Matthew Dong
mdong@berkeley.edu

Machine Learning Meets Education:
SURF 2018 Research Proposal

1 Research Statement

The education system has been left behind in our modern society, as teaching methodologies and web infrastructure in most schools have remained static in a rapidly evolving world. The Computational Approaches to Human Learning (CAHL) Lab works to bring the vanguard of machine learning to education in order to study and individualize the learning experience through a variety of contexts. Currently the lab is working closely with the University of California, Berkeley as a testing ground for a state of the art course recommendation system, available as the website AskOski. This system intends to help students receive a more fulfilling education and achieve their academic goals while graduating on time and within budget by providing greater personalized course guidance. This research project proposes the addition of an intelligent search bar in addition to the main features of the site, allowing students to query general subjects or specific concepts before returning the most relevant courses that go beyond simple keyword matching. To achieve this level of abstraction, we will be studying and leveraging machine learning models not typically used in an education setting, as well as evaluating their efficacy in this novel context to set a precedence in the field of learning analytics for both researchers and policymakers alike. AskOski can be scaled to other colleges and universities and offers immediate relevance to students in their academic careers, potentially impacting faculty and administration as well.

2 Background

2.1 Guidance in Higher Education

With an average student loan debt in America of $22,135 [1] students are paying more and taking longer than the targeted 4 years to graduate, especially at public schools. DeAngelo et al. [2] found that 6 years after matriculation, only 49.5% of students at public colleges had earned their degree compared with 78.2% at private universities. Part of the problem is attributable to matters of guidance, with a national adviser to student ratio of one to 400 [3]. Improved course selection navigation needs to be provided not only for post-secondary students at large, but also to those who would benefit more from additional guidance, such as transfer students from two year degree programs to a Bachelor's program, as well as underprivileged and first generation college students.

2.2 Representation Learning and Collaborative Filtering

Representation learning is the ability to learn the relational semantics of items based on behavioral data, and collaborative filtering is the process of evaluating items using the opinions of the crowd [4]. This project will leverage both ideas in representing courses as objects living in the same space, as learned through 2.2M course enrollment records. The primary models we will use to establish our intelligent search are skip grams and continuous-bag-of-words, which have obtained breakthrough results on a plethora of Natural Language Processing tasks [5]. In the natural language context, the idea is that words close together in the vector space formed by the model share a similar location due to the similarity of their word contexts and can be considered alike. Furthermore, the word vectors can be arithmetically manipulated while retaining its semantic properties, allowing the expression of relationships such as the archetype "King - Man + Woman = Queen." These ideas and models will be ported to the recommendation of classes, where courses close together in the vector space can be assumed to share features with one another and can also be arithmetically manipulated, e.g. "Biology Course" + " Engineering" should return a "Bioengineering Course." Recent research at CAHL have already identified skip-grams as useful in the context of MOOCs (Massive Open Online Courses) for identifying knowledge components in problem solving tasks [6], which are descriptions of the mental process a learner uses accomplish steps in a problem.

2.3 Synthesis

Different approaches have been taken in a range of institutions to improve the student course enrollment experience. One example takes into account institutional degree requirements as well as scheduling constraints of the student for courses being recommended [7], whereas [8] asked students to give their career goals and then rate courses for their workload and relevance to those goals, allowing other students to select courses based on those characteristics. However, our approach leverages representation learning in an attempt to portray courses and queries at a higher level, potentially revealing relationships hidden in course enrollment patterns not captured by explicit features. The addition of the intelligent search bar to the recommendation interface provides additional flexibility and personalization for designing a student's educational pathway, but more significantly, this project lifts cutting edge modeling paradigms from their native fields and potentially shows their utility and effectiveness in a pragmatic application to the education system.

3 Research Plan

The research project will consist of 3 main stages: data collection and preprocessing, machine learning, and lastly web development to incorporate the search feature into the website. Weekly or biweekly progress check-ins with my research supervisor will be held via video calls or in person meetings.

3.1 Data Collection

The goal of this stage is to collect and format the datasets necessary to train the machine learning models. The requisite data will be information on all classes currently and previously offered at UC Berkeley, including departments, course number, and their posted course descriptions. UCB's Student Information Systems offers access to this information through their APIs, which take a month to acquire the proper credentials, but can be attained before the summer. An older version of this dataset previously utilized by the lab can be used if there exist any delays, and the next two stages would proceed normally until more updated course descriptions can be secured. The other major dataset, anonymized student enrollment histories, has already been secured by the lab.

3.2 Machine Learning

3.2.1 Word2Vec

The first step in this phase is to gain an abstract representation of all the courses relative to each other by converting them to vectors living in the same space. This is completed using a skip-gram model known as Word2Vec that can map courses to vectors through student enrollment sequences. Figure 2 is a sample illustration of what we would like to achieve, replacing the students glyphs in the plot with courses. To evaluate the effectiveness of our model, we need to determine if closeness in this vector space translates back to courses that are actually similar in content. Since an objective notion of "close" does not exist in course catalogs, we can try to use classes deemed equivalent by the registrar by checking if the model accurately categorizes those as similar, and use heuristic domain knowledge about UCB courses to provide an additional sanity check.

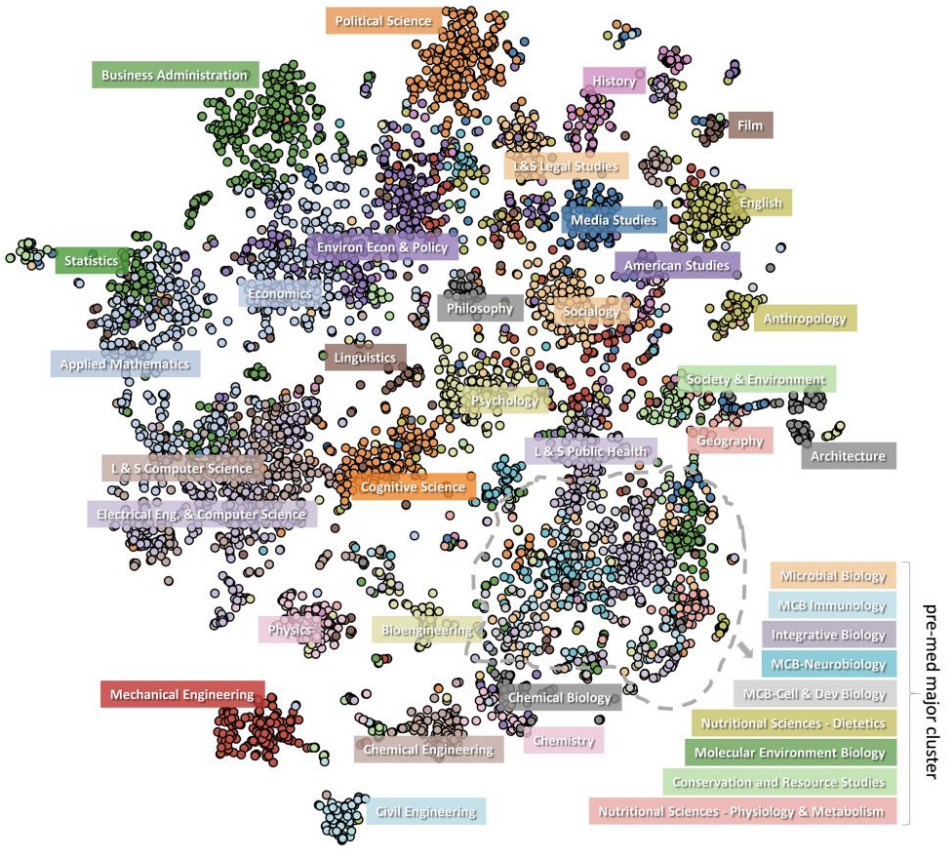

Fig 2: Projection of high dimensional vectors representing students enrollments onto a 2-dimensional space.

3.2.2 Bag-of-Words (BOW)

In order to bestow interpretability to this embedded space, the course vectors will be mapped using logistic regression to their corresponding BOW description, a way to represent text interpretable by the model. The specificity of words outputted by this mapping can be controlled by a bias parameter, so at different biases a separate set of keywords describing a particular course is outputted. For example, an economics course can be described at one level by (Markets, Pricing, Policy) and at another with (Industrial, Game Theory, Income), and so forth. The proper levels of granularity that will capture the essence of a course still have to be determined, but the ultimate objective of this stage is to collect these outputted keywords for each course at different bias levels and compile them into a separate keyword dictionary. This dictionary offers a semantic portrayal of the courses learned from behavioral inputs (student enrollment histories), and hopefully captures details beyond course titles and descriptions.

3.2.3 Application

Given a certain query from the user, our intelligent search can check for matches not only in the course description and title, but also in the newly produced keyword dictionary as well to return the best results. This additional layer can help mediate the difference in terminology students would use to describe the same subject or concept - e.g. an incoming student using more generic terms or a student from a different department unfamiliar with specialized jargon. A student interested in cross disciplinary fields now has the flexibility of searching multiple keywords such as “economics” + “sociology” or “genetics” + “programming” - “chemistry”, and our search offers the ability to generate vector representations of the query(s), perform vector arithmetic, and return the most relevant matches. Determining whether our intelligent search is effective represents a unique challenge, and some ideas include evaluating whether the keyword dictionary captured more than the course description by some metric, or if the courses produced by the query are more highly ranked as useful by an objective source than courses produced by a typical keyword search.

3.3 Web Development

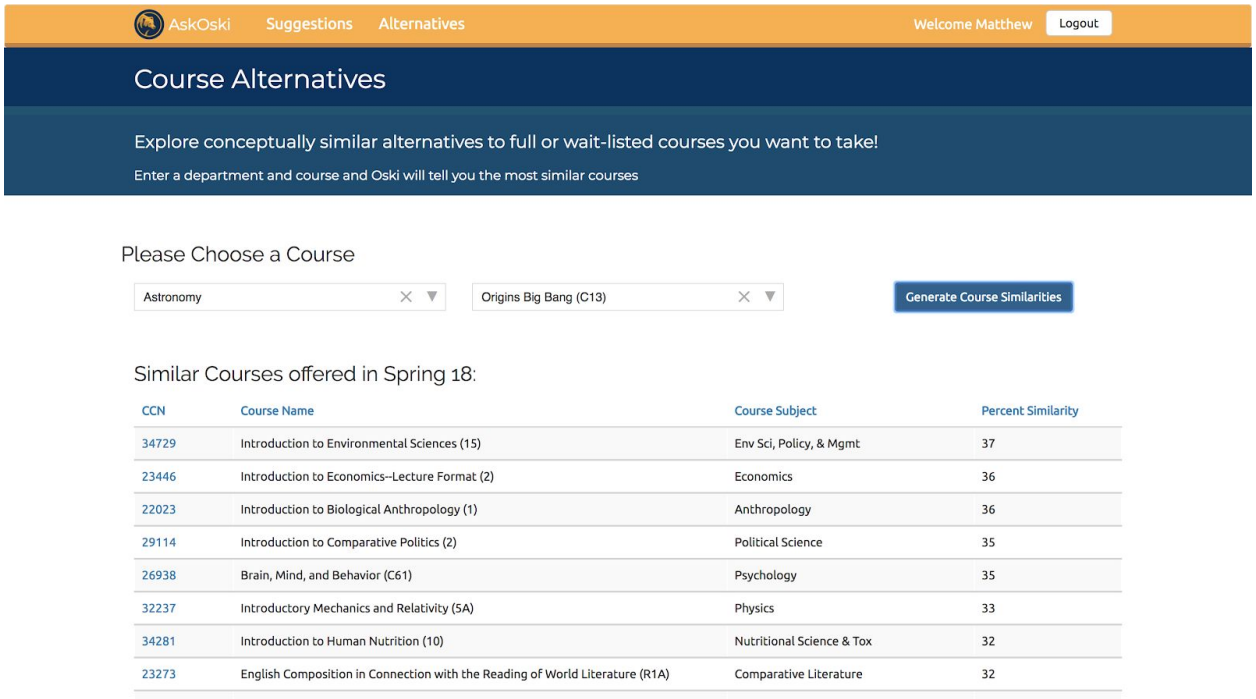

Fig 3: Beta testing of the course recommendation system.

In the current version of the the recommendation system (Fig 3), students have the option of receiving personalized course suggestions based on their enrollment history, similar student

profiles, and specified features, and can also view suggested course alternatives. Backend infrastructure already exists to support the learned skip-gram vector representations of the courses, so we would like to incorporate a search bar to the user interface under "Alternatives" where students can type in their own input(s) and receive a drop down list of top 10 courses most relevant to their query.

4 Qualifications

In previous summers I have engaged in related research projects in the same lab (CAHL) under the guidance of Zachary Pardos, the faculty sponsor for this proposal and the principal investigator of the lab. Thus, I am comfortable working under his tutelage and understand the expectations; I am also familiar with the requisite programming language (Python) as well as the additional packages (Pandas, NumPy, Gensim) and environments (Bash, IPython) needed to complete this endeavor. Furthermore, this semester I am enrolled in a research seminar led by Dr. Pardos to gain additional domain expertise and familiarity with pertinent research methods for this summer. This will also be a chance to form close connections with other members of the lab who may be able to provide feedback or guidance if needed. Additionally, I am taking a web development course, and the UI design skills and principles I learn will be useful during that stage of the proposal. As a statistics major, I have taken related classes in computer science, mathematics, statistics, and data science to ensure I have the technical skills and proper background necessary to carry out this endeavor. My personal engagement in this project is attributed to a belief in education as a human right, the intention of pursuing educational data mining as a career path, and a desire to use the revolutionary advances in our technological age to reform the shortcomings of the current education system.

References

1. Hess, Abigail. “Here's How Much the Average American in Their 20s Has in Student Debt.” CNBC, CNBC, 14 June 2017, www.cnbc.com/2017/06/14/heres-how-much-the-average-american-in-their-20s-has-in-student-debt.html.
2. DeAngelo, L., Franke, R., Hurtado, S., Pryor, J.H., Tran, S.: Completing college: Assessing graduation rates at four-year institutions. Los Angeles: Higher Education Research Institute, UCLA (2011)
3. Four-year myth. Complete College America. Indianapolis, IN (2014)
4. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: The adaptive web, pp. 291–324. Springer (2007)
5. Goldberg, Y.: A primer on neural network models for natural language processing. J. Artif. Intell. Res.(JAIR) 57, 345–420 (2016)
6. Pardos, Z.A., Dadu, A. (2017) Imputing KCs with Representations of Problem Content and Context. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP). Bratislava, Slovakia. ACM. Pages 148-155.
7. Parameswaran, A., Venetis, P., Garcia-Molina, H.: Recommendation systems with complex constraints: A course recommendation perspective. ACM Transactions on Information Systems (TOIS) 29(4), 20 (2011)
8. Farzan, R., Brusilovsky, P.: Encouraging user participation in a course recommender system: An impact on user behavior. Computers in Human Behavior 27(1), 276–284 (2011)