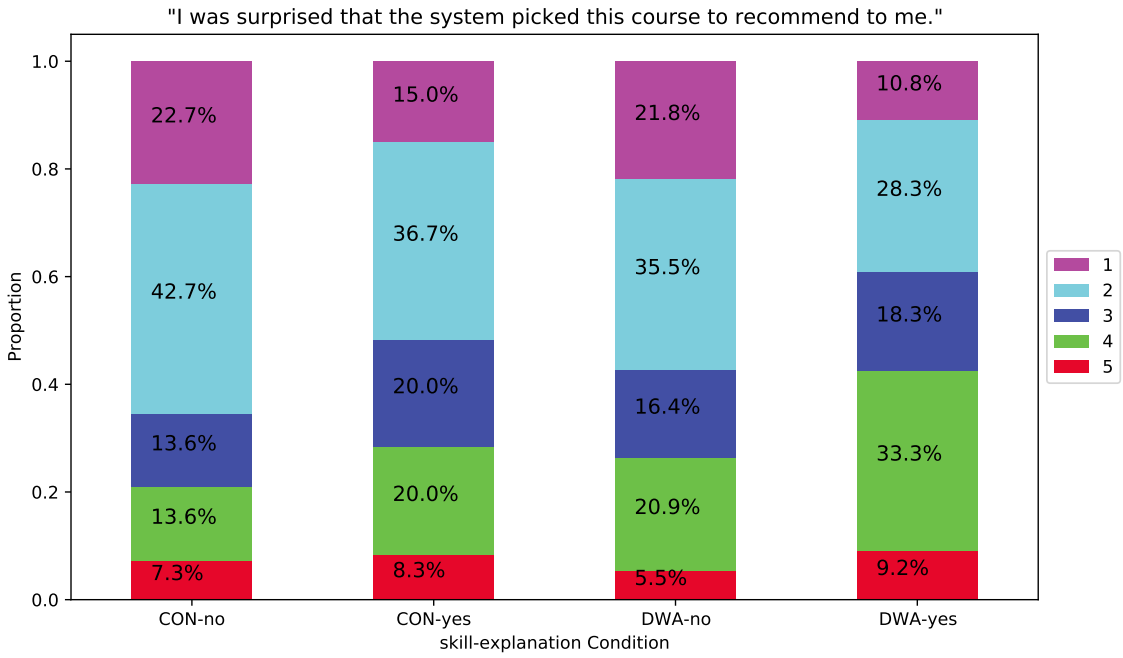


Figure 38: Proportional distribution of user responses to the statement 'I was surprised that the system picked this course to recommend to me.' across different skill-explanation conditions: *Concept* system without *Explanation* (CON-no), *Concept* system with *Explanation* (CON-yes), *DWA* system without *Explanation* (DWA-no), and *DWA* system with *Explanation* (DWA-yes). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'.

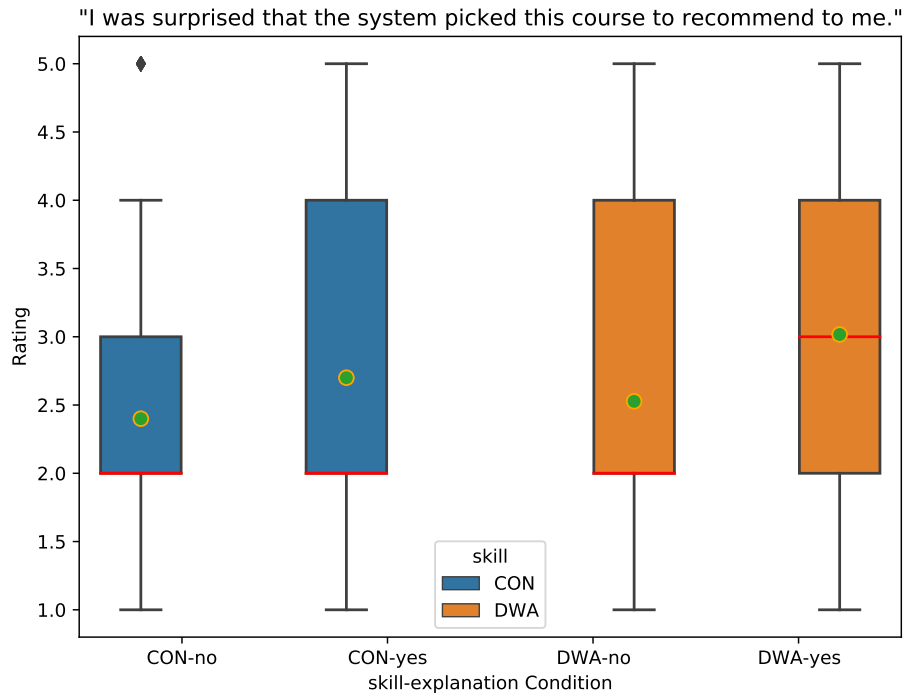


Figure 39: Graph illustrating the ratings in response to the statement 'I was surprised that the system picked this course to recommend to me.' across four conditions: *Concept* system without *Explanation* (CON-no), *Concept* system with *Explanation* (CON-yes), *DWA* system without *Explanation* (DWA-no), and *DWA* system with *Explanation* (DWA-yes). The red lines indicate the median ratings, while the green circles depict the average ratings.

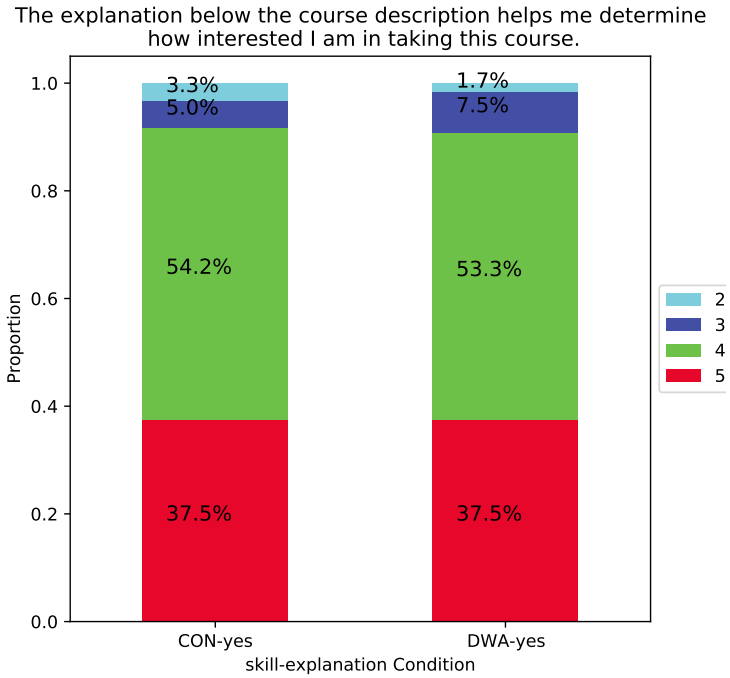


Figure 40: Proportional distribution of user responses to the statement 'The explanation below the course description helps me determine how interested I am in taking this course.' across different skill conditions with explanations: *Concept* system with *Explanation* (CON-yes), and *DWA* system with *Explanation* (DWA-yes). Ratings: 1 - 'Strong Disagree', 2 - 'Disagree', 3 - 'Neutral', 4 - 'Agree', 5 - 'Strong Agree'.

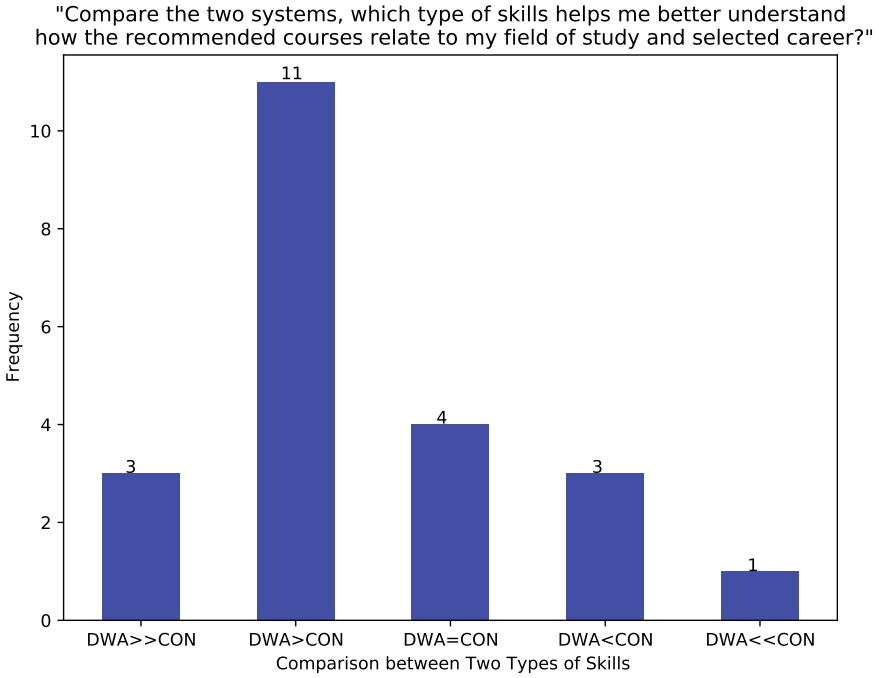


Figure 41: Frequency distribution of skill comparisons between the *DWA* system and the *CON* system, indicating their relevance to recommended courses for specific fields of study and careers.

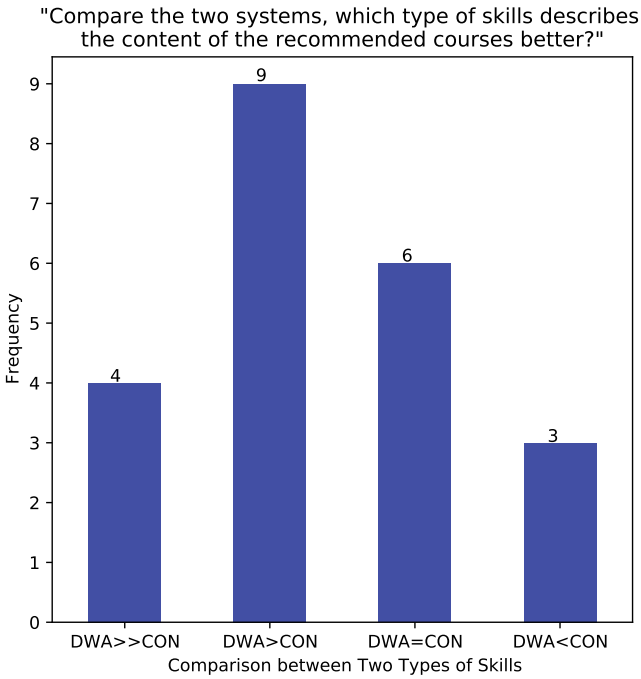


Figure 42: Frequency distribution of skill comparisons between the *DWA* system and the *CON* system in terms of how well they describe the content of recommended courses.

approach may result in less diversity and overlapping semantics. Conversely, DWAs are a diverse set of skills curated by human experts to describe work in the US labor market. To address this limitation, I have proposed an approach (described in Algorithm 1) to diversify the list of explaining concepts. However, the diversity threshold is manually selected and fixed, limiting its effectiveness. To improve the diversity of the concept set, a more rigorous process for fine-tuning the threshold is necessary. Additionally, applying clustering techniques to group similar concepts can create concept clusters that enhance the vector representation, resulting in more effective concept-based explanations.

Skill Quantity. The recommendation explanation aims to connect the skills required for the career to target course recommendations. It features a list of skills that the student has not acquired or mastered. The top ten most important skills are presented to the student along with a description of the courses. To determine the effectiveness of presenting ten skills, participants were asked to rate the number of skills as Too Few, Good, or Too Many. The results in Fig 43 indicate that the majority of participants believe that ten skills are a sufficient number for the explanation. A higher percentage of participants agreed with concept-based explanation (70.8%) than those with DWA-based explanation (66.7%). However, some participants expressed a desire for more concepts, with 8.3% rating the number as too few. This may be due to the fact that DWAs tend to be longer in length, with an average of 6 words compared to 2 for concepts. In the written feedback, the majority of the participants stated that five to ten skills are sufficient for them to evaluate the recommendations, with seven being the most common suggestion. By optimizing the production of our explainable course recommender, I can present the top seven most crucial skills as the default view, while providing students with the option to expand the list for a more comprehensive view. Ideally, the system would be designed to dynamically adapt to the individual preferences of each user, thereby providing a personalized and relevant experience by presenting the most useful information for their unique needs.

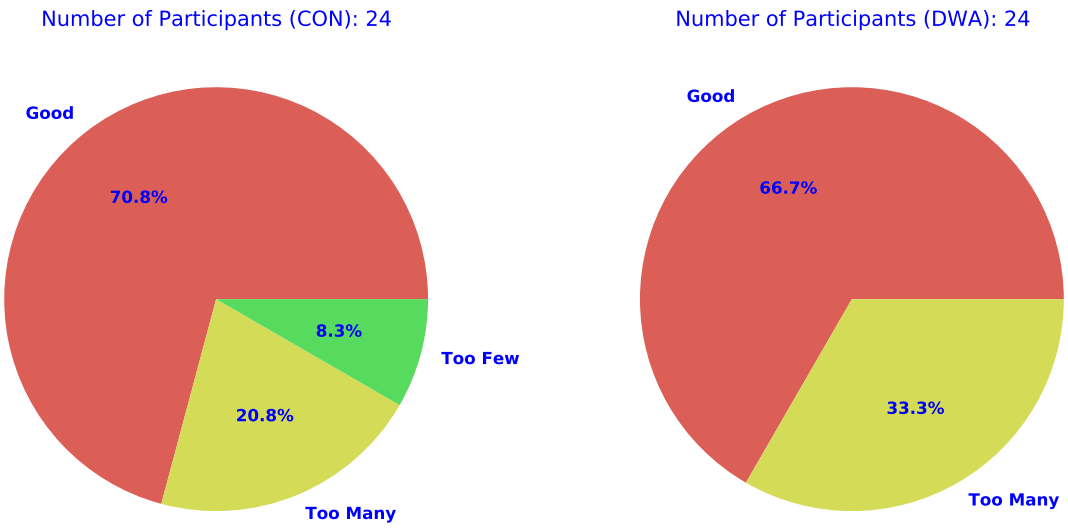


Figure 43: Percentage distribution of user responses to the statement 'The system presents to you a list of 10 skills to explain the recommendations. The number of skills is:', comparing between the *CON* system and the *DWA* system.

7.6 Summary and Discussion

In the preceding chapters, we have developed insights into skill-based explanation and effective methods for modeling courses. Building on this foundation, this chapter introduces a novel personalized and explainable course recommender system to help students explore courses that will provide them with the knowledge and skills required for their future careers. The system recommends courses based on the student's enrollment history and career preferences and provides explanations for the recommendations. I apply the approach that connects college education to job markets presented in Chapter 6, and evaluate the usefulness of job information in course recommendations, and validate the effectiveness of explanation in improving user perception of recommendations. The study is the first of its kind to use actual job information for career-oriented explainable course recommendations using advanced NLP techniques, and it is expected to help users make better decisions and increase their

trust and acceptance of recommendations. This study also compares two different models for representing skills: concepts automatically extracted from course descriptions, as detailed in Chapter 4, and O*NET DWAs manually constructed by experts to describe work in the US labor market, as discussed in Chapter 6. Both recommender models yielded promising results, as indicated by user feedback. Participants generally found the recommendations helpful and expressed interest in taking the recommended courses. For instance, a participant stated, "*The assigned courses were very interesting and intriguing and I will consider taking all of them.*". Additionally, the explanations positively impacted the recommendation. The majority of subjects explicitly agreed or strongly agreed that the explanations helped them determine their level of interest in the recommended courses.

In comparing the two types of skills, the Concept system tended to provide more relevant recommendations to users, whereas the *DWA* system tended to introduce an element of surprise to the recommendations. However, subjects explicitly expressed a preference for the *DWA* system over the *Concept* system, stating that it helped them better understand how the recommended courses related to their field of study and preferred career, and provided a more detailed description of the course content. For example, one participant stated, "*I liked the skill explanations that went into a little more detail, such as those in the DWA system.*" Interestingly, another participant suggested that including both systems could "get the best of both worlds!".

My recommendation and explanation approach has several limitations that need to be addressed for future research. The recommender systems are shown to be able to identify courses that align with students' interests and needs. However, they do not currently account for course level or order, potentially leading to suboptimal recommendations. To address this issue, incorporating prerequisite information can improve recommendations. Advanced recommender engines such as deep sequential models or Transformer-based models like PLAN-BERT have the ability to learn the intrinsic order of courses based on students' enrollment patterns. Moreover, advanced placement credits and equivalent courses (e.g., CMPINF0401 and CS0008) are not considered when generating recommendations; as a result, students receive a course recommendation that is relevant but not useful. Incorporating these features in real-world applications could significantly enhance the quality of recommen-

dations. Secondly, in order to enhance the clarity and effectiveness of the explanations with extracted concepts, the diversification process, as detailed in Algorithm 1, is aimed to provide a diverse set of skills used for explanation by removing concepts that are too similar to those already in the current explaining skill set. However, it's important to highlight that the diversity of the concept set may be constrained by the predefined threshold used in the study's algorithm. To overcome this issue, a more rigorous threshold fine-tuning process can be implemented to increase the diversity of the concept set. Another approach could apply clustering techniques to group similar concepts to enhance the vector representation, leading to more effective concept-based explanations.

There are also several limitations in the user study. It is worth noting a potential limitation related to the sample size in the *between*-subjects experiment, which consisted of only 46 subjects. As such, the study may have lacked sufficient power to justify the real effect of the explanation on the recommendation, at least for the measures of overall *success* and *unexpectedness*. The *within*-subjects conditions, on the other hand, eliminates the *between*-participant variability, thereby fewer participants needed to attain an adequate level of statistical power. Nonetheless, despite randomizing the order in which participants saw the conditions and separating the two sessions three days apart, potential carryover effects may persist. To address these limitations, future research could consider a larger sample size for *between*-subjects studies to provide more robust and reliable results. Lastly, due to constraints on data availability and resources, my study focused only on undergraduate students in the School of Computing and Information. Future research could consider to generalize for all majors within the institute and for graduate students to provide more robust results.

Finally, due to data availability and resource constraints, my study focused only on undergraduate students within the School of Computing and Information. To enhance the generalizability of our findings, it is imperative for subsequent research endeavors to encompass a broader range of majors within the institute, including graduate students. This comprehensive approach will provide a more comprehensive and substantiated set of results.

8.0 CONCLUSIONS, DISCUSSION AND FUTURE WORK

In this chapter, I will begin by summarizing the key findings and conclusions derived from all the studies and analyses. I will then articulate the significant contributions that my dissertation makes to the field. Following this, I will discuss limitations and avenues for future research.

8.1 Summary & Contribution

In this dissertation, I tackle the challenges surrounding academic decision-making and course exploration within higher education, while also addressing the disconnect between learning and work in the U.S. Leveraging the cutting-edge technologies of machine learning and deep learning, I have proposed several approaches for modeling courses and jobs. These approaches serve as the foundation for the development of advanced course guidance and educational information systems, designed to assist students in choosing courses that seamlessly align with their individual interests, abilities, educational trajectories, and career aspirations.

Automatic concept extraction with deep learning. Taking a step toward the ambitious goal of automated educational ontology construction, I have undertaken the development of a concept extraction model for course descriptions, devoid of manual data labeling for training. I approached concept extraction from course descriptions as a sequence labeling task, leveraging state-of-the-art deep learning architectures, namely BERT and BI-LSTM-CRF. These models were trained on various publicly available datasets and subsequently consolidated into an ensemble model for concept extraction, aimed at improving overall model efficacy. Crucially, this final model is flexible to extract concepts from diverse document types, such as course descriptions, without necessitating specialized domain adaptation techniques.

I conducted a comprehensive assessment and comparison of BERT and BI-LSTM-CRF

models for concept extraction, employing widely recognized keyphrase extraction metrics. Remarkably, the stacking ensemble model consistently delivered the highest performance based on the F1 score, regardless of the architectural approach employed. Notably, the BERT ensemble model outperformed the BI-LSTM-CRF ensemble model in this regard. The most superior performance across all metrics was achieved by combining the BERT and BI-LSTM-CRF models. Furthermore, I conducted an expert evaluation to gauge the quality of the concepts extracted, particularly for their application in course recommendation systems. The results indicated that both experts exhibited a high level of consensus in their assessment of the concepts extracted by the model, with a proportional agreement of 92.88%, signifying strong agreement between them. Moreover, the Kappa agreement between the two experts was 0.57, suggesting a good level of agreement. In summary, the study's findings demonstrate the effectiveness of using a combination of BERT and BI-LSTM-CRF models for concept extraction from descriptions in the context of explainable course recommendation systems. The expert evaluation further validates the quality of the extracted concepts, highlighting the practical applicability of these models in the education field.

Explanation for serendipitous course recommendation. I investigated the impact of skill-based explanations on a serendipitous course recommendation system. This system's primary objective is to furnish students with comprehensive insights into available courses, encompassing their alignment with existing knowledge and the acquisition of new skills. This, in turn, empowers students to evaluate a course's relevance more effectively and increases their confidence when making choices. I utilized the trained concept extraction model to extract multi-gram skills from course catalog descriptions. In a collaborative effort with the CAHL lab at the University of California, Berkeley, we embarked on an in-depth exploration of the impact of skill-based explanations within this serendipitous course recommendation system. This investigation was conducted through an online user study at the same institution, utilizing the capabilities of the AskOski system, powered by PLAN-BERT—an advanced deep neural network model, further enriched with a diversification strategy.

Our study represents a pioneering effort in scrutinizing the effect of skill-based explanations on serendipitous course recommendations in higher education. While our overarching findings did not conclusively demonstrate a clear impact of the explanation on PLAN-BERT's

recommendations, they did reveal a noteworthy surge in participant interest in courses that exhibited high levels of unexpectedness. It is evident that individuals who received explanatory information displayed a favorable attitude towards the utility of these explanations in influencing their interest in the recommendations. Furthermore, our research uncovered another crucial aspect: the substantial impact of explanations in bolstering users' confidence in their decision-making processes. Consequently, this reduced their inclination to provide 'neutral' opinions. A detailed statistical analysis illuminated a compelling interaction between participants' major declaration status and the presence of explanations. Specifically, among participants who had not declared a major, the absence of explanations was significantly associated with an increase in their likelihood to express neutral opinions.

Connecting higher education to workplace activities and earnings. I have developed an effective methodology that bridges the gap between the curriculum taught in colleges and universities across the United States and the detailed work activities (DWAs) as defined by the Department of Labor to characterize the American workforce. This innovative approach opens up new avenues for tracking the ever-evolving landscape of higher education and workforce development. For instance, the emergence of DWAs within the syllabi of a field of study (FOS), or major, corresponds to the co-occurrence of DWA pairs across all of academia. These insights can prove invaluable to educators, policymakers, and course recommendation systems as they seek to craft educational programs that align closely with students' career goals.

However, it's important to recognize that not every FOS will cover every skill or ability. This discrepancy arises from the varying labor market demands for specific DWAs, which can differ significantly by industry, region, and employer. Consequently, understanding which course topics correlate with increased or decreased post-graduation earnings can significantly enhance the relevance of educational programs and policies. This, in turn, enhances students' success as they transition into the workforce. For example, academic programs could adapt to include new high-demand skills while reducing emphasis on outdated topics. These insights could revolutionize goal-based learning within course recommendation systems, leading to more personalized and effective recommendations. By identifying relevant topics based on students' predefined objectives, such as maximizing post-graduation earnings, we can proac-

tively equip students with the skills (e.g., "Prompt Engineering") needed to meet the growing demand for roles like Software Engineering or Business Intelligence Analytics in the labor market.

While DWAs offer a means to bridge the gap between work and learning, they also reveal profound distinctions across various courses, fields of study, and universities. For instance, the relevance scores of DWAs have enhanced our ability to predict graduate earnings within many fields, though not universally across all disciplines. It is worth noting that the ONET*DWAs may not provide the most precise taxonomy for characterizing the nuanced knowledge embedded in different courses. This limitation arises, in part, because ONET data primarily serves to delineate workers' attributes rather than higher education programs. The absence of a standardized knowledge base describing more granular concepts and skills in higher education and the labor market underscores the urgent need for future educational research. This research should strive to construct a knowledge base capable of standardizing and enhancing our understanding of how educational foundations influence workforce development and the skills of individuals.

Career-oriented course recommendation with explanation. With compelling results emerging from our course representation methods and the concept extraction model for course descriptions, I have pioneered the development of a personalized and explainable course recommendation system. This system aims to empower students in their journey to explore courses that align with their future career aspirations, equipping them with the requisite knowledge and skills. By leveraging students' enrollment history and career preferences, our innovative system provides tailored course recommendations while also offering comprehensive explanations for these suggestions.

This study marks a significant milestone as it leverages real job-related data to enable explainable course recommendations through advanced Natural Language Processing techniques. This approach is expected to enhance user decision-making, instill trust, and foster acceptance of the recommendations. Notably, my research also explore the comparison of two distinct skill representation models: concept-based (*Concept*) and O*NET DWAs (*DWA*). Feedback from participants underscores the system's efficacy, with users expressing the value of the recommendations and a keen interest in enrolling in the suggested courses. Further-

more, the provision of explanations has been a key driver of recommendation acceptance. A substantial majority of respondents explicitly agreed that these explanations aided them in assessing their interest in the recommended courses. When comparing the two skill representation models, the Concept system excels in offering more relevant recommendations, while the DWA system introduces an element of surprise. Remarkably, despite this difference, participants voiced a preference for the DWA system. They found it to be more informative in illustrating how recommended courses align with their academic and career goals, providing a detailed course content overview.

Contributions. In summary, my thesis makes valuable contributions to multiple research domains within the fields of computational social science and artificial intelligence in education. These contributions encompass the interconnection of higher education and graduate career pathways, the extraction of knowledge, the provision of course recommendations, and the generation of insightful explanations.

- Leveraging cutting-edge deep learning architectures, I have developed a concept extraction model for educational documents (e.g., course catalog descriptions). This endeavor, aimed at automating the construction of educational ontologies, could enhance course guidance and educational information systems, and standardize insights into how educational foundations shape workforce development and the skills of workers. The evaluation results demonstrate the model's ability to efficiently extract concepts from course descriptions and highlight the quality and reliability of the extracted concepts, confirming the practical applicability of these models in the field of education. Notably, this model has successfully been employed to extract skills for two distinct explainable course recommendation systems in this thesis.
- By using the trained concept extraction model to extract multi-gram skills from course catalog descriptions, I contribute to a pioneering effort aimed at investigating the impact of skill-based explanations on a Transformer-based course recommendation system in higher education. The outcomes of our user study highlight a substantial increase in participant interest in courses that exhibited high levels of unexpectedness. Participants displayed a positive attitude towards the value of these explanations in shaping their interest in the recommendations. Additionally, the research showed a pivotal insight:

the influential role of explanations in bolstering the confidence of participants who had not yet declared a major in their decision-making process.

- This thesis represents a pioneering effort in bridging the gap between workplace activities and higher education. We analyze a large, novel, extensive dataset comprising over one million syllabi from more than eight hundred bachelor's degree-granting institutions in the United States to establish connections between the curriculum taught in higher education and the detailed work activities outlined by the US Department of Labor. Our unified information system connecting workplace skills to the skills taught during higher education holds the potential to enhance the workforce development of highly-skilled individuals, provide valuable insights into educational programs and course recommendation systems regarding future trends, and enable employers to quantify the skill profiles of potential candidates.
- Finally, through our course representation methods and the concept extraction model for course descriptions, I have led the way in developing a personalized and explainable course recommendation system, with a focus on career alignment and skill-based explanations. This innovative system is designed to empower students on their educational journey by helping them discover courses that align with their future career goals while providing them with the necessary knowledge and skills. My research represents a pioneering effort in the field, as it is the first to utilize real job information for career-focused, explainable course recommendations. The results from the user study have been highly promising, as indicated by positive user feedback. These findings pave the way for future research endeavors aimed at harnessing job data and skill-based explanations to enhance course recommendations within higher education.

8.2 Discussion, Limitations & Future Work

My dissertation inevitably presents certain limitations due to constraints in methodology, resources, and current technological capabilities. Acknowledging these limitations is crucial not only for providing a comprehensive understanding of the study's scope but also for guid-

ing future research directions. In this section, I outline the primary limitations encountered during this research and provide recommendations for future work. These insights aim to pave the way for subsequent studies to further build upon our findings for improvement and investigate deeper into areas not yet explored.

Automatic knowledge extraction. Over the past few decades, it has consistently been demonstrated that skills play a vital role in numerous educational AI systems. Not only skills help recommender systems in making sound recommendations but also display skills is one of the most intuitive ways to explain the documents. The lack of a standardized knowledge base for higher education and the job market underscores the need for one to better understand how education influences careers and individual skills. In this thesis, I have embarked on the development of a cutting-edge concept extraction model for educational documents. Employing state-of-the-art deep learning architectures, I have conducted a comprehensive evaluation of the model's capability to efficiently extract concepts from course descriptions. The evaluation results robustly demonstrate the model's effectiveness in this task, shedding light on the quality and reliability of the extracted concepts. This affirmation solidifies the practical applicability of these models within the field of education. Nonetheless, it is important to note that this achievement represents only the initial stride toward the ultimate objective of automating the construction of educational ontologies.

The construction of educational ontologies presents a multifaceted challenge, demanding not only the recognition of concepts within textual documents but also the extraction of relationships, the construction of hierarchical structures, and the disambiguation of concepts [66, 67]. With the emergence of Large Language Models [200, 201], there is a substantial promise for advancing the field. These LLMs exhibit remarkable capabilities in generalization across a wide range of tasks through multi-task training and unified encoding, underpinned by their comprehensive understanding of linguistic intricacies, semantics, and contextual nuances. The vast training data that LLMs are exposed to empowers them to identify and extract information with higher accuracy and contextual relevance, such as ChatGPT [1]. Additionally, with the advent of instruction fine-tuning techniques, recent studies [202, 203, 204] have shown significant progress in information extraction using LLMs.

[1] https://openai.com/blog/chatgpt

This progress underscores the potential of LLMs to play a pivotal role in automating the construction of educational ontologies. Consequently, further research and exploration of LLMs in this context hold the promise of transforming the task of automatic educational ontology construction from aspiration to reality.

Connecting college education to career through detailed work activities. In my dissertation, I have successfully presented a proof-of-concept that harnesses the power of innovative syllabus data and cutting-edge natural language processing (NLP) techniques. This approach facilitates the crucial connection between labor market data and higher education by effectively predicting the evolution of skills taught within a Field of Study (FOS) and establishing a correlation between Detailed Work Activities (DWAs) and graduate earnings. To extend and strengthen the findings of this study, there is a rich landscape of potential future research avenues. One notable direction is the exploration of causal relationships, particularly with regard to skill-level adjustments influencing course content. It is essential to acknowledge that my current study cannot fully address the selection bias inherent in students' university enrollment choices. However, future investigations could leverage natural experiments to overcome this limitation. These experiments might include analyzing the effects of significant events such as the appointment or retirement of faculty members, the creation of a new academic department, the rise of a prominent employer due to fiscal incentives, or substantial donations targeting specific educational results that could serve as potent avenues of exploration.

Additionally, my investigation into the College Scorecard earnings data is confined to merely two graduation cohorts. Likewise, the Post-Secondary Employment Outcomes data encompasses only a select number of institutions. The scope of my research only includes earnings up to a year post-graduation, potentially missing out on longer-term career progressions [196]. A more comprehensive approach would be to integrate workers' resume data, forging direct links between workers' educational foundations during college and their subsequent career trajectories. Such an approach would encapsulate dimensions like worker adaptability, job tenure, mobility, and more. Another exciting avenue for future exploration lies in the analysis of job postings. Comparing employer demands to the DWAs identified in our study can help pinpoint the educational programs that are most or least adaptive. This

comparative analysis, as demonstrated in previous studies (e.g., [24]), promises to shed light on the evolving landscape of high-skilled workers and their impact on job polarization and urbanization [26, 174, 185].

In my research, I have also demonstrated the presence of discernible variations in graduate earnings at the cohort level, highlighting the influence of skills imparted through educational courses. My methodology has primarily concentrated on assessing the outcomes for various groups of graduates, such as those categorized by their majors or universities. While this approach provides valuable insights, future research could delve deeper into the nuances of labor market outcomes at the individual level. For instance, within the same major, students often select different courses, thereby acquiring distinct skill sets. This variation in course selection may, in turn, lead to divergent career paths and income levels. The intriguing questions that remain unanswered revolve around the extent to which individual course choices shape occupational trajectories and earnings, and the degree to which acquired skills contribute to the observed variations in career outcomes. These are compelling areas for further investigation. However, it's important to acknowledge that conducting such research presents certain challenges. One significant obstacle is the availability and accessibility of datasets suitable for studying individual-level outcomes, as privacy concerns surrounding personal data continue to grow. Overcoming these privacy barriers will be crucial for advancing our understanding of these issues. Furthermore, my analyses have primarily focused on graduates with bachelor's degrees. Future research could extend its scope to explore the skills cultivated during graduate education or even investigate the influence of undergraduate education on the path to graduate school admission. By broadening the scope of inquiry in this way, we can gain a more comprehensive understanding of the relationship between education, skills, and career outcomes at all levels of academic attainment.

Skill-based explanation for course recommendation. The central focus of my thesis centers on the enhancement of course recommendation systems in higher education through the incorporation of explanations. The discussion of my first study highlights several limitations that warrant attention in future research. Although PLAN-BERT has proven its effectiveness in leveraging past sequence information, and the incorporation of user and item features to generate recommendations, it's essential to recognize that our strategy for

diversifying the recommendation list, applied on top of PLAN-BERT's output, is relatively simplistic. We currently limit recommendations to one course per department, which may lead to the inclusion of irrelevant course suggestions. In the context of academic course offerings, some departments naturally share stronger connections with related disciplines, while others may have fewer connections. To address this concern, future studies should consider relaxing this constraint and establishing a relevance threshold for course recommendations. If a course from a department fails to meet this threshold, multiple courses from the same department can then be recommended.

Another avenue for enhancing recommendation diversity and optimizing for serendipity is to frame the problem as a multi-task/multi-label optimization problem during the recommender system's training phase [171, 172]. This approach allows the recommendation engine to simultaneously optimize for both relevance and unexpectedness, striking a balance and effectively constraining them to achieve a unique, non-dominated solution. To implement this, collecting labels for the unexpectedness of relevant courses is imperative for training the models.

Furthermore, although this study did not conclusively establish the influence of providing explanations on the course recommendation systems, it leaves open the possibility that offering comprehensive insights into why specific courses are suggested and how they align with students' abilities and interests could enhance students' comprehension of the value of these recommendations. The effectiveness of such explanations may vary depending on a student's academic field and stage of progression. Further exploration of these factors has the potential to refine the system's design. Notably, our study did identify a positive impact of explanations on generating interest in unexpected courses, particularly among students who have not yet declared a major. Future research efforts may prioritize this demographic to expand the sample size and draw more robust conclusions.

In this study, I chose to exclusively employ skills extracted from course catalog descriptions as the primary knowledge components for course representation in our explanations. While this approach is conventional for content-based methods and has been employed in numerous prior studies, an alternative approach involves representing courses based on their relationship with individual skills. This alternative approach has demonstrated its effec-

tiveness and yielded promising results, as elaborated in Chapters 6 and 7. Notably, it has proven robust across various skill taxonomies, including O*NET DWAs and extracted concepts. Consequently, the exploration of this course modeling approach, along with experimentation with various skill types for generating explanations in serendipitous course recommendations, stands as a promising avenue for future research.

My second study represents a pioneering effort within the field, leveraging real job-related data to facilitate career-oriented course recommendations and explanations. By doing so, I aim to bolster the adaptability of serendipitous course recommendations, thereby fostering a more holistic course exploration experience for students. The promising results derived from my user studies lay a strong foundation for future research endeavors in this domain. However, it is essential to acknowledge several limitations identified in my first course recommendation study, which must be addressed in subsequent research. While the recommender systems demonstrated the capacity to identify courses aligning with students' interests and needs, they fell short in accounting for the course level or sequence, potentially resulting in suboptimal recommendations. To rectify this issue, integrating prerequisite information into the recommendation process holds the potential to significantly improve the quality of recommendations. Advanced recommender engines, such as deep sequential models or Transformer-based models, possess the capability to discern the intrinsic course order based on students' enrollment patterns. Additionally, overlooking advanced placement credits and equivalent courses in the recommendation process can lead to relevant but ultimately unhelpful recommendations. Incorporating these features into real-world applications could substantially enhance recommendation quality.

Furthermore, the threshold fine-tuning process employed in my study might have limited the diversity of the concept set for explanation. To overcome this limitation, a more rigorous threshold fine-tuning process could be implemented to increase the diversity of the concept set. Alternatively, employing clustering techniques to group similar concepts could enhance vector representations, resulting in more effective concept-based explanations. This underscores the significance of the educational ontology discussed earlier, as it can play a pivotal role in improving various downstream applications, including but not limited to course recommendation and explanation.

Another limitation of my research lies in the sample size of the between-subjects experiments, consisting of only 46 subjects for the first study and 53 subjects for the second study. This constraint may have affected the study's statistical power, potentially obscuring the true impact of explanations on the recommendation. To mitigate this limitation, future research endeavors should consider a larger sample size to yield more robust and reliable results. It is also important to note that my studies primarily focused on the use of explanations to enhance recommendation effectiveness. Future studies can explore the utility of explanations in other dimensions, such as *Transparency*, *Trust*, *Persuasiveness*, and *Satisfaction* [110, 107, 109]. Investigating these aspects could broaden our understanding of the potential impact of explanations in the realm of recommender systems, offering a more comprehensive perspective on their influence.

Lastly, while user experiments that involve collecting feedback through questionnaires are a prevalent technique for assessing the user experience in recommender systems [170], they come with inherent limitations. The subjective nature of user preferences, influenced by mood, personal biases, or external factors, can distort the data, leading to a misrepresentation of their actual preferences. Moreover, human preferences are often complex and multidimensional, which might not be adequately captured by simplistic questionnaire responses. Additionally, users may struggle to express their preferences accurately, resulting in incomplete or misleading data. There's also the tendency for users to answer questionnaires in a manner they perceive as expected, favorable, or socially acceptable, rather than truthfully, especially in cases where subjects claim to like suggested courses but don't follow through with enrollment. To enhance the evaluation method, tracking students' course enrollment records could provide a more accurate measure of the effectiveness of recommendations. This approach, however, comes with its own set of challenges, including privacy concerns and other factors influencing students' decisions, like schedule conflicts or course demand. Despite the limitations of the current evaluation approach, introducing students to interesting and relevant courses still can be beneficial, as it informs them and aids in making more informed decisions regarding course selection.

Large language models. In this dissertation, I apply natural language processing methods across all the studies, encompassing aspects of document representation, knowl-

edge extraction, recommendation systems, and explanation generation. In its early stages, Large Language Models (LLMs), most notably ChatGPT developed by OpenAI, have played a transformative role in the technology industry and research landscape.[2] Their profound impact stems from their extraordinary capacity for generalization across a diverse spectrum of tasks, achieved through multi-task training and unified encoding. This capability is grounded in their comprehensive understanding of linguistic intricacies, semantics, and contextual nuances. It is clear that leveraging the power of LLMs, particularly through effective prompting and instruction fine-tuning, holds immense potential for enhancing the methods employed in this research. LLMs can be strategically integrated into intermediate tasks to improve downstream tasks. Alternatively, they can be directly employed for downstream tasks, such as providing user and item details and prompting Language Models to create explanations. One notable avenue of improvement lies in leveraging embeddings generated by LLMs, such as the OpenAI embeddings [205], which have shown promise in enhancing vector representations [206]. These augmented representations offer significant benefits in the domains of course, job, and skill analysis. They can contribute to more accurate skill detection and comparison, facilitate relationship calculations, and ultimately enhance the functionality of educational information systems. Moreover, they can help elevate the quality of course recommendations and the overall explainability of the system. Furthermore, the potential of LLMs extends to aiding educational ontology development by either curating training datasets or directly querying LLMs for relationship determinations. For instance, Figures 44 & 45 demonstrate using ChatGPT 4.0 for concept relation detection. Such strategies could be pivotal in refining the skill diversification process, leading to more nuanced, skill-centric explanations in higher education course recommendations.

In essence, the incorporation of LLMs, like ChatGPT, holds immense potential for advancing the field of natural language processing and its applications, particularly within the scope of this dissertation. As these models continue to evolve, they are poised to play an increasingly pivotal role in shaping the future of technology and research.

[2]https://hbr.org/2023/06/discovering-where-chatgpt-can-create-value-for-your-company

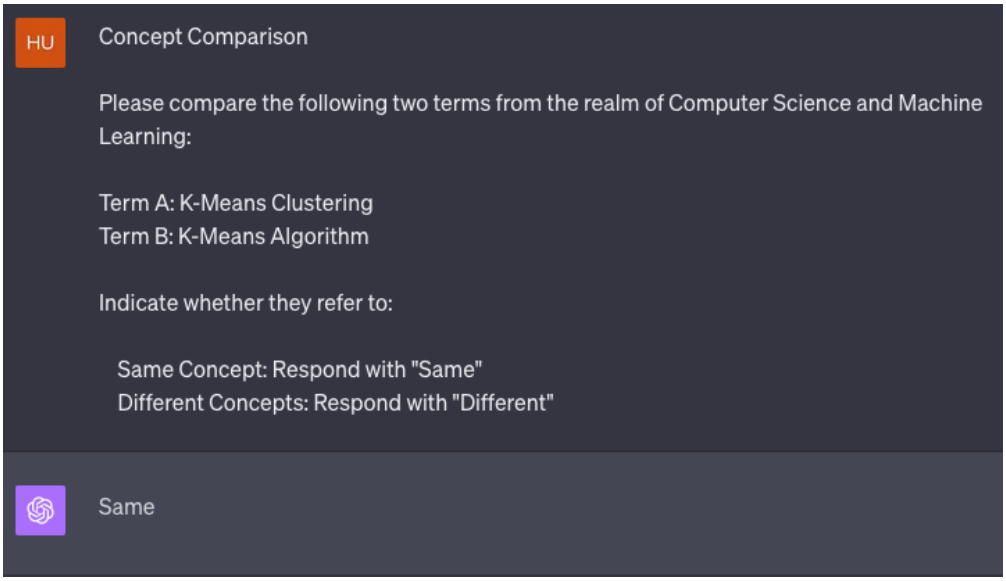


Figure 44: This example demonstrates using ChatGPT 4.0 interface for concept relation detection *without* explanation.

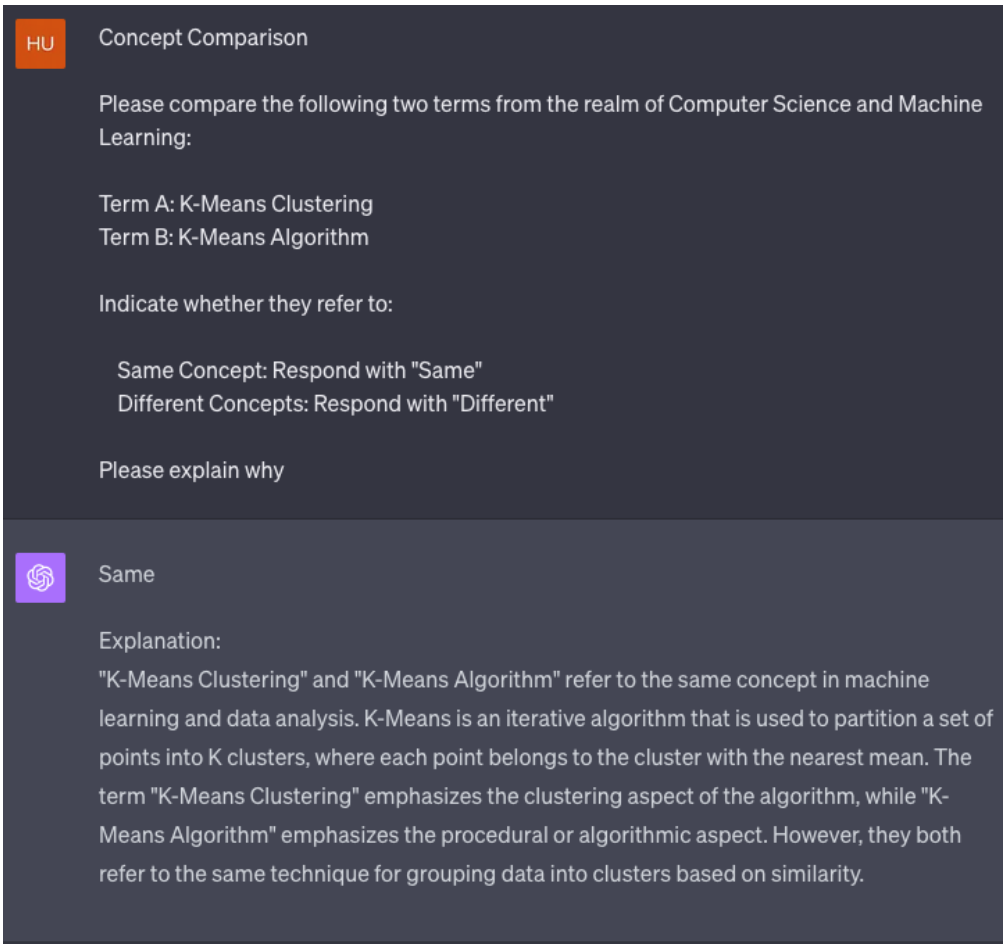


Figure 45: This example demonstrates using ChatGPT 4.0 interface for concept relation detection *with* explanation.

Appendix A Connecting Higher Education to Workplace Activities and Earnings

A.1 OSP Data Processing

The OSP dataset contains roughly 3 million syllabi in the US. Each syllabus has multiple attributes. These include: a unique syllabus ID, the probability it is a syllabus (this is due to the automated process with which syllabi were originally scraped), the year, the FOS (and corresponding level of certainty of being that FOS), the institution, the location of the institution (latitude/longitude), the language, as well as metadata on the process (e.g., method used to collect the syllabus info). Our analysis only used a small subset of these, hence the potential for significant further study.

A.1.1 The Identification of Course Descriptions

There is a multi-stage process to identify course descriptions from the raw syllabus data. To begin, the text associated with each syllabus includes multiple elements. Some are useful (e.g., course objectives, course descriptions, outline of the class), and others are less so (e.g., office hour times, contact details, administrative information). The first task is to keep the former (i.e., the elements that contribute to an understanding of the content taught in the class), and to exclude the latter.

Each syllabus text in the dataset is structured with multiple headings, including "Overview", "About the course", "Course content", "Description", "Course outline", "Outcome", "Objective", "Aim", and "Goal". We begin by splitting the syllabus text into these groups (using roughly a dozen 'useful' headings), so that we end up with text under each such heading. After some processing (e.g., removing duplicate text and pieces of text that are too similar: we use *SequenceMatcher* from the *difflib* (Python) library and do not include text that has a similarity ¿ 0.8 with another already-added text). The concatenation of these groups forms the "course description" of the syllabus. Note, this method is imperfect due to the noise and

complication of the format texts in OSP: there is text that is not being captured and likely small amounts of irrelevant administrative information included. However, our hypothesis is that the administrative text (e.g., "Office hours are between 3-5pm on Thursdays") is 'neutral' and does not influence the DWA analysis that follows. There is clearly scope for improvement with this process with more advanced natural language processing techniques, but we do believe the outcome of this process is efficient for our following analyses and do not cause misleading results.

A.1.2 The Language Embeddings to Compute DWA Syllabus Similarity

Once the text has been cleaned, the next process is to generate a vector to represent each DWA, using a Wikipedia-trained word embeddings (*fasttext-wiki-news-subwords-300*) in Gensim language models[1].

As the first step, we tokenize each DWA in the full list of 2070 DWAs. This tokenized list is used to create a (*genism.corpora*) dictionary. Then, we take that dictionary of DWAs to run bag-of-words (BOWs) on the tokenized syllabus text. We also generate the (*genism.models*) 'word embedding similarity index' from the pre-trained embeddings we used (trained on Wikipedia pages). Then, we create a sparse term similarity matrix between the language model similarity index and the dictionary based off of DWAs. We generate one final index by taking the *soft cosine similarity* between that similarity matrix and the bag-of-words representation of DWAs.

Once the initial processing has occurred, we move to analyzing each syllabus text in turn. First, we take the processed syllabus text and tokenize it a BOWs representation (removing words like *days of the week*, *months*, and *common words* like "http", "hour", "assignment", "college", "university", "emails", etc.). Using the final generated similarity index, we calculate the *soft cosine similarities* [190] between the BOWs representations of the DWAs and the syllabus. As a result, each syllabus is represented as a vector of 2070 dimensions, showing how strongly each of the 2070 DWAs is associated with the course description. With this representation of course syllabi, we now can easily compute the

[1]https://github.com/RaRe-Technologies/gensim-data

relationship between each pair of course syllabi, representations of FOS and universities and so on.

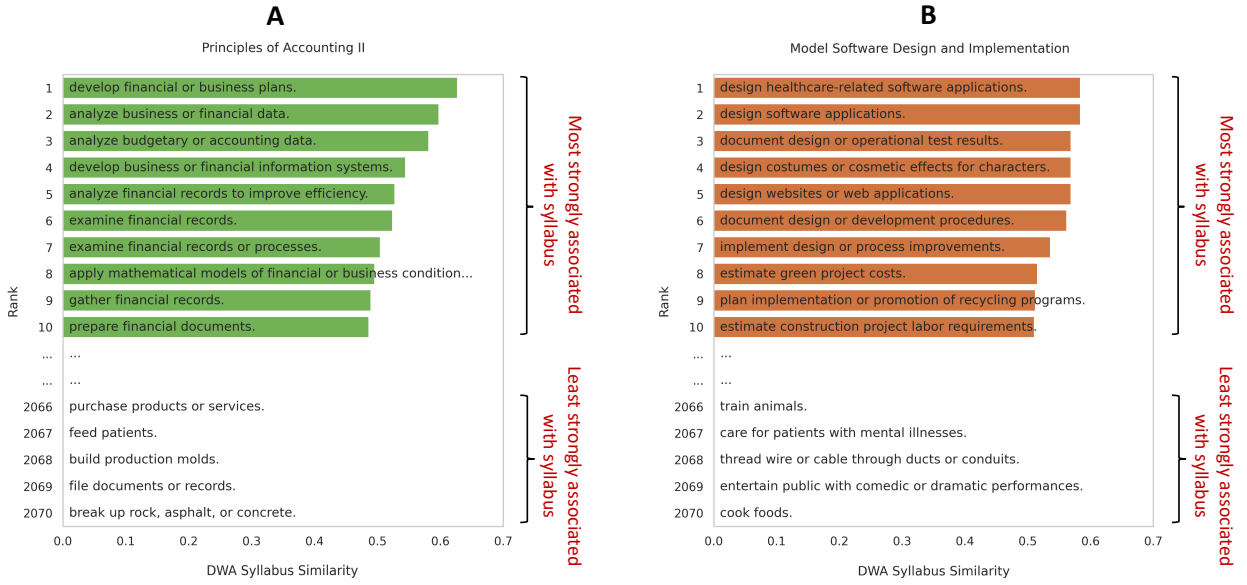


Figure 46: (A) An example accounting syllabus and the activities that are most and least strongly associated with its course description; and (B) An example computer science syllabus and the activities that are most and least strongly associated with its course description. The course description and learning objectives are extracted and embedded into a pre-trained language space. DWA syllabus similarity scores (from *0* to *1*) are calculated for each detailed workplace activity against the syllabus.

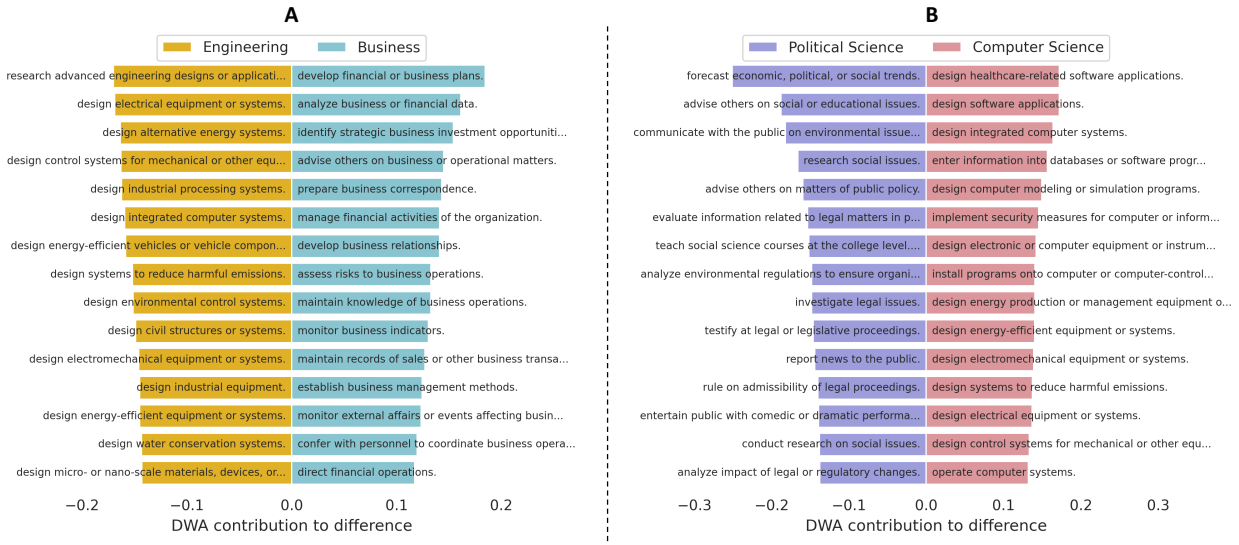


Figure 47: (A) The DWAs that most significantly distinguish Engineering syllabi from Business syllabi. (B) The DWAs that most significantly distinguish Political Science syllabi from Computer Science syllabi.

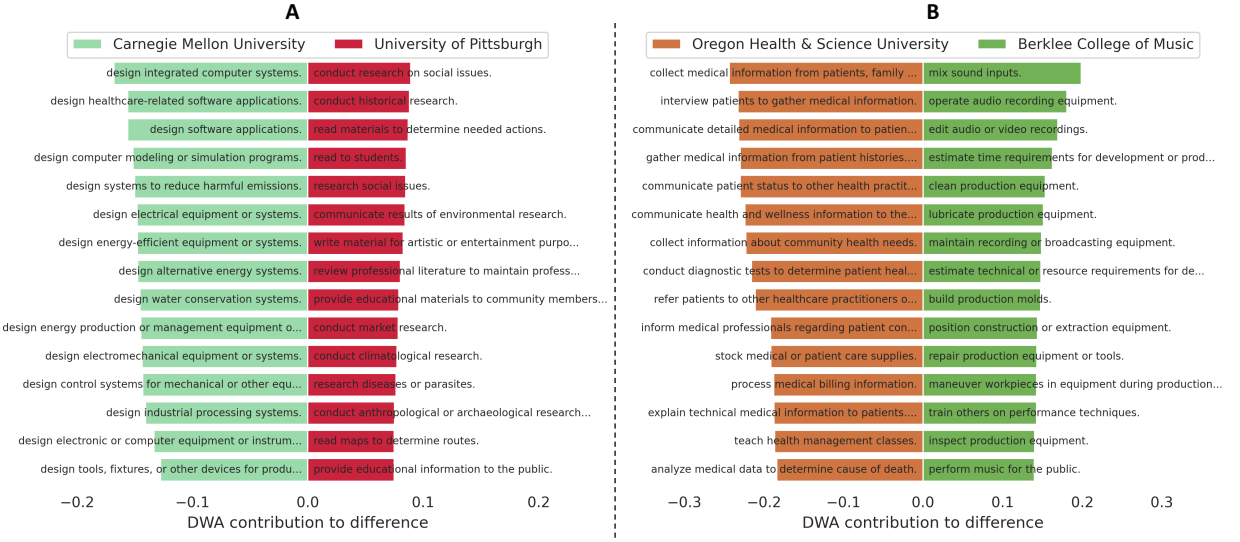


Figure 48: (A) The DWAs that most strongly separate Carnegie Mellon University syllabi from University of Pittsburgh syllabi. (B) The DWAs that most strongly separate Oregon Health & Science University syllabi from Berklee College of Music syllabi.

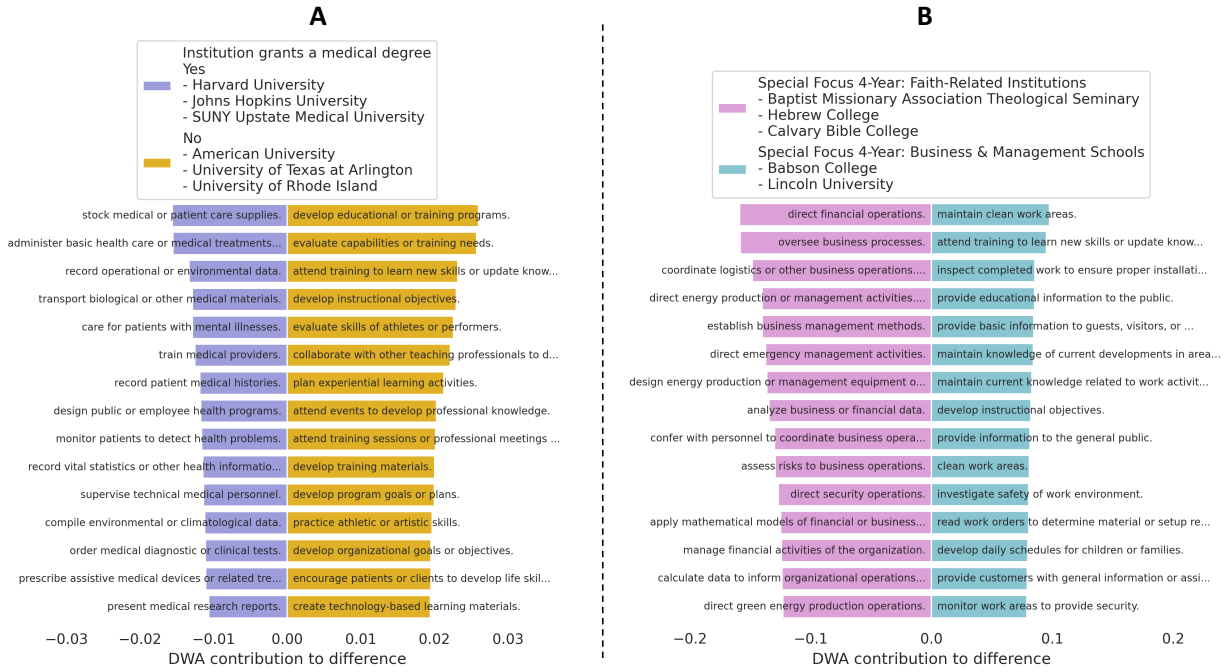


Figure 49: (A) The DWAs that most strongly separate Medical Degree-Granting Schools syllabi from Non-Medical Degree-Granting Schools syllabi. (B) The DWAs that most strongly separate Special Focus 4-Year Faith-Related Schools syllabi from Business & Management Schools syllabi.

A.2 Distance Metric Correlation

To calculate DWA relationships, we experiment two different methods: (1) *Direct* - compute directly the cosine similarity of the embedding vectors; and (2) via course syllabi - based on the co-occurrence of dwa_1 and dwa_2 in course syllabi. For the second method, we experiment with four different similarity and distance metrics: *Cosine* similarity, *Euclidean* distance, *Manhattan* distance and *Jaccard* similarity. Figure 50 shows the correlations of these methods and distance metrics.

	Direct	Cosine	Euclidean	Manhattan	Jaccard
Direct	1.000000	0.520041	-0.302950	-0.278611	0.493206
Cosine	0.520041	1.000000	-0.214060	-0.225266	0.894886
Euclidean	-0.302950	-0.214060	1.000000	0.990004	-0.460567
Manhattan	-0.278611	-0.225266	0.990004	1.000000	-0.491437
Jaccard	0.493206	0.894886	-0.460567	-0.491437	1.000000

Figure 50: The correlation matrix of the two methods and four distance metrics to calculate DWA relationships.

A.3 Predicting Educational Trends

A.3.1 Comparing Distance Metrics

For robustness checks, we run Models 2, 3, 4 & 5 (explained in the main manuscript) with two different methods and four distance metrics for computing the DWA relationships. The first method (called *Direct)* computes the DWA relationships by directly measuring the Cosine similarity of their language embedding vectors. This approach measures a static relationship between DWAs and can not distinguish the dynamics of how one DWA relates to another locally (i.e., within a FOS or a university) and globally (i.e., across all of academia). The second method calculates the relationship between each pair of DWAs based on the co-occurrence of the two DWAs in course syllabi. The relationships can be measured locally as well as globally. We experiment with four different distance metrics for the second method: Cosine similarity (*Cosine*), Euclidean distance (*Euclidean*), Manhattan distance (*Manhattan*) and Jaccard similarity (*Jaccard*).

Figure 51 and 52 show the RMSE and R-Squared performance comparisons of these methods and distance metrics for each of the models involving inter-DWA relationships. As can be seen from the figures, *Jaccard* performs best consistently across all the models. Second from the best is *Cosine* similarity metric. *Direct* method fails to distinguish the dynamics of how one DWA relates to another locally and globally; as the results, for the best model (i.e., Model 5), it performs worst among all the variations. On the other hand, the global relationships captured by *Manhattan* and *Euclidean* help them surpass *Direct* performances.

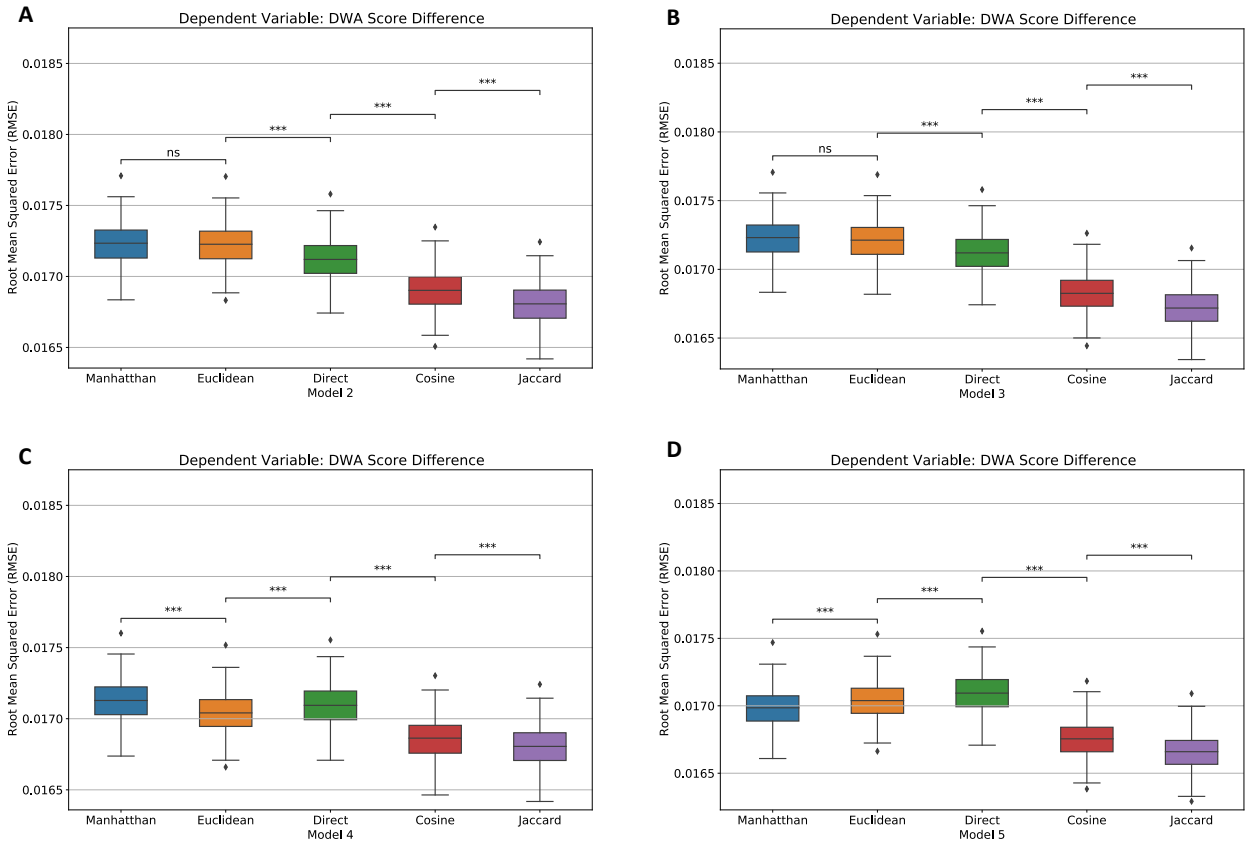


Figure 51: **Workplace activities detected from syllabi predicting teaching dynamics within a field of study and earnings of college graduates**. We perform 5-fold cross-validation and repeat 40 times (i.e., 200 trials in total) for each model and measure RMSE by the resulting models applied to the test set. Asterisks indicate the statistically significant difference between the two models' performances with Bonferroni correction. (A) Predicting the importance of DWAs changing in 10 years (2008 vs. 2017). (A), (B), (C) and (D) show the performance comparisons of different distance metrics calculating DWA relationships for Models 2, 3, 4 and 5, respectively.

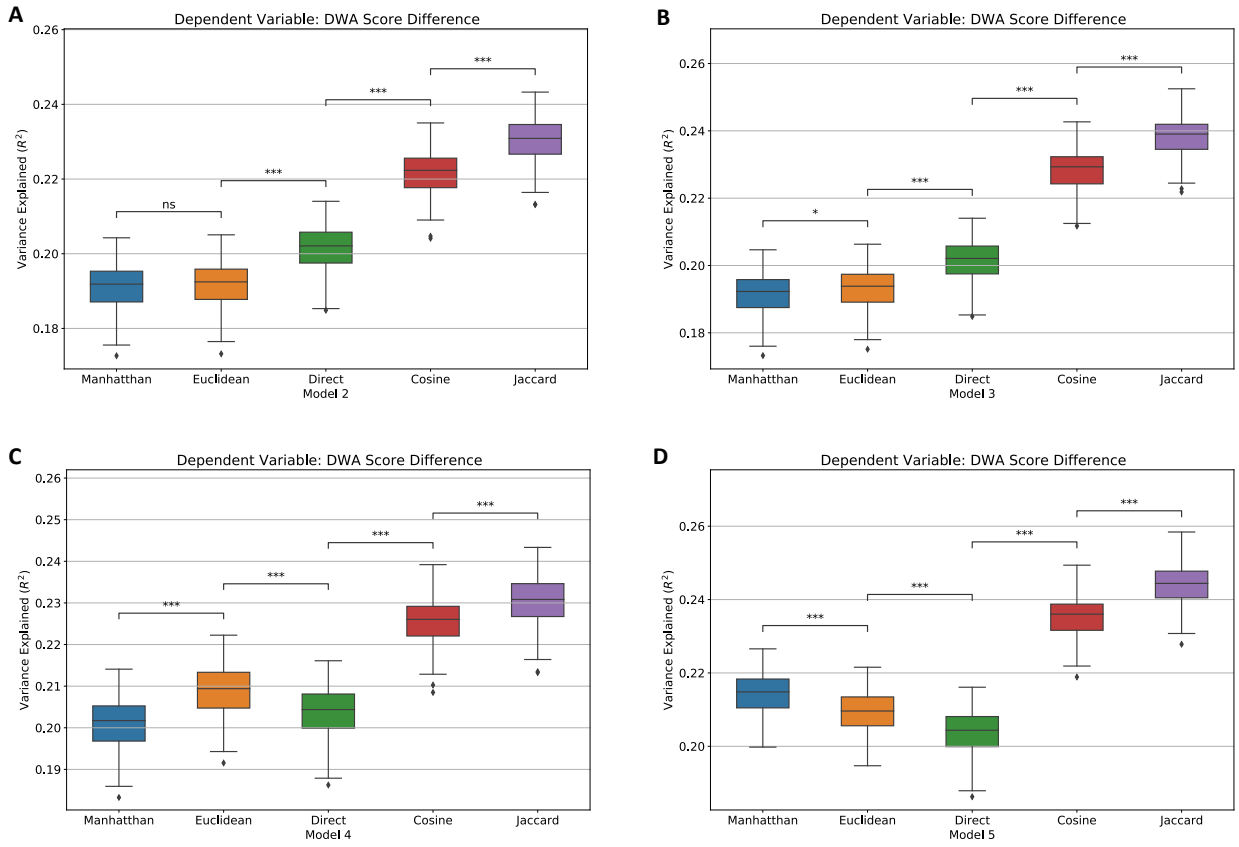


Figure 52: **Workplace activities detected from syllabi predicting teaching dynamics within a field of study and earnings of college graduates**. We perform 5-fold cross-validation and repeat 40 times (i.e., 200 trials in total) for each model and measure the variance explained (i.e., R^2) by the resulting models applied to the test set. Asterisks indicate the statistically significant difference between the two models' performances with Bonferroni correction. (A) Predicting the importance of DWAs changing in 10 years (2008 vs. 2017). (A), (B), (C) and (D) show the performance comparisons of different distance metrics calculating DWA relationships for Models 2, 3, 4 and 5, respectively.

A.3.2 Classification Analysis

In addition to the regression analysis for educational trends presented in the main text, we perform a classification analysis for this problem. The task is to predict which DWAs become "important" in the future, meaning those DWAs are not considered important at the current time but potentially are important in future (10 years later). This helps to understand how fields of study evolve over time, enabling proactive course design by educators and informing educational incentives from policy makers.

"Important" DWAs are the DWAs that are the most prevalent ones for a FOS. DWA (a) is labeled as "important" in a FOS (f) when it satisfies the condition below:

$$r_f(dwa) >= \mu_f + 2 * \sigma_f \tag{23}$$

Where μ_f and σ_f are the mean and the standard deviation of the relationships between the DWAs and the FOS, respectively. On average, there are around 59 and 58 "important" DWAs per FOS in 2008 and 2017, respectively. The number of DWAs that are important in 2017 but not important in 2008 is 15. These DWAs are positive labels in our classification analysis.

We build a logistic regression model to classify whether a DWA is important (1) or not (0). We use the information about the current propensity score of the DWA and its relationships with currently important DWAs calculated with the Jaccard similarity metric. Based on the principle of relatedness [183], our assumption is that skills co-taught with currently important skills are likely to become more important in future. To evaluate how well the model performs, we report the ROC curves and AUC scores for each individual FOS (see Figure 53). Since there is an unbalance in number of data points in the two classes (0 vs. 1), we, in addition, measure the model performance in terms of *precision*, *recall* and *F1-score* at top N (see Figure 54).

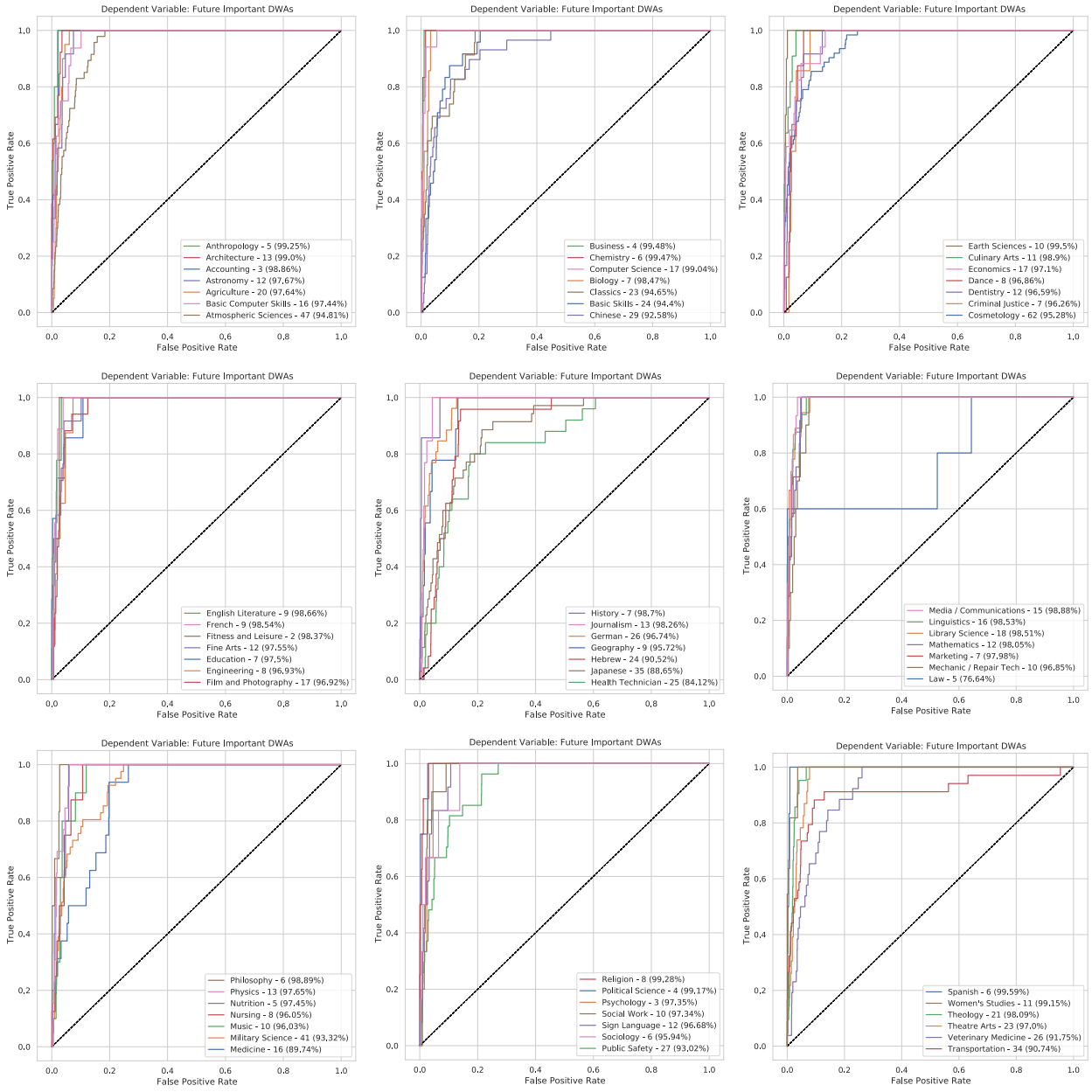


Figure 53: ROC curves of the important-DWA classification model for each individual FOS. The legends display the field name, the numbers of important DWAs, and the AUC scores.

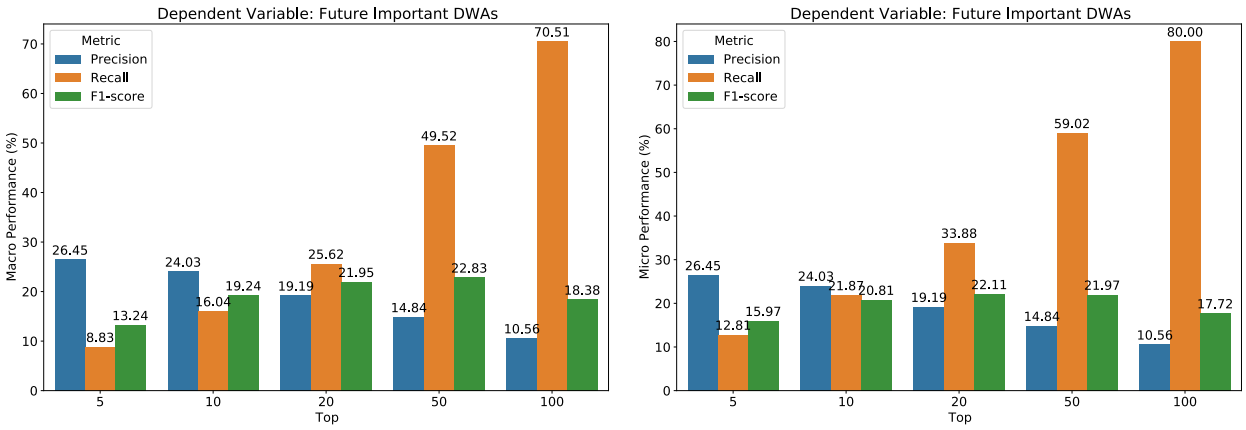


Figure 54: Precision, recall and F-scores of the important-DWA classification model at top *N*. Macro performance is calculated when considering the prediction for all FOS together; while micro performance is the average performance of each of individual FOS.

A.4 Selection of Graduate Earnings Records

College Scorecard Earnings provides transparency and consumer information related to individual institutions of higher education and individual fields of study within those institutions. We only process earnings records for Baccalaureate colleges and universities. We map College Scorecard CIP codes to OSP CIP codes. As a result, each earnings record includes the field name and institution information. There are 9007 graduate earnings records in 54 fields-of-study at 832 institutions.

To understand how workplace activities extracted from course syllabi contribute to the earnings of graduates. We aggregate DWAs from the course syllabi taught in individual academic fields at specific institutions. Those DWAs are the features to predict graduate earnings presented in the main text. Though large, the OSP course syllabus data is not distributed evenly across fields-of-study and institutions. Some fields and institutions have much fewer course syllabi. Thus, to sufficiently estimate work activities taught in a FOS at a university, we limit earnings records for FOS (in an institution) that have at least 10 course syllabi. As a result, we obtain 2872 earnings records in 47 FOS at 347 institutions. Furthermore, we select FOS that have at least 30 earnings records across institutions for prediction tasks, resulting in the remaining 2601 earnings records in 26 FOS at 343 institutions (see Table 11 for details of numbers of observations of FOS in our analysis before and after filtering).

It is possible that a subset of earnings records of a FOS does not effectively represent the distribution of the entire population. We perform the Kolmogorov-Smirnov (KS) statistical test to make sure the remaining earnings records representative for the entire population of the field at the institute. If the remaining earnings observations pass the KS test (*p-value* > 0.05), we will keep that FOS for the earnings prediction analyses. As an example, Figure 55 plots the distribution of median earnings of graduates in *Business* against a number of syllabi. Table 11 shows the *p-values* of the KS test for the FOS in our "*Within Field-of-Study Skill Variation and the Earnings of Recent College Graduates*" analysis in the main text.

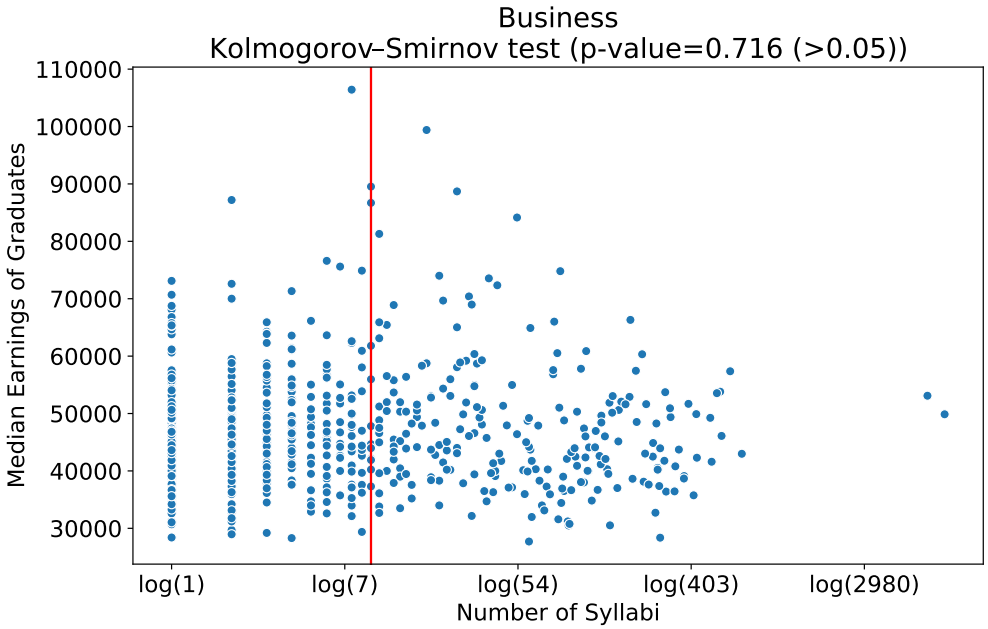
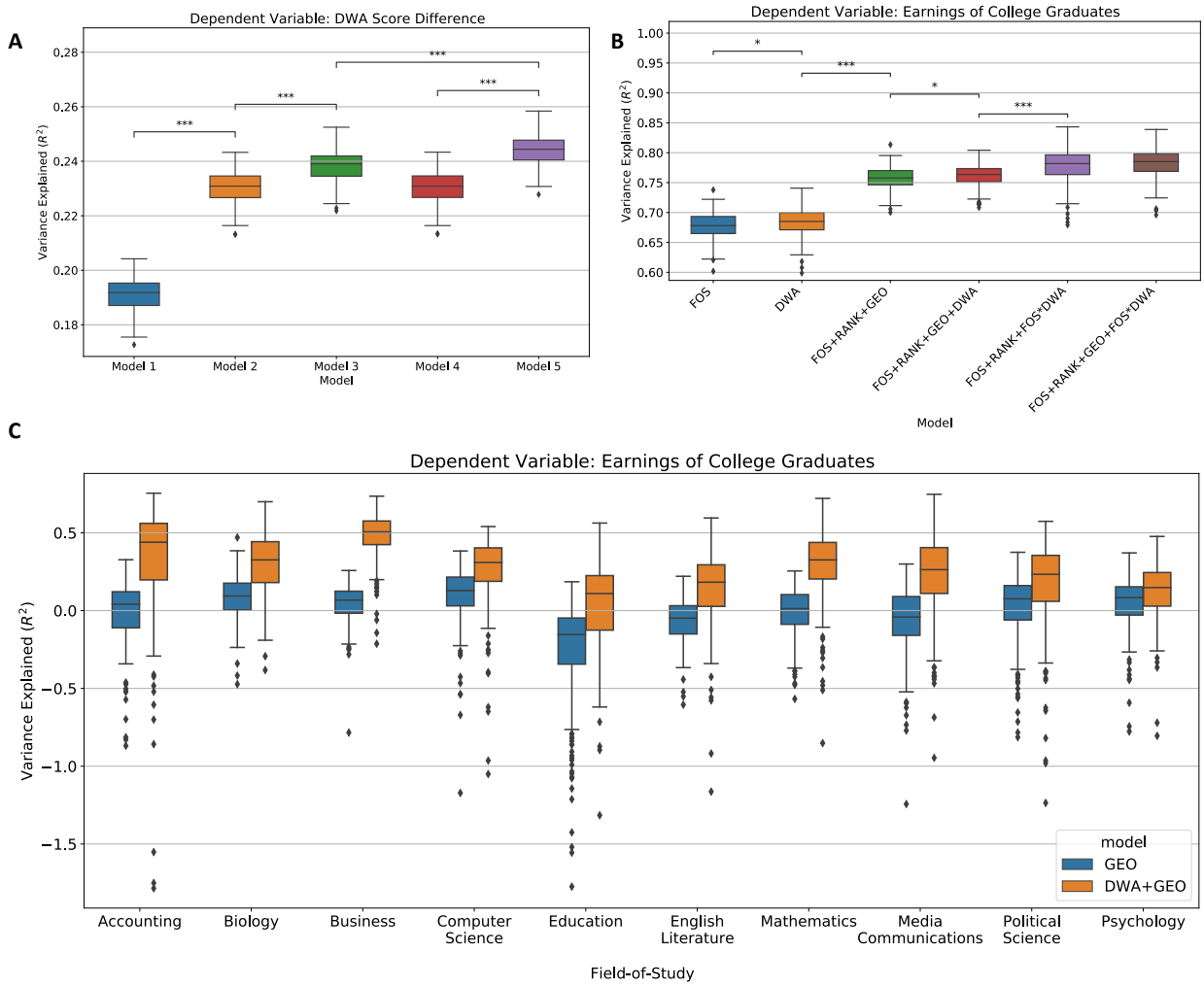


Figure 55: Kolmogorov-Smirnov (KS) statistical test for the subset of median earnings of graduates in *Business*. The subset distribution passes the test with the $p\text{-}value = 0.716$ (> 0.05). For visualization purposes, we use the natural logarithm of the number of the syllabi on the x-axis. The data points are on the red line and the right of the red line belongs to the selected subset used in our analysis.

Table 11: Numbers of earnings records of the top ten FOS that have passed the Kolmogorov–Smirnov test with the $p - values < 0.05$.

Field-of-Study	Number of records (after filtering)	Number of records (before filtering)	P-value (Kolmogorov–Smirnov test)
Business	246	683	0.716
Computer Science	198	640	0.06
Biology	181	563	0.89
Psychology	173	539	0.742
Mathematics	142	399	0.53
English Literature	135	526	0.81
Political Science	132	424	0.981
Education	122	424	0.419
Media / Communications	119	375	0.737
Accounting	109	224	0.542



$p <= 0.05$*, $p <= 0.01$**, $p <= 0.001$***

Figure 56: **Workplace activities detected from syllabi predicting teaching dynamics within a field of study and earnings of college graduates**. We perform 5-fold cross-validation and repeat 40 times for each model and measure the variance explained (i.e., R^2) by the resulting model applied to the test set. Asterisks denote significant differences between model performances with Bonferroni correction. (A) Predicting changing DWAs' importance over 10 years (2008 vs. 2017). Model 1 considers current DWA scores and FOS fixed effects as a baseline, while other models explore DWAs' relationships via Jaccard similarity. (B) Predicting median earnings of graduates across all FOS using FOS and RANK fixed effects as the baseline. (C) Predicting median earnings within FOS, with mean earnings as the baseline. DWA models outperform baseline models with $p - values < 0.001$ across all FOS. $R^2 < 0$ occurs in cross-validation when the model overfits or encounters outlier issues.

Table 12: DWAs that have significant coefficients in the OLS regression analysis of the Earnings of Recent College Graduates.

Field-of-Study	Detailed Work Activity	Coefficient	P-value
Business	advise others on career or personal development.	1.647	0.00957
	complete documentation required by programs or regulations.	1.805	0.00238
	conduct health or safety training programs.	-2.674	0.00005
	direct criminal investigations.	-2.175	0.00546
	evaluate program effectiveness.	2.392	0.00002
	explain project details to the general public.	-1.62	0.01865
	explain use of products or services.	-1.702	0.0421
	position construction forms or molds.	2.434	0.04461
	research methods to improve food products.	1.514	0.01467
	review laws or regulations to maintain professional knowledge.	-1.25	0.04926
Computer Science	estimate labor or resource requirements for forestry, fishing, or agricultural operations.	-1.509	0.0367
	explain technical medical information to patients.	-1.536	0.0097
Biology	coordinate personnel recruitment activities.	-2.101	0.00095
	direct technical activities or operations.	-1.714	0.0174
	plant greenery to improve landscape appearance.	-2.294	0
	prepare outgoing mail.	1.898	0.04589
	test characteristics of materials or structures.	0.906	0.03263
Psychology	diagnose neural or psychological disorders.	0.635	0.04753
	distribute instructional or library materials.	1.213	0.03618
	evaluate patient functioning, capabilities, or health.	1.931	0.00049
	plan menu options.	1.621	0.01161
	refer clients to community or social service programs.	-0.973	0.01262
	select resources needed to accomplish tasks.	-1.44	0.03091
Mathematics	conduct diagnostic tests to determine patient health.	-0.869	0.04136
	schedule activities or facility use.	-0.965	0.03244
	teach online courses.	-1.019	0.02433
English Literature	adjust routes or speeds as necessary.	-1.504	0.021
	design energy production or management equipment or systems.	1.875	0.01623
Political Science	–	–	–
Education	design integrated computer systems.	1.024	0.02689
	teach social science courses at the college level.	0.526	0.02646
Media/Communications	confer with managers to make operational decisions.	1.911	0.00424
	review art or design materials.	1.137	0.02602
	serve on institutional or departmental committees.	1.751	0.00426
Accounting	advise others on career or personal development.	3.863	0
	develop artistic or design concepts for decoration, exhibition, or commercial purposes.	-1.803	0.00111
	make decisions in legal cases.	1.35	0.02197
	participate in staffing decisions.	1.664	0.03567
	process animal carcasses.	-1.113	0.02748
	promote educational institutions or programs.	-1.738	0.01258
	promote environmental sustainability or conservation initiatives.	-2.167	0.03087

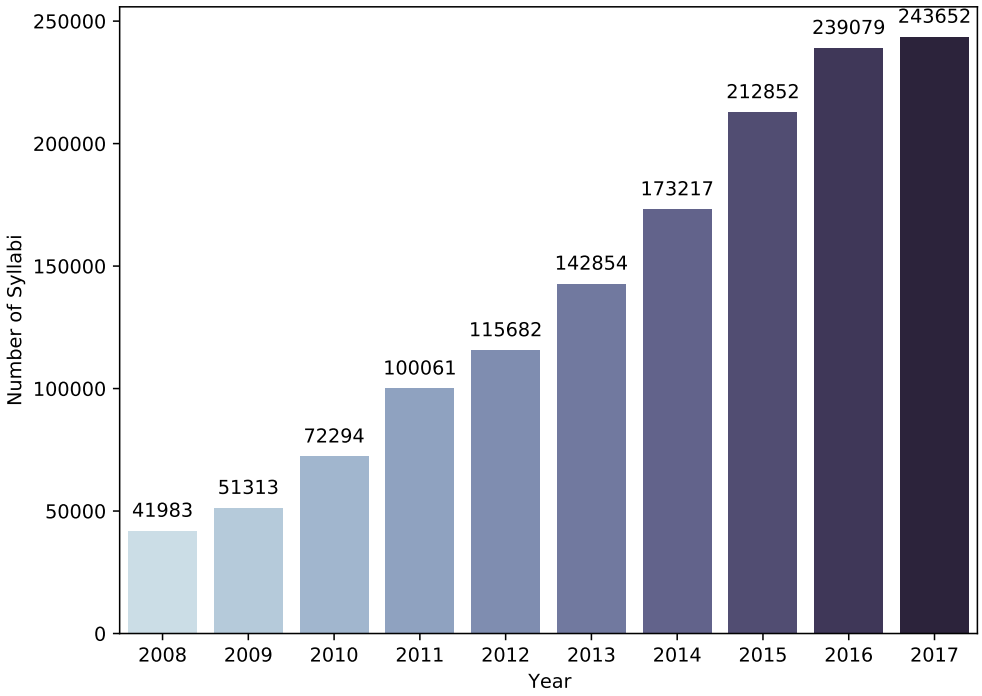


Figure 57: Course statistics per year in OSP data.

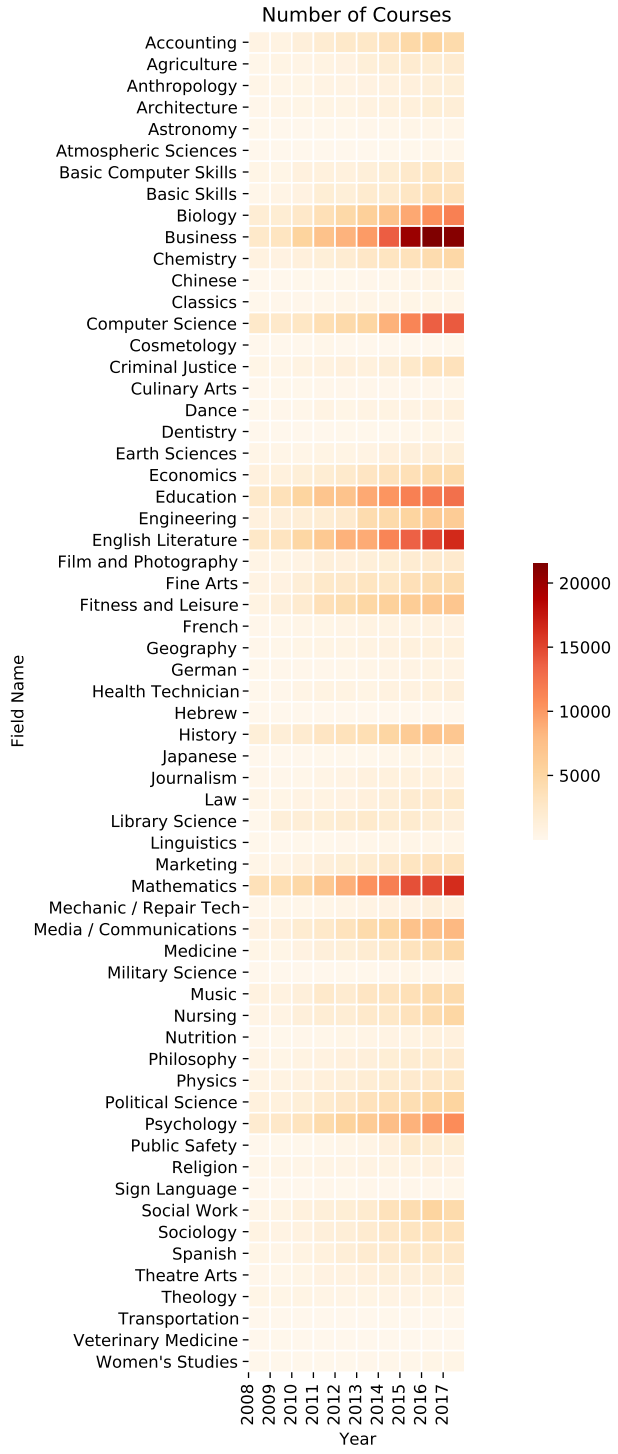


Figure 58: Course statistics per year and per FOS in OSP data.

Appendix B Course Recommendation

Table 13: Summary of Majors of Participants.

Majors	**Number of Subjects**
'Letters & Sci Undeclared'	15
'Molecular & Cell Biology'	3
'Mechanical Engineering'	3
'L&S Computer Science'	3
'Electrical Eng & Comp Sci'	2
'Chemistry'	2
'L&S Public Health'	2
'Bioengineering'	2
'Industrial Eng & Ops Rsch'	1
'Molecular Environ Biology'	1
'Media Studies'	1
'Mathematics'	1
'L&S Data Science'	1
'Integrative Biology'	1
'Info & Data Science-MIDS'	1
'Applied Mathematics', 'L&S Computer Science'	1
'Engineering Physics'	1
'Economics'	1
'Economics', 'L&S Data Science'	1
'Economics', 'L&S Data Science', 'L&S Ops Research & Mgmt Sci'	1
'Economics', 'French', 'History', 'Philosophy'	1
'Economics', 'Electrical Eng & Comp Sci', 'L&S Data Science'	1
'EECS/MSE Joint Major'	1
'Cognitive Science'	1
'Cognitive Science', 'L&S Computer Science'	1
'Civil Engineering'	1
'Business Administration'	1
'Business Administration', 'Electrical Eng & Comp Sci'	1
'Statistics'	1

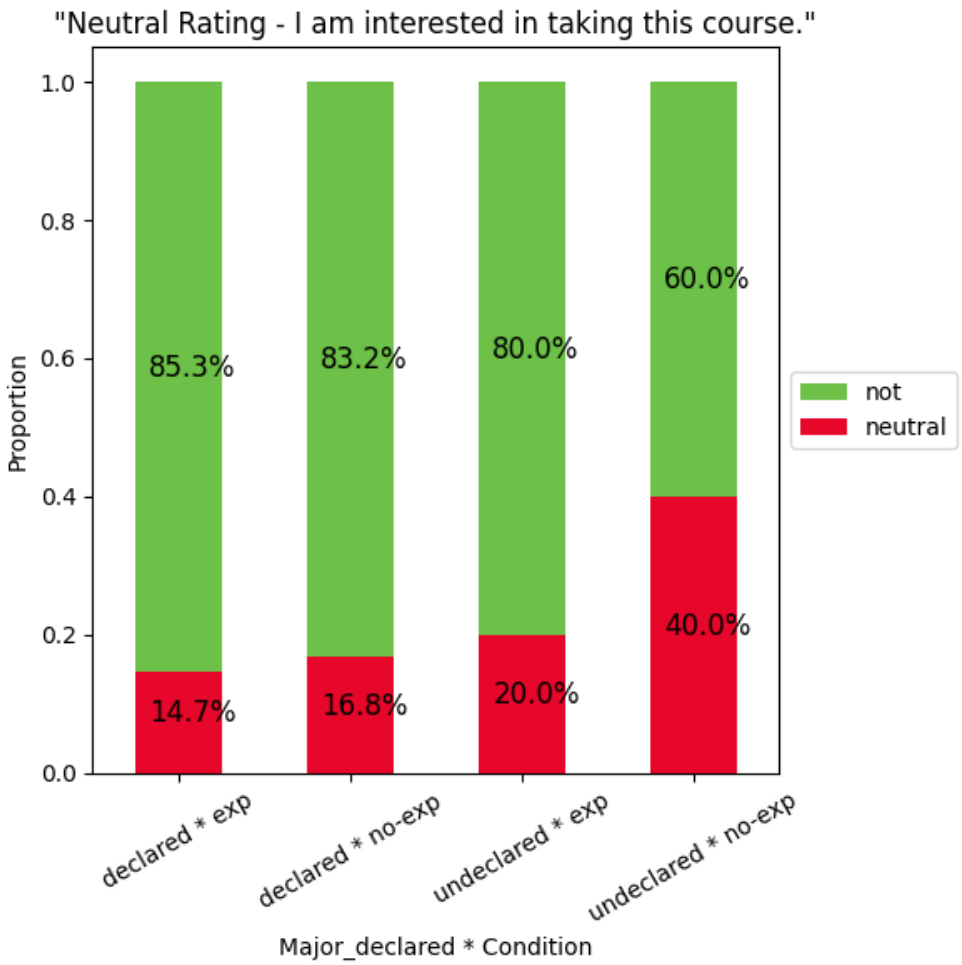


Figure 59: Distribution of 'Neutral' ratings for the statement 'I am interested in taking this course.' among four groups based on the interactions between major (declared vs. undeclared) and the presence of an explanation (vs. no explanation): declared * exp (N=95), declared * no-exp (N=95), undeclared * exp (N=45), undeclared * no-exp (N=30). The 'Neutral' ratings are aggregated from the responses to the three primary research questions: Q1, Q2, and Q3. The percentage of 'Neutral' ratings is 19.24% (51 'Neutral' ratings of 265).

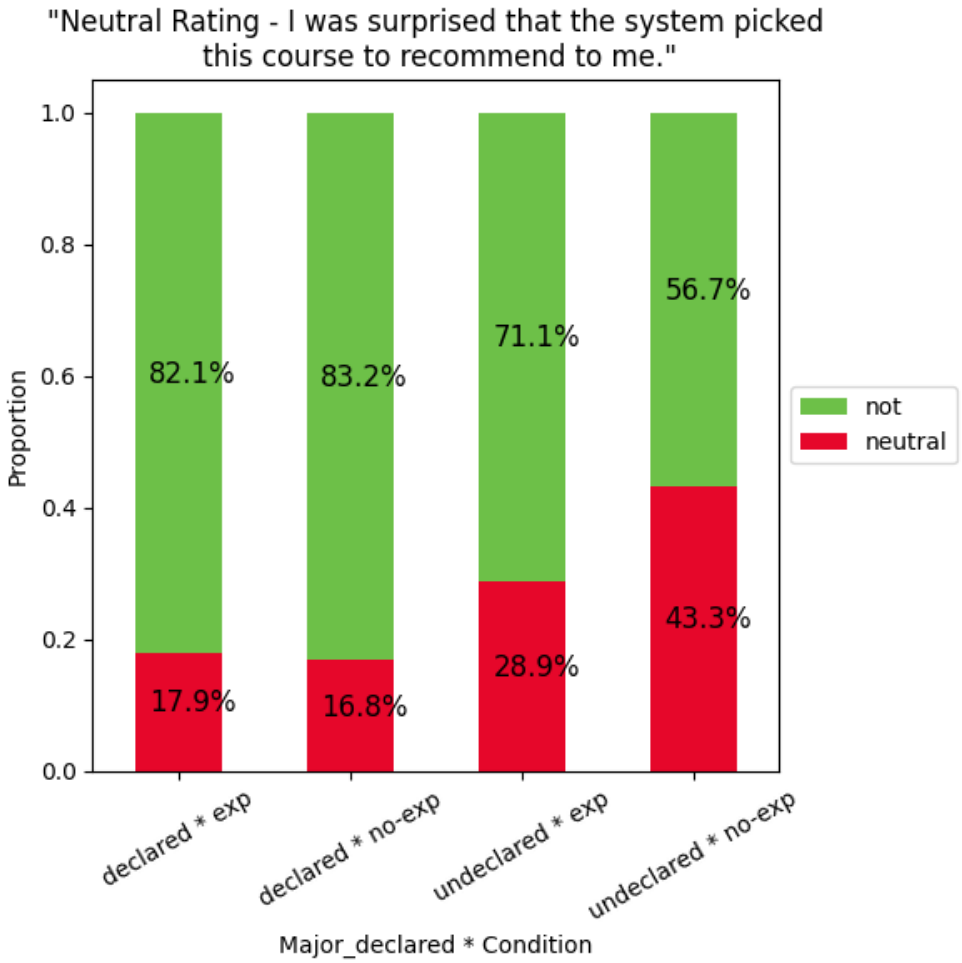


Figure 60: Distribution of 'Neutral' ratings for the statement 'I was surprised that the system picked this course to recommend to me.' among four groups based on the interactions between major (declared vs. undeclared) and the presence of an explanation (vs. no explanation): declared * exp (N=95), declared * no-exp (N=95), undeclared * exp (N=45), undeclared * no-exp (N=30). The percentage of 'Neutral' ratings is 22.26% (59 'Neutral' ratings of 265).

Bibliography

[1] Raj Chetty, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. Mobility report cards: The role of colleges in intergenerational mobility, 2017.

[2] Bryce Loo, N. Luo, and Ziyi Ye. Career prospects and outcomes of u.s.-educated international students: Improving services, bolstering success. New York: World Education Services. Retrieved from wes.org/partners/research/, 2017.

[3] Dirk Witteveen and Paul Attewell. The earnings payoff from attending a selective college. *Social Science Research*, 66:154–169, 2017. ISSN 0049-089X.

[4] Judith Scott-Clayton. The shapeless river: Does a lack of struc- ture inhibit students' progress at community colleges? In B. L. Castleman, S. Schwartz, and S. Baum, editors, *Decision-making for student success: Behavioral insights to improve college access and persistence*, pages 102–123. Routledge, 2015.

[5] Doug Shapiro, Afet Dundar, Faye Huie, Phoebe Khasiala Wakhungu, Xin Yuan, Angel Nathan, and Youngsik Hwang. Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees. (signature report no. 13). national student clearinghouse (2017), 2017.

[6] D. Shapiro and A. Dundar. Completing college: A national view of student attainment rates. url https:// nscresearchcenter.org/signaturereport12/, 2016.

[7] David Wilezol and William Bennett. *Is College Worth It?* Thomas Nelson, Nashville, TN, 2013.

[8] William Bonvillian and Sanjay Sarma. *Workforce Education: A New Roadmap*. MIT Press, 2021.

[9] Sara Goldrick-Rab. Following their every move: An investigation of social-class differences in college pathways. *Sociology of Education*, 79(1):67–79, 2006.

[10] Four-year myth: Make college more affordable. restore the promise of graduating on time. Complete College America. Indianapolis, 2014. Retrieved from `https://files.eric.ed.gov/fulltext/ED558792.pdf`.

[11] Thomas Bailey, Shanna Smith Jaggars, and Davis Jenkins. Redesigning america's community colleges. *Harvard University Press.*, 2015.

[12] Sorathan Chaturapruek, Thomas S. Dee, Ramesh Johari, René F. Kizilcec, and Mitchell L. Stevens. How a data-driven course planning tool affects college students' gpa: Evidence from two field experiments. In *Proceedings of the Fifth Annual ACM*

Conference on Learning at Scale, L@S '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358866.

[13] Sorathan Chaturapruek, Tobias Dalberg, Marissa E. Thompson, Sonia Giebel, Monique H. Harrison, Ramesh Johari, Mitchell L. Stevens, and Rene F. Kizilcec. Studying undergraduate course consideration at scale. *AERA Open*, 7: 2332858421991148, 2021.

[14] Ivana Ognjanovic, Dragan Gaević, and Shane Dawson. Using institutional data to predict student course selections in higher education. *Internet High. Educ.*, 29:49–62, 2016.

[15] Weijie Jiang, Zachary A. Pardos, and Qiang Wei. Goal-based course recommendation. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge*, LAK19, page 36–45, New York, NY, USA, 2019. Association for Computing Machinery.

[16] Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160, 2020.

[17] Rosta Farzan and Peter Brusilovsky. Encouraging user participation in a course recommender system: An impact on user behavior. *Computers in Human Behavior*, 27 (1):276–284, 2011. Current Research Topics in Cognitive Load Theory.

[18] Aditya Parameswaran, Petros Venetis, and Hector Garcia-Molina. Recommendation systems with complex constraints: A course recommendation perspective. *ACM Trans. Inf. Syst.*, 29(4), December 2011. ISSN 1046-8188.

[19] Zachary A Pardos, Zihao Fan, and Weijie Jiang. Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, pages 1–39, 2019.

[20] Daniel F. Chambliss and Christopher G. Takacs. How college works. *Harvard University Press.*, 2014.

[21] Zachary A Pardos and Weijie Jiang. Designing for serendipity in a university course recommendation system. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 350–359, 2020.

[22] Mustafa Bilgic and Raymod Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization 2005, Workshop on the Next Stage of Recommender Systems Research at IUI'05*, 2005.

[23] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. Inspectability and control in social recommenders. In *6th ACM Conference on Recommender System*, pages 43–50, 2012.

[24] Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, 115(50):12630–12637, 2018. ISSN 0027-8424.

[25] Sarah H. Bana. work2vec: Using language models to understand wage premia. Working paper, Stanford Digital Economy Lab., March 2022.

[26] Ahmad Alabdulkareem, Morgan R. Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. Unpacking the polarization of workplace skills. *Science Advances*, 4(7):eaao6030, 2018. doi: 10.1126/sciadv.aao6030.

[27] David J. Deming. The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics*, 132(4):1593–1640, 06 2017.

[28] Morgan R. Frank, David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. 116(14):6531–6539, 2019. ISSN 0027-8424.

[29] Aurora Esteban, Amelia Zafra Gómez, and Cristobal Romero. A Hybrid Multi-Criteria approach using a Genetic Algorithm for Recommending Courses to University Students. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, 2018.

[30] Weijie Jiang and Zachary Pardos. Evaluating sources of course information and models of representation on a variety of institutional prediction tasks. In *Proceedings of the 13th International Conference on Educational Data Mining*, pages 115–125. International Educational Data Mining Society, 2020.

[31] Jialu Liu, Jingbo Shang, and Jiawei Han. *Phrase Mining from Massive Text and Its Applications*. Morgan & Claypool, 2017. doi: 10.2200/S00759ED1V01Y201702DMK013.

[32] Jesse Vig, Shilad Sen, and John Riedl. Tagsplanations: Explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, page 47–56, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605581682.

[33] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '14, page 83–92, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450322577.

[34] Annika Waern. User involvement in automatic filtering - an experimental study. *User Modeling and User Adapted Interaction*, 14(201-237), 2004.

[35] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. Open user profiles for adaptive news systems: help or harm? In *the 16th international conference on World Wide Web, WWW '07*, pages 11–20. ACM, 2007.

[36] Zachary A. Pardos, Hung Chau, and Haocheng Zhao. Data-assistive course-to-course articulation using machine translation. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, L@S '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368049. Article No.: 22.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

[38] Run Yu, Zachary A. Pardos, Hung Chau, and Peter Brusilovsky. Orienting Students to Course Recommendations Using Three Types of Explanation. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct)*, page 238–245, 2021.

[39] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1441–1450, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763.

[40] Rada. Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, July 2004.

[41] Adrien Bougouin, Florian Boudin, and Béatrice Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan, 2013. Asian Federation of Natural Language Processing.

[42] Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 620–628, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1.

[43] Xiaojun Wan and Jianguo Xiao. Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976. Coling 2008 Organizing Committee, 2008.

[44] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, pages 254–255, New York, NY, USA, 1999. ACM. ISBN 1-58113-145-3.

[45] Olena Medelyan, Eibe Frank, and Ian H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1318–1327, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3.

[46] Shuting Wang, Chen Liang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C. Lee Giles. Concept hierarchy extraction from textbooks. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 147–156, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3307-8.

[47] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 661–670, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4.

[48] Corina Florescu and Cornelia Caragea. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115. Association for Computational Linguistics, 2017.

[49] Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu. Unsupervised keyphrase extraction: Introducing new kinds of words to keyphrases. In Byeong Ho Kang and Quan Bai, editors, *AI 2016: Advances in Artificial Intelligence*, pages 665–671, Cham, 2016. Springer International Publishing. ISBN 978-3-319-50127-7.

[50] Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel. Corephrase: Keyphrase extraction for document clustering. In Petra Perner and Atsushi Imiya, editors, *Machine Learning and Data Mining in Pattern Recognition*, pages 265–274, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31891-0.

[51] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. In Michael W. Berry and Jacob Kogan, editors, *Text Mining: Applications and Theory*, pages 1–20. John Wiley and Sons, Ltd, 2010.

[52] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[53] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 213–222, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9.

[54] Thuy Dung Nguyen and Min-Yen Kan. Keyphrase extraction in scientific publications. In Dion Hoe-Lian Goh, Tru Hoang Cao, Ingeborg Torvik Sølvberg, and Edie Rasmussen, editors, *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[55] Pinaki Bhaskar, Kishorjit Nongmeikapam, and Sivaji Bandyopadhyay. Keyphrase extraction in scientific articles: A supervised approach. In *Proceedings of COLING 2012*, 2012.

[56] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991, 2015. URL `https://api.semanticscholar.org/CorpusID:12740621`.

[57] Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In *The World Wide Web Conference*, WWW '19, page 2551–2557, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748.

[58] Dhruva Sahrawat, Debanjan Mahata, Haimin Zhang, Mayank Kulkarni, Agniv Sharma, Rakesh Gosangi, Amanda Stent, Yaman Kumar, Rajiv Ratn Shah, and Roger Zimmermann. Keyphrase extraction as sequence labeling using contextualized embeddings. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 328–335, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5.

[59] Seoyeon Park and Cornelia Caragea. Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5409–5419, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[60] Yansen Wang, Zhen Fan, and Carolyn Rose. Incorporating multimodal information in open-domain web keyphrase extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1790–1800, Online, November 2020. Association for Computational Linguistics.

[61] Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. Open domain web keyphrase extraction beyond language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5175–5184, Hong Kong, China, November 2019. Association for Computational Linguistics.

[62] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[63] Igor Labutov, Yun Huang, Peter Brusilovsky, and Daqing He. Semi-supervised techniques for mining learning outcomes and prerequisites. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 907–915, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4887-4.

[64] Isaac Alpizar-Chacon and Sergey Sosnovsky. Order out of chaos: Construction of knowledge models from pdf textbooks. In *Proceedings of the ACM Symposium on Document Engineering 2020*, DocEng '20, pages 1–10, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380003. Article No.:8.

[65] Khushboo Thaker, Peter Brusilovsky, and Daqing He. Student modeling with automatic knowledge component extraction for adaptive textbooks. In *Proceedings of First Workshop on Intelligent Textbooks at 20th International Conference on Artificial Intelligence in Education (AIED 2019)*, 2019.

[66] Mehrnoush Shamsfard and Ahmad Abdollahzadeh Barforoush. Learning ontologies from natural language texts. *Int. J. Human-Computer Studies*, 60:17–63, 2004.

[67] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Ontology learning from text: A look back and into the future. *ACM Comput. Surv.*, 44(4), September 2012. ISSN 0360-0300.

[68] Amal Zouaq, Roger Nkambou, and Claude Frasson. Building domain ontologies from text for educational purposes. In Erik Duval, Ralf Klamma, and Martin Wolpers, editors, *Creating New Learning Experiences on a Global Scale*, pages 393–407, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[69] Angel Conde, Mikel Larrañaga, Ana Arruarte, and Jon A. Elorriaga. Testing language independence in the semiautomatic construction of educational ontologies. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, pages 545–550, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07221-0.

[70] Angel Conde, Mikel Larrañaga, Ana Arruarte, Jon A. Elorriaga, and Dan Roth. Litewi: A combined term extraction and entity linking method for eliciting educational

ontologies from textbooks. *J. Assoc. Inf. Sci. Technol.*, 67(2):380–399, February 2016. ISSN 2330-1635.

[71] Hung Chau, Igor Labutov, Khushboo Thaker, Daqing He, and Peter Brusilovsky. Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*, 31:820–846, 2021.

[72] Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain, December 2020. Association for Computational Linguistics.

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186. Association for Computational Linguistics, June 2019.

[74] Boxuan Ma, Min Lu, Yuta Taniguchi, and Shin'ichi Konomi. Exploring the design space for explainable course recommendation systems in university environments. In *Companion Proceedings 10th International Conference on Learning Analytics Knowledge*, 03 2020.

[75] Deepani B. Guruge, Rajan Kadel, and Sharly J. Halder. The state of the art in methodologies of course recommender systems—a review of recent research. *Data*, 6 (2), 2021.

[76] Sanjog Ray and Anuj Sharma. A collaborative filtering based approach for recommending elective courses. In Sumeet Dua, Sartaj Sahni, and D. P. Goyal, editors, *Information Intelligence, Systems, Technology and Management*, pages 330–339, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[77] Hana Bydžovská. Course enrollment recommender system. In Mingyu Feng Tiffany Barnes, Min Chi, editor, *Proceedings of the 9th International Conference on Educational Data Mining*, pages 312–317, Raleigh, NC, USA, 2016. International Educational Data Mining Society.

[78] Boxuan Ma, Yuta Taniguchi, and Shin'ichi Konomi. Course recommendation for university environment. In *Educational Data Mining*, pages 460–466, 2020.

[79] Narimel Bendakir and Esma Aimeur. Using association rules for course recommendation. AAAI Workshop Technical Report, 2006.

[80] Behdad Bakhshinategh, Gerasimos Spanakis, Osmar R Zaiane, and Samira ElAtia. A course recommender system based on graduating attributes. In *International Conference on Computer Supported Education*, pages 347–354, 2017.

[81] Zameer Gulzar, A. Anny Leema, and Gerard Deepak. Pcrs: Personalized course recommender system based on hybrid approach. *Procedia Computer Science*, 125:518–524, 2018. ISSN 1877-0509. The 6th International Conference on Smart Computing and Communications.

[82] Raphaël Morsomme and Sofia Vazquez Alferez. Content-based course recommender system for liberal arts education. In *Educational Data Mining*, 2019.

[83] Elham S. Khorasani, Zhao Zhenge, and John Champaign. A markov chain collaborative filtering model for course enrollment recommendations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3484–3490, 2016. doi: 10.1109/BigData.2016.7841011.

[84] Agoritsa Polyzou, Athanasios N. Nikolakopoulos, and George Karypis. Scholars walk: A markov chain framework for course recommendation. In *Educational Data Mining*, pages 396–401, 2019.

[85] A. Esteban, A. Zafra, and C. Romero. Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. *Knowledge-Based Systems*, 194:105385, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.105385.

[86] Ammar A. Neamah and Amer S. El-Ameer. Design and evaluation of a course recommender system using content-based approach. In *2018 International Conference on Advanced Science and Engineering (ICOASE)*, pages 1–6, 2018. doi: 10.1109/ICOASE.2018.8548789.

[87] Mara Houbraken, Chang Sun, Evgueni Smirnov, and Kurt Driessens. Discovering hidden course requirements and student competencies from grade data. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, page 147–152, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350679.

[88] Michael Backenköhler, Felix Scherzinger, Adish Kumar Singla, and Verena Wolf. Data-driven approach towards a personalized curriculum. *EasyChair Preprints*, 2018.

[89] Asmaa Elbadrawy and George Karypis. Domain-aware grade prediction and top-n course recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 183–190, 2016.

[90] Mohammed E. Ibrahim, Yanyan Yang, David L. Ndzi, Guangguang Yang, and Murtadha Al-Maliki. Ontology-based personalized course recommendation framework. *IEEE Access*, 7:5180–5199, 2018. doi: 10.1109/ACCESS.2018.2889635.

[91] Sara Morsy and George Karypis. Will this course increase or decrease your gpa? towards grade-aware course recommendation. *ArXiv*, abs/1904.11798, 2019.

[92] Jie Xu, Tianwei Xing, and Mihaela van der Schaar. Personalized course sequence recommendations. *IEEE Transactions on Signal Processing*, 64(20):5340–5352, 2016. doi: 10.1109/TSP.2016.2595495.

[93] Zhen Li, David Tinapple, and Hari Sundaram. Visual planner: Beyond prerequisites, designing an interactive course planner for a 21st century flexible curriculum. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 1613–1618, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310161.

[94] Chris Wong. Sequence based course recommender for personalized curriculum planning. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 531–534, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93846-2.

[95] Erzhuo Shao, Shiyuan Guo, and Zachary A. Pardos. Degree planning with plan-bert: Multi-semester recommendation using future courses of interest. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14920–14929, May 2021.

[96] Jason L. Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. Dynamics of human trust in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, page 305–308, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681.

[97] Denis Parra and Peter Brusilovsky. User-controllable personalization: A case study with setfusion. *International Journal of Human-Computer Studies*, 78:43–67, 2015. ISSN 1071-5819.

[98] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, page 158–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131763.

[99] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, page 241–250, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132220.

[100] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, page 830–831, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134541.

[101] Weiquan Wang and Izak Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4):217–246, 2007.

[102] Nava Tintarev and Judith Masthoff. The effectiveness of personalized movie explanations: an experiment using commercial meta-data. In Wolfgang Nejdl, Judy Kay, Pearl Pu, and Eelco Herder, editors, *5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2008)*, volume 5149 of *Lecture Notes in Computer Science*, pages 204–213. Springer Verlag, 2008.

[103] Nava Tintarev and Judith Masthoff. Evaluating recommender explanations: Problems experienced and lessons learned for the evaluation of adaptive systems. In Stephan Weibelzahl, Judith Masthoff, Alexandros Paramythis, and Lex van Velsen, editors, *the Sixth Workshop on User-Centred Design and Evaluation of Adaptive Systems, held in conjunction with the International Conference on User Modeling, Adaptation, and Personalization (UMAP2009), Trento, Italy*, CEUR Workshop Proceedings, ISSN 1613-0073, pages 54–63, 2009.

[104] Cong Yu, Laks VS Lakshmanan, and Sihem Amer-Yahia. Recommendation diversification using explanations. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1299–1302. IEEE, 2009.

[105] Jesse Vig, Shilad Sen, and John Riedl. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56. ACM, 2009.

[106] Pearl Pu and Li Chen. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556, 2007.

[107] Nava Tintarev and Judith Masthoff. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*, pages 353–382. Springer, 2015.

[108] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5):393–444, 2017.

[109] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020. ISSN 1554-0669.

[110] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *3rd International Workshop on Web Personalisation, Recommender Systems and Intelligent User Interfaces at IEEE 23rd International Conference on Data Engineering (ICDE 2007)*, pages 801–810, 2007.

[111] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[112] Masahiro Sato, Budrul Ahsan, Koki Nagatani, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. Explaining recommendations using contexts. In *23rd International Conference on Intelligent User Interfaces*, pages 659–664. ACM, 2018.

[113] Sidra Naveed, Benedikt Loepp, and Jürgen Ziegler. On the use of feature-based collaborative explanations: An empirical comparison of explanation styles. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, page 226–232, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379502.

[114] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 297–305, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346528.

[115] Darius Afchar and Romain Hennequin. Making neural networks interpretable with attribution: Application to implicit signals prediction. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, page 220–229, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832.

[116] Nava Tintarev and Judith Masthoff. Beyond explaining single item recommendations. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 711–756, New York, NY, 2022. Springer US.

[117] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendatkaty2018ions using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.

[118] Georgina Peake and Jun Wang. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2060–2069. ACM, 2018.

[119] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie. A reinforcement learning framework for explainable recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 587–596, 2018. doi: 10.1109/ICDM.2018.00074.

[120] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. Post-hoc explanations for complex model recommendations using simple methods. In *IntRS '20 - Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, 2020.

[121] Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction*, 31:629–673, 2020.

[122] Noemi Mauro, Zhongli Filippo Hu, and Liliana Ardissono. Justification of recommender systems results: a service-based approach. *User modeling and user-adapted interaction*, page 1—43, October 2022. ISSN 0924-1868. doi: 10.1007/s11257-022-09345-8.

[123] Guojing Zhou, Xi Yang, Hamoon Azizsoltani, Tiffany Barnes, and Min Chi. Improving student-system interaction through data-driven explanations of hierarchical reinforcement learning induced pedagogical policies. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 284–292, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612.

[124] Jordan Barria-Pineda, Kamil Akhuseyinoglu, Stefan Želem-Ćelap, Peter Brusilovsky, Aleksandra Klasnja Milicevic, and Mirjana Ivanovic. Explainable recommendations in a personalized programming practice system. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 64–76, Cham, 2021. Springer International Publishing.

[125] Kyosuke Takami, Yiling Dai, Brendan Flanagan, and Hiroaki Ogata. Educational explainable recommender usage and its effectiveness in high school summer vacation assignment. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 458–464, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450395731.

[126] Don Hossler, Doug Shapiro, Afet Dundar, Mary Ziskin, Jin Chen, Desiree Zerquera, and Vasti Torres. Transfer and mobility: A national view of pre-degree student movement in postsecondary institutions. signature report 2. *National Student Clearinghouse*, 2012.

[127] Alexandria Walton Radford, Lutz Berkner, Sara C Wheeless, and Bryan Shepherd. Persistence and attainment of 2003-04 beginning postsecondary students: After 6 years. first look. nces 2011-151. *National Center for Education Statistics*, 2010.

[128] Doug Shapiro, Afet Dundar, Faye Huie, Phoebe Khasiala Wakhungu, Xin Yuan, Angel Nathan, and Youngsik Hwang. Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees.(signature report no. 13). *National Student Clearinghouse*, 2017.

[129] United States. Government Accountability Office (GAO). Higher education: students need more information to help reduce challenges in transferring college credits. 2017. URL `https://www.gao.gov/assets/690/686530.pdf`.

[130] David B Monaghan and Paul Attewell. The community college route to the bachelor's degree. *Educational Evaluation and Policy Analysis*, 37(1):70–91, 2015.

[131] Thomas Bailey, Shanna S. Jaggars, and Davis Jenkins. Redesigning america's community colleges: A clearer path to student success. *Cambridge, MA: Harvard University Press.*, 2015.

[132] Davis Jenkins, Amy E. Brown, John Fink, Hana Lahr, and Takeshi Yanagiura. Building guided pathways to community college student success: Promising practices and early evidence from tennessee. *Community College Research Center (CCRC)*, 2018.

[133] Jennifer B. Schanker and Erica L. Orians. Guided pathways: the scale of adoption in michigan. *Michigan Community College Association (MCCA)*, 2018. Retrieved from `http://www.mcca.org/uploads/ckeditor/files/SOA%20Publication%20Final(1).pdf`.

[134] Jack E Smith. Articulation and the chief instructional ocfficer. *New Directions for Community Colleges*, 1982(39):41–49, 1982.

[135] California Intersegmental Articulation Council. Handbook of california articulation policies and procedures. 2013. URL `https://www.csusb.edu/sites/csusb/files/CIAC_Handbook_Spring_2013.pdf`.

[136] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013. URL `http://arxiv.org/abs/1309.4168`.

[137] Zachary A Pardos and Andrew Joo Hun Nam. A university map of course knowledge. *PloS one*, 15(9):e0233207, 2020.

[138] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[139] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[140] Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474. Association for Computational Linguistics, 2018.

[141] Matthew Dong, Run Yu, and Zachary A Pardos. Design and deployment of a better course search tool: Inferring latent keywords from enrollment networks. In *European Conference on Technology Enhanced Learning*, pages 480–494. Springer, 2019.

[142] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[143] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, Rene Witte, Greg Butler, and Adrian Tsang. An approach to controlling user models and personalization effects in recommender systems. In *international conference on Intelligent user interfaces, IUI '2013*, pages 49–56. ACM Press, 2013.

[144] Dorota Glowacka, Tuukka Ruotsalo, Ksenia Konuyshkova, Kumaripaba Athukorala, Samuel Kaski, and Giulio Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *international conference on Intelligent user interfaces, IUI '2013*, pages 117–127. ACM Press, 2013.

[145] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[146] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6297–6308, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[147] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

[148] Xianjie Shen, Yinghan Wang, Rui Meng, and Jingbo Shang. Unsupervised deep keyphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11303–11311, Jun. 2022.

[149] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y.

[150] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. ISSN 0925-2312.

[151] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), jul 2020. ISSN 2157-6904.

[152] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-1030.

[153] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[154] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[155] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank F. Xu, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5253–5260. AAAI Press, 2018.

[156] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[157] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks, 2013.

[158] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, 1996.

[159] Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 591–598, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[160] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

[161] Yansen Wang, Zhen Fan, and Carolyn Rose. Incorporating multimodal information in open-domain web keyphrase extraction. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 1790–1800, Online, November 2020. Association for Computational Linguistics.

[162] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[163] Mengdi Wang, Hung Chau, Khushboo Thaker, Peter Brusilovsky, and Daqing He. Knowledge annotation for intelligent textbooks. *Technology knowledge and learning*, 28:1–22, 2023.

[164] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837, 2018. doi: 10.1109/TKDE.2018.2812203.

[165] Behnam Rahdari, Peter Brusilovsky, Khushboo Thaker, and Jordan Barria-Pineda. Knowledge-driven wikipedia article recommendation for electronic textbooks. In Carlos Alario-Hoyos, María Jesús Rodríguez-Triana, Maren Scheffel, Inmaculada Arnedillo-Sánchez, and Sebastian Maximilian Dennerlein, editors, *Addressing Global Challenges and Quality Education*, pages 363–368, Cham, 2020. Springer International Publishing. ISBN 978-3-030-57717-9.

[166] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[167] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 157–164, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306836.

[168] Denis Kotkov, Joseph A. Konstan, Qian Zhao, and Jari Veijalainen. Investigating serendipity in recommender systems based on real user feedback. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, page 1341–1350, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351911.

[169] Bart P. Knijnenburg and Alfred Kobsa. Helping users with information disclosure decisions: Potential for adaptation. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, page 407–416, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319652.

[170] Bart P. Knijnenburg and Martijn C. Willemsen. *Evaluating Recommender Systems with User Experiments*, pages 309–352. Springer US, Boston, MA, 2015.

[171] Krysta M. Svore, Maksims N. Volkovs, and Christopher J.C. Burges. Learning to rank with multiple objective functions. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 367–376, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324.

[172] Debabrata Mahapatra, Chaosheng Dong, Yetian Chen, and Michinari Momma. Multi-label learning to rank through multi-objective optimization. KDD '23, page 4605–4616, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030.

[173] Esteban Moro, Morgan R Frank, Alex Pentland, Alex Rutherford, Manuel Cebrian, and Iyad Rahwan. Universal resilience patterns in labor markets. *Nature communications*, 12(1):1–8, 2021.

[174] David Autor. Work of the Past, Work of the Future. Technical report, National Bureau of Economic Research, 2019.

[175] Peter Arcidiacono. Affirmative action in higher education: How do admission and financial aid rules affect future earnings? *Econometrica*, 73(5):1477–1524, 2005. ISSN 00129682, 14680262.

[176] Stephanie Riegg Cellini and Nicholas Turner. Gainfully employed? assessing the employment and earnings of for-profit college students using administrative data. *Journal of Human Resources*, 54:342—-370, 2019.

[177] Raj Chetty, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. Income Segregation and Intergenerational Mobility Across Colleges in the United States. *The Quarterly Journal of Economics*, 135(3):1567–1633, 02 2020. ISSN 0033-5533.

[178] Zachary Bleemer and Aashish Mehta. Will studying economics make you rich? a regression discontinuity analysis of the returns to college major. *American Economic Journal: Applied Economics*, 14(2):1–22, April 2022.

[179] Xiaoxiao Li, Sebastian Linde, and Hajime Shimao. Major Complexity Index and College Skill Production. Technical report, 2021. Available at SSRN 3791651.

[180] Barbara Biasi and Song Ma. The education-innovation gap. Working Paper 29853, National Bureau of Economic Research, March 2022.

[181] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. Skills2job: A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing*, 101:107049, 2021. ISSN 1568-4946.

[182] S C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[183] César A. Hidalgo, Pierre-Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, Canfei He, Dieter F. Kogler,

Andrea Morrison, Frank Neffke, David Rigby, Scott Stern, Siqi Zheng, and Shengjun Zhu. The principle of relatedness. In Alfredo J. Morales, Carlos Gershenson, Dan Braha, Ali A. Minai, and Yaneer Bar-Yam, editors, *Unifying Themes in Complex Systems IX*, pages 451–457, Cham, 2018. Springer International Publishing. ISBN 978-3-319-96661-8.

[184] Daron Acemoglu and David Autor. Chapter 12 - skills, tasks and technologies: Implications for employment and earnings. In David Card and Orley Ashenfelter, editors, *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier, 2011.

[185] Morgan R Frank, Lijun Sun, Manuel Cebrian, Hyejin Youn, and Iyad Rahwan. Small cities face greater impact from automation. *Journal of the Royal Society Interface*, 15 (139):20170946, 2018.

[186] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1–4, 2018.

[187] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, pages 1–13, 2019.

[188] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2020.

[189] Zenun Kastrati, Ali Shariq Imran, and Arianit Kurti. Integrating word embeddings and document topics with deep learning in a video classification framework. *Pattern Recognition Letters*, 128:85–92, 2019. ISSN 0167-8655.

[190] Helena Gomez-Adorno Grigori Sidorov, Alexander Gelbukh and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computacion y Sistemas*, 18(3):491–504, 2014.

[191] Emile Cammeraat and Mariagrazia Squicciarini. Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment. *OECD*, 2021.

[192] Eric R. Eide, Michael J. Hilmer, and Mark H. Showalter. Is it where you go or what you study? the relative influence of college selectivity and college major on earnings. *Contemporary Economic Policy*, 34(1):37–46, 2016.

[193] ChangHwan Kim, Christopher R. Tamborini, and Arthur Sakamoto. Field of study in college and lifetime earnings in the united states. *Sociology of Education*, 88(4): 320–339, 2015.

[194] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[195] Weijie Jiang, Zachary A. Pardos, and Qiang Wei. Goal-based course recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, page 36–45, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362566.

[196] David J Deming and Kadeem Noray. Earnings Dynamics, Changing Job Skills, and STEM Careers. *The Quarterly Journal of Economics*, 135(4):1965–2005, 06 2020. ISSN 0033-5533.

[197] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[198] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[199] Shade T Shutters, Rachata Muneepeerakul, and José Lobo. Constrained pathways to a creative urban economy. *Urban Studies*, 53(16):3439–3454, 2016.

[200] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[201] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[202] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. Universalner: Targeted distillation from large language models for open named entity recognition. *ArXiv*, 2023.

[203] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. Instructuie: Multi-task instruction tuning for unified information extraction. 2023.

[204] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130.

[205] Jimmy Lin, Ronak Pradeep, Tommaso Teofili, and Jasper Xian. Vector search with openai embeddings: Lucene is all you need. *ArXiv*, 2023.

[206] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. *ArXiv*, 2023.