Paper Outline for Course Registrar

Vanessa Tang, Sangseok Lee, Yiran Chen (Becky)

10/16/2020

Hypothesis/Goal/Aim:

We aim to understand how past student class enrollment and completion data can help guide current students in class selection by constructing a personalized course recommendation system.

Objectives with (short, one-sentence) justification:

With a personalized course recommendation system, the current students will be able to see what classes the past students took in each major to help guide their own course selection. The information of the order as well as the combination of the classes taken in the past will help facilitate course load planning, order, and selection for future students leading to academic success.

Audience: Academic Audience

Style: White Paper

Abstract (Place holder; to be written only after rest of paper is complete and not in the outline)

Introduction

The abundance of courses available and a variety of academic requirements is often overwhelming for students. Undergraduate students must select courses that are relevant to their interest while being suitable for their academic background and graduation requirements. Furthermore, course selection can often depend on time, location, availability, professors, and recommendations from peers, to name just a few factors that contribute to course selection. In this project, we intend to provide a course recommender system integrated with existing DukeHub course scheduler to help students make better-informed course selections.

The proposed system aims to recommend courses whose content best matches the student's academic interests (major, minor, certificate, secondary), identify the student's academic pathways, and also presents predicted grades with respect to different course selections according to the student's academic history. Traditional recommendation systems such as movie or book recommenders produce a score for each item independently based on a variety of similarity metrics. On top of using the scoring scheme to show how desirable a certain course is for a student, the proposed recommender system will also try to integrate the academic requirements such as major or graduation requirements to make the recommender more practical and personalized.

Data

From Duke Registrar, we have two main datasets containing the same information but split into two time periods. The first time period is 2000 to 2012, and the second time period is 2012 to current. This is due to significant course renumbering and restructuring happening between Spring 2012 and Fall 2012. Unfortunately, this course renumbering in 2012 cannot be linked between the two datasets, meaning that there is no sure way to relate courses taken prior to 2012 to those taken after 2012. Because we have more complete data for the 2000-2012 segment, we are first focusing EDA and model-building on this segment of data. Furthermore, having a cleaner, more complete dataset facilitates model building. However, future practical applications of a course recommender system must integrate current data and course structure.

The degree dataset includes graduated students' de-identified ID's, graduation year, and major, minor, certificate, or secondary degree descriptions. The course enrollment dataset includes de-identified ID numbers that correspond to those in the degrees dataset in addition to course descriptions, grades received, and academic year for every course each student took at Duke. The data dictionary can be found in the Appendix.

Preprocessing Because there are two datasets with student information, we combined these datasets by merging on student ID (`Calculation ID`) to create a dataset with one row for every course taken per student and separate columns for major, minor, certificate, and secondary degree information merged from the degrees dataset. This is helpful because this dataset now has one row per item, which facilitates model-building for most recommender systems. Furthermore, this better integrates degree information as Duke students can have any three combination of major, minor, certificate, and secondary. This degree information may be helpful in future modeling as user-based features.

Exploratory Data Analysis

There are about 6,000 to 7,000 undergrad students enrolled at Duke every year. However, due to the structure of our datasets, the number of students incrementally decreases in the last few years as students "drop off" the dataset given the range of dates as students in 2010 graduate in 2014 but the dataset ends at 2012. These students are not included in this dataset. Thus, EDA and model building must account for inconsistencies at each end of the dataset's years.

Every student graduates with a major, but at Duke, students have the option of combining three of these options: major (MAJ), minor (MIN), secondary (SEC), and certificate (CER). Because there are many different combinations of majors, minors, secondary, and certificates and all students must graduate with a major, we plan to first focus on course planning for a student's major.

Focusing on major only, we start by using two subsets of data to build a prototype: one with students majoring in Economics who started at Duke in 2005 and graduated in 2009, and one with students who major in the five most popular majors also from year 2005 to 2009. Using these smaller, cleaner datasets can help facilitate model building in the early stages.

Dynamic Course Dashboard

One intuitive way to map students' academic pathway is based on the most most popular classes taken during a given semester for a given major. We created a Tableau dashboard that shows popular classes for a selected semester for a selected major. It also provides an option to select courses for a given graduation year, as course patterns change over time, so viewing only recent graduates' course history is most reliable. This helps students in course selection to see what most recent graduates with the same major have taken in the semester of interest.

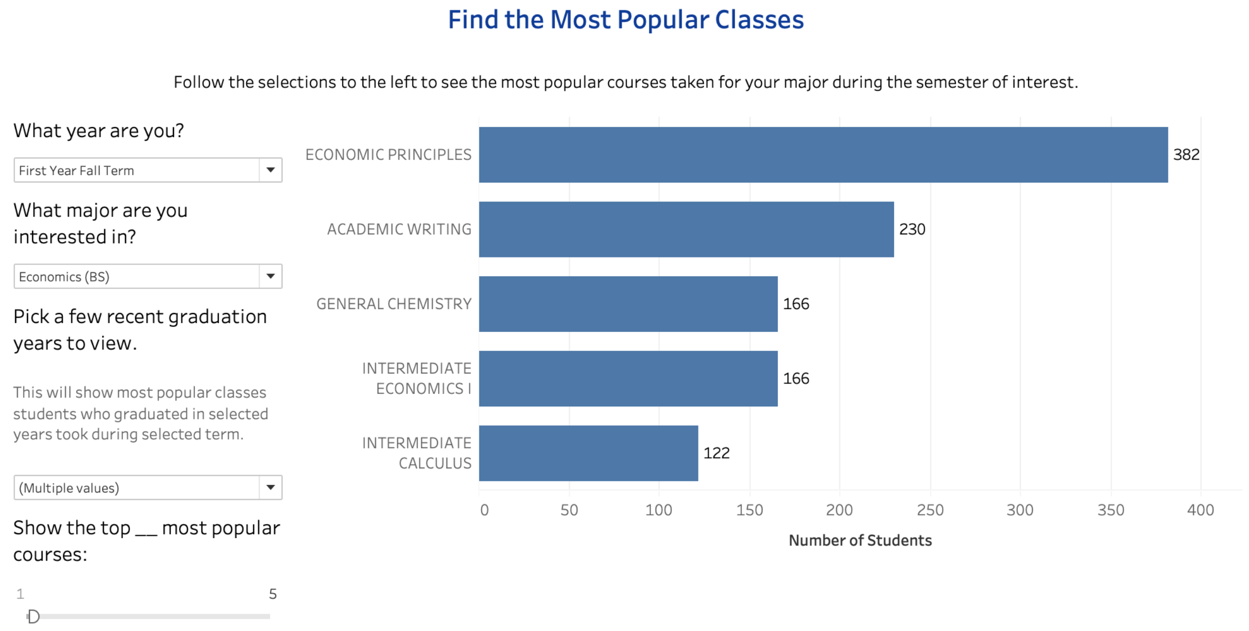
To facilitate long-term course planning, a Pathway dashboard is also created. This shows the 5 most popular courses to take for each term from Freshman to Senior year given a certain major, which can help a current student plan course selection for their entire time at Duke accordingly. Likewise, this dashboard provides an option to only select recent graduates' course history to account for course pattern changes that occur over time.

Methods

Item-based

Purpose of section: We start with computing the item-item relationships of the classes. Our final goal here is to construct a new item by item matrix containing the weights (relationships) between each of our classes where a perfect correlation equals 1 and no correlation at all equals 0.

Preprocessing 1. Normalize data The normalization should be applied before we do our similarity calculations, so instead of having a rating as ones or zero we have some value between 0 and 1 representing the



Find the Most Popular Classes

Follow the selections to the left to see the most popular courses taken for your major during the semester of interest.

What year are you?

First Year Fall Term

What major are you interested in?

Economics (BS)

Pick a few recent graduation years to view.

This will show most popular classes students who graduated in selected years took during selected term.

(Multiple values)

Show the top __ most popular courses:

Figure 1: Dashboard to view most popular classes for the selected major and term.

importance of that users rating.

$$magnitude = \sqrt{(x^2 + y^2 + z^2 + ...)}$$

2. Caculate similarities We then create the new unit vector by dividing the rating by the magnitude:

$$vector = (\frac{x}{magnitude}, \frac{y}{magnitude}, \frac{z}{magnitude}, ...)$$

User-based

1. Define a neighborhood of items from the item-based approach as above.
2. Calculate the score for all items for a specific user. To get the score, we use this formula.

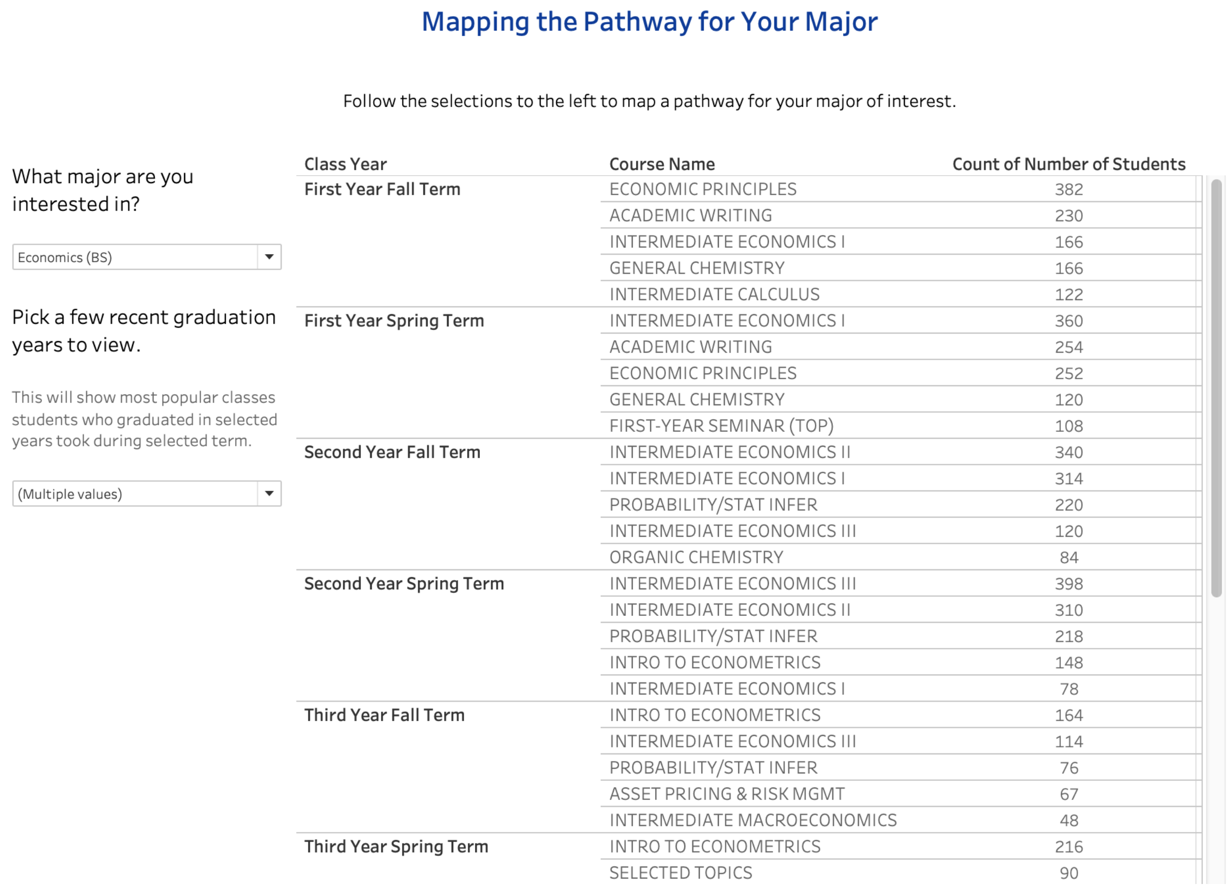
$$S(u,i) = \frac{\sum_{j \in N} W_{ij} r_{ui}}{\sum_j |W_{ij}|}$$

we get the score for user u and item i by summing together all the weights for that item W_{ij} multiplied with the users rating for that item r_{ui}. We then divide by the sum of all the weights for that item W_{ij}. We limit our calculations to a neighborhood of only the n most similar items to the users own ratings and then do the scoring on those. That's what the j ϵ N in the numerator above describes. It says “for items j in neighborhood N”.

3. Sort by the n highest scores (most recommended)

Hybrid Algorithms

Purpose of section: We will explore hybrid recommendation algorithms which combine both the concepts of content-based recommendation models and collaborative filtering. These algorithms are able to solve the cold start problem, where recommender systems struggle to recommend for new users, which is commonly seen in recommendation settings. They also work better for new data utilizing similarities based on user features.



Mapping the Pathway for Your Major

Follow the selections to the left to map a pathway for your major of interest.

What major are you interested in?

Economics (BS) ▾

Pick a few recent graduation years to view.

This will show most popular classes students who graduated in selected years took during selected term.

(Multiple values) ▾

Class Year	Course Name	Count of Number of Students
First Year Fall Term	ECONOMIC PRINCIPLES	382
	ACADEMIC WRITING	230
	INTERMEDIATE ECONOMICS I	166
	GENERAL CHEMISTRY	166
	INTERMEDIATE CALCULUS	122
First Year Spring Term	INTERMEDIATE ECONOMICS I	360
	ACADEMIC WRITING	254
	ECONOMIC PRINCIPLES	252
	GENERAL CHEMISTRY	120
	FIRST-YEAR SEMINAR (TOP)	108
Second Year Fall Term	INTERMEDIATE ECONOMICS II	340
	INTERMEDIATE ECONOMICS I	314
	PROBABILITY/STAT INFER	220
	INTERMEDIATE ECONOMICS III	120
	ORGANIC CHEMISTRY	84
Second Year Spring Term	INTERMEDIATE ECONOMICS III	398
	INTERMEDIATE ECONOMICS II	310
	PROBABILITY/STAT INFER	218
	INTRO TO ECONOMETRICS	148
	INTERMEDIATE ECONOMICS I	78
Third Year Fall Term	INTRO TO ECONOMETRICS	164
	INTERMEDIATE ECONOMICS III	114
	PROBABILITY/STAT INFER	76
	ASSET PRICING & RISK MGMT	67
	INTERMEDIATE MACROECONOMICS	48
Third Year Spring Term	INTRO TO ECONOMETRICS	216
	SELECTED TOPICS	90

Figure 2: Dashboard to view top 5 most popular classes for the selected major from First Year Fall Term to Fourth Year Spring Term.

MatchBox MatchBox is an on-line learning algorithm capable of incrementally taking account of new data so the system can immediately reflect the latest user preferences. Developed by Microsoft, it is an easily implemented algorithm that can provide similar users as well as similar items.

LightFM LightFM is one established hybrid matrix factorization model. It learns embeddings for users and items in a way that encodes user preferences over items. When multiplied the two matrices together, these representations produce scores for every item for a given user.

Results

We will be focusing on validation metric of accuracy for top k recommended courses, using the prior three years' data of a successfully graduated student and predict for the fourth year course selections.

From the LightFM model, the accuracy is fairly low to predict the correct courses in the fourth year. It might be due to the fact that there is a limited number of courses taken each year even though many course predictions can be made. In addition, the model might find it difficult to distinguish between courses with similar names but in different domains, or courses that share different names but lie in the same subject. This would require further investigation on the word embeddings on the courses to form representative cluster structures.

Unlike normal recommendation system where there are no constraints on the items recommended, in the academic setting, we need to exclude the courses already taken, and take the temporal aspect of courses into consideration. The sequence of courses across the four year have significant importance, mainly due to prerequisites. We aim explore other modeling techniques that can account for this temporal aspect.

Figures/tables: > Figure of the dashboard included above

Future Work

We will be investigating different types of models as well as explore better representations of course text information using NLP techniques.

1. Explore text embeddings: Get Word2vec using SKipgram, RNN, or LSTM
2. Insert temporal variation (per semester) in Word2Vec
3. Connect 2000-2012 data with 2013 - 2020 data since the course description and code is different from each other

Conclusion/Discussion

(Place holder)

References

1. Morsomme, R., & Alferez, S. V. (2019). Content-Based Course Recommender System for Liberal Arts Education. International Educational Data Mining Society.

Appendix

Data Dictionary

- Calculation ID: anonymized student ID
- Acad Plan: shorthand/abbreviation for major/minor/certificate/secondary
- Acad Plan Descr: description of acad plan (major/minor/certificate/secondary)
- Plan Type: denotes major/minor/certificate/secondary
- Degree: abbreviated type of degree (AB, BS, BSE)
- Descr_completions: type of degree (Bachelor of Arts, Bachelor of Science, Bachelor of Science in Engineering)

- Comp Term Descr: academic year and term of graduation
- Acad Year_completions: academic year of graduation
- Subject: subject of course
- Catalog: course number
- Descr_enrollment: name of course
- Grade: grade received in descr_enrollment
- Term Descr: academic year and term of enrollment in class
- Acad Year_enrollment: academic year of course enrollment
- Enrollment Start: first year student has registered for classes at Duke
- Term Year: year student is taking course
- Semester Term: term student is taking course (Fall, Spring, Summer 1, Summer 2)
- Class Year: indicates freshman through senior year (or more) and term (First Year Fall Term, First Year Spring Term, etc.)