Duke Registrar Recommender System
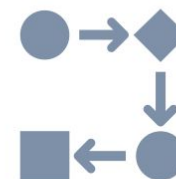
Vanessa Tang, Sangseok Lee, Yiran Chen

Project Overview

Course recommendation system for undergraduate students

Current course selection process is complex and time-consuming

Personalized course recommender for course planning

How?

Student's planned majors and minors

Course history from other students within the same major

Academic pathway

Challenges

- No information on:
  - Graduation/major requirements
  - Course: prerequisites, availability, ratings, professors
  - Success metric: career interests

- Students don't have to declare until beginning of junior year

- Time: course order matters

- Separated two time frames
  - Course renumbering
  - No explicit linkage

Data

- Student level data including:
  - All courses taken at duke
  - Grade received
  - Department
  - Course number/name
  - Academic year taken
  - Major(s) // Minor(s) // Certificate(s) // Secondary
  - Academic year of graduation
- Added the following features:
  - Enrollment year
  - Semester course taken (First Year Fall)
  - Numerical grades

Popular Classes for Selected Major and Semester

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 ▼

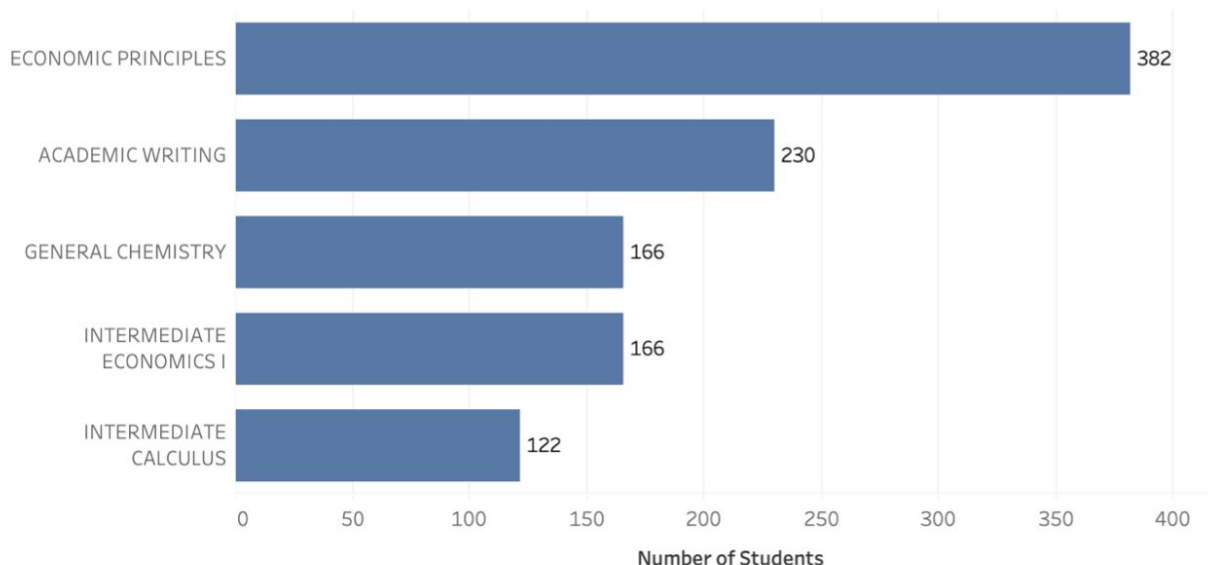**What major are you interested in?**

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:

Course	Number of Students
ECONOMIC PRINCIPLES	382
ACADEMIC WRITING	230
GENERAL CHEMISTRY	166
INTERMEDIATE ECONOMICS I	166
INTERMEDIATE CALCULUS	122

Number of Students

Pathway

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

Modeling

- Data
  - Economics BS 2005-2009
  - Top 5 Majors 2005-2009:
    - Biology, Economics, Public Policy, Biomedical Engineering, Political Science
  - Train and test sets split by semester

- Collaborative Filtering: Neighborhood Methods

- Metric
  - Accuracy = $\dfrac{\text{Number of Courses Taken in Top 10 Recommendations}}{\text{Total Number of Courses Taken (Senior Year)}}$

Neighborhood Methods

$$cos(\theta) = \frac{A \cdot B}{\|A\| \ \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

Item Based → Neighborhood → User Based

Cosine Similarity

	ACADEMIC WRITING	ACCEL GENERAL CHEMISTRY	ADV FIN & MGRL ACCOUNTING	ADV INTERMED FR LANG/CUL	ADV INTERMEDIATE SPANISH	ADV SPANISH WRITING	ADVANCED CALCULUS I	ADVANCED CHINESE	ADVANCED KOREAN
ACADEMIC WRITING	1.000000	0.481208	0.362603	0.231847	0.392254	0.484160	0.499679	0.164869	0.145704
ACCEL GENERAL CHEMISTRY	0.481208	1.000000	0.031193	0.131527	0.075656	0.712739	0.735604	0.000000	0.038905
ADV FIN & MGRL ACCOUNTING	0.362603	0.031193	1.000000	0.037772	0.110384	0.047303	0.149293	0.000000	0.049743
ADV INTERMED FR LANG/CUL	0.231847	0.131527	0.037772	1.000000	0.000000	0.000000	0.048076	0.000000	0.000000

Neighborhood Methods

	1	2	3	4	5	6	7	8
ACADEMIC WRITING	ACADEMIC WRITING	INTRO TO ECONOMETRICS	INTERMEDIATE ECONOMICS III	PROBABILITY/STAT INFER	INTERMEDIATE ECONOMICS II	ECONOMIC PRINCIPLES	INTERMEDIATE ECONOMICS I	INTERMEDIATE CALCULUS
ACCEL GENERAL CHEMISTRY	ACCEL GENERAL CHEMISTRY	FIRST-YEAR GERMAN I	HONORS SEMINAR II	LINEAR ALGEBRA & DIFF EQUATION	HONORS SEMINAR I	ADVANCED CALCULUS I	LECTURES SPECIAL TOPICS	ELEMENTARY ITALIAN 2
ADV FIN & MGRL ACCOUNTING	ADV FIN & MGRL ACCOUNTING	FINANCIAL ACCOUNTING	MANAGERIAL FINANCE	FIRST-YEAR SEMINAR (TOP)	MANAGERIAL EFFECTIVENESS	INTERMEDIATE ECONOMICS II	ACADEMIC WRITING	CORPORATE FINANCE
ADV INTERMED FR LANG/CUL	ADV INTERMED FR LANG/CUL	FR FOR CURRENT AFFAIRS	INTRO TO OPERATING SYSTM	COMPUTER ORGANIZA/PROG	SOFTWARE DESIGN/IMPLEMEN	SP TOP: INTRO TO LIT	INTERMED FRENCH LANG/CUL	COMPETITIVE STRAT & INDUS ORG
ADV INTERMEDIATE SPANISH	ADV INTERMEDIATE SPANISH	INTERMEDIATE SPANISH	INTERMEDIATE ECONOMICS II	FIRST-YEAR SEMINAR (TOP)	SPAN FOR ORAL COMMUNICA	INTERMEDIATE ECONOMICS I	ACADEMIC WRITING	INTERMEDIATE CALCULUS

Neighborhood Methods

ADV FIN & MGRL ACCOUNTING	0.814720
INTRO TO CULTURAL ANTHRO	0.695917
SELECTED TOPICS IN ECON (TOP)	0.677204
THE CREATIVE MIND	0.666135
INTERMEDIATE CALCULUS	0.660741
LABORATORY CALCULUS I	0.657436
ASSET PRICING & RISK MGMT	0.654018
INTERNATL ECONOMY, 1850-2000	0.652133
ORGANIC CHEMISTRY	0.626705
ENTREPRENEURSHIP	0.616361

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

u = student
i = course
j = course taken
W = weights (0/1)
r = item-based matrix

Results

- Accuracy = $\dfrac{\text{Number of Courses Taken in Top 10 Recommendations}}{\text{Total Number of Courses Taken (Senior Year)}}$

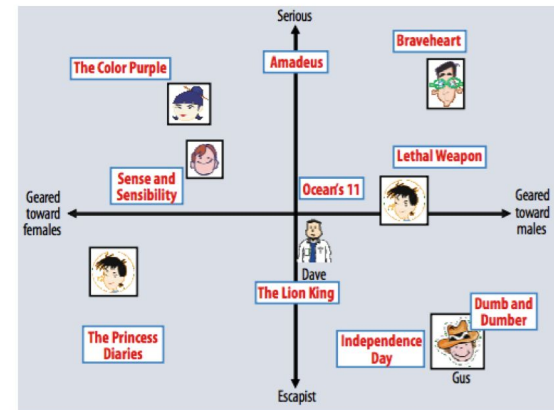|	Fourth Year Spring	Fourth Year Fall	Third Year Spring
Econ Subset	67%	17%	16%
Top 5 Majors	56%	13%	13%

Future Work: Short Term

- Cross Validation

- Latent Factor Methods

- Recommenders for implicit datasets
  - Implicit library
  - Dataiku

1. Neighborhood Methods
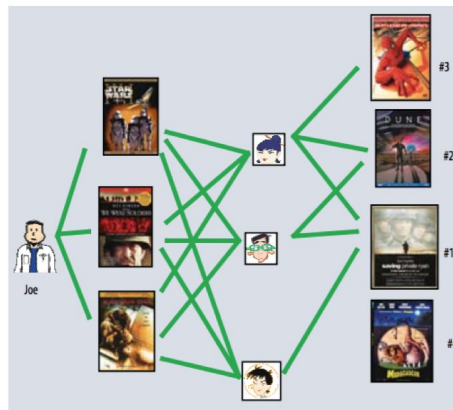
2. Latent Factor Methods

Future Work: Long Term

- Narrow down search space for recommending courses

- Incorporate temporal aspect of data

- Efficient vector representations for courses

- Increase generalizability of the model

- Integrate with Dataiku platform

Thank You!

Appendix

Matrix Factorization

- Singular Value Decomposition (SVD)

$$R = U\Sigma V^T$$

R: predicted student grades
U: student grades
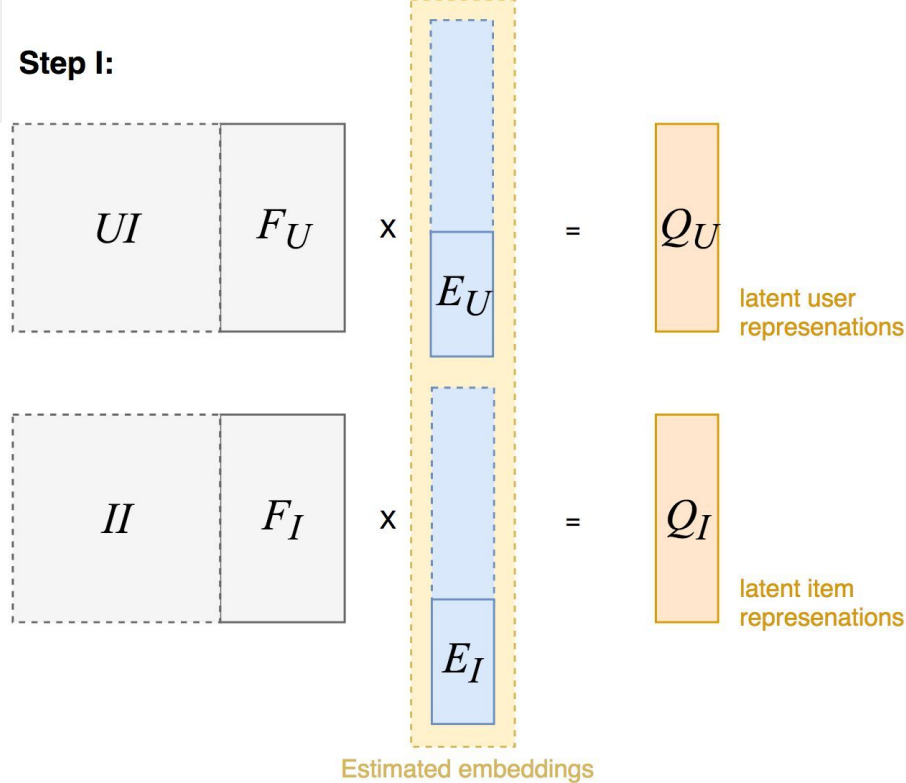Σ: diagonal matrix of singular values (weights)
V^T: courses

- LightFM
  - Incorporate both item and user metadata into the traditional MF algorithm
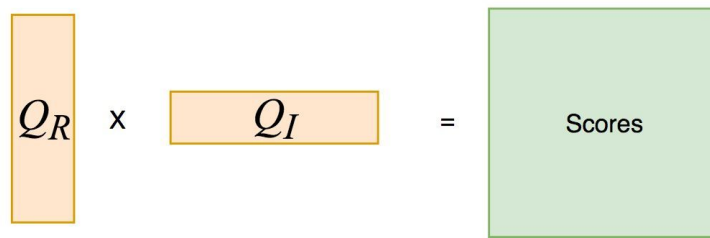
LightFM

- Hybrid matrix factorization model
- Content-based: User // Item feature
- Collaborative: Interaction matrix
- Learns embeddings for users and items in a way that encodes user preferences over items
  - When multiplied together, these representations produce scores for every item for a given user
- Pros:
  - Cold start

Step I:

$$UI \quad F_U \quad \times \quad E_U \quad = \quad Q_U$$

latent user represenations

$$II \quad F_I \quad \times \quad = \quad Q_I$$

$$E_I$$

latent item representations

Estimated embeddings

Step II:

$$Q_R \quad \times \quad Q_I \quad = \quad \text{Scores}$$

Results

- Accuracy = $\dfrac{\text{Number of Courses Taken in Top 10 Recommendations}}{\text{Total Number of Courses Taken (Senior Year)}}$

	Fourth Year Spring	Fourth Year Fall	Third Year Spring
Econ Subset	67%	17%	16%
Top 5 Majors	56%	13%	13%

- Rarely-captured: Research Independent Study, Projects, Selected Topics
- Well-captured: Intro Biochemistry, Asset Pricing & Risk Mgmt, Electrobiology

Neighborhood Methods

ADV FIN & MGRL ACCOUNTING	0.814720
INTRO TO CULTURAL ANTHRO	0.695917
SELECTED TOPICS IN ECON (TOP)	0.677204
THE CREATIVE MIND	0.666135
INTERMEDIATE CALCULUS	0.660741
LABORATORY CALCULUS I	0.657436
ASSET PRICING & RISK MGMT	0.654018
INTERNATL ECONOMY, 1850-2000	0.652133
ORGANIC CHEMISTRY	0.626705
ENTREPRENEURSHIP	0.616361

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

u = student
i = course
j = course taken
W = weights (0/1)
r = item-based matrix