Duke Class Recommendation

- AIPI 540

Secret Santa: Abhishek, Jared, Mrinoy, Samyukta

Table of contents

Santa!

Santa is coming to
bring you all classes!

1.
Introduction

Motivation

1. Vast Choices
2. Time Consuming
3. Need for a Proactive Approach

Objective

Simplify the course selection process by implementing a course recommendation system that is

- Smart
- Intuitive
- Tailored

Previous Efforts

01

BridgeU

Online tool matching prospective international students with universities based on courses offered. Universities must subscribe to be included.

02

leverage edu

Assists students in finding study abroad opportunities, and associated courses

03

Individual schools - Class lookups based on filtering and keyword search

No results in public domain for this type of recommendation system

2.

Dataset

Data Sources

Student

Built sample data including:

- Desired career
- Program of study
- Top 2 hobbies
- Gender
- Country of origin

Courses

Scraped from Duke catalog API:

- Subject name
- Course catalog
- Course title
- Course description
- Prerequisites

Matched 1 student to 5 courses

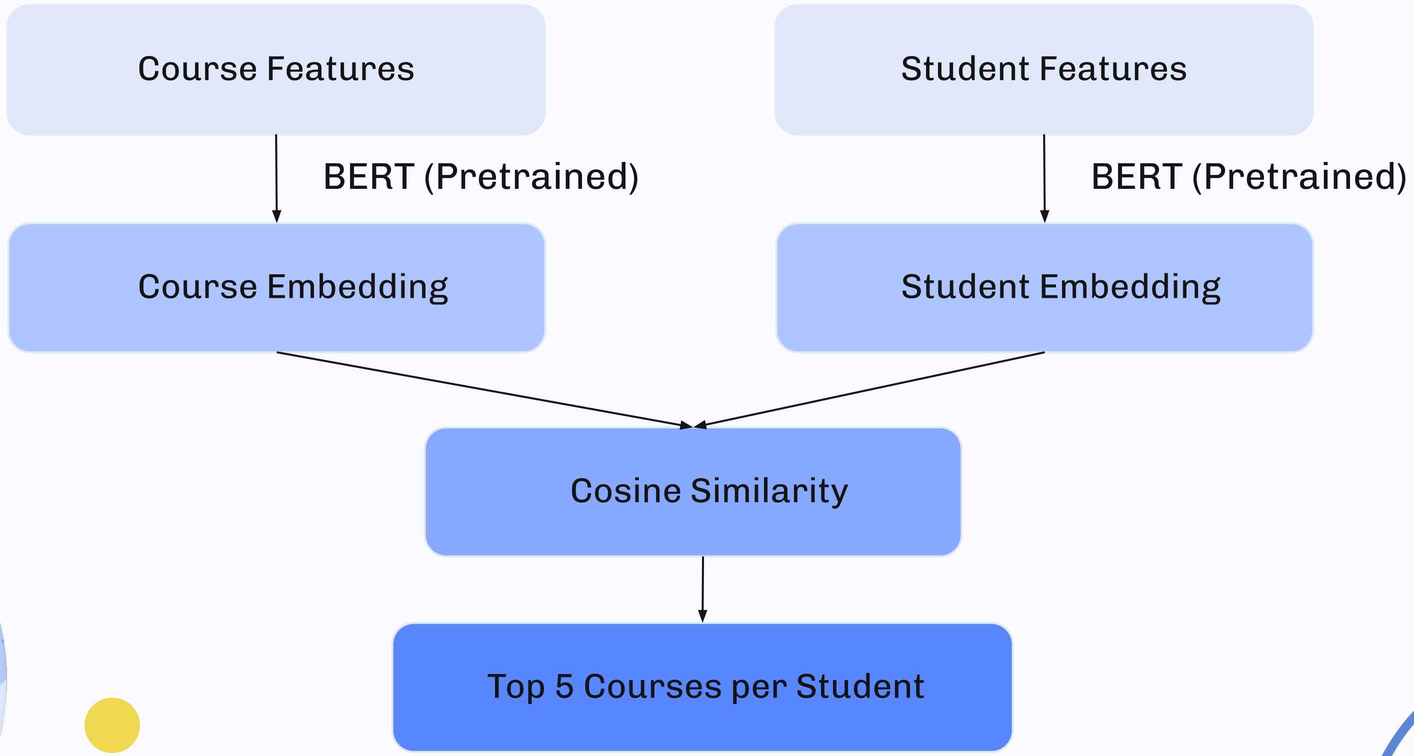
3.

Model Architecture

Non-Deep Learning Approach: TF-IDF

1. Extract Course Descriptions
2. Stop Words & Lemmatization
3. TF-IDF Matrix
4. Cosine Similarity with Student Features
5. Top 5 Courses

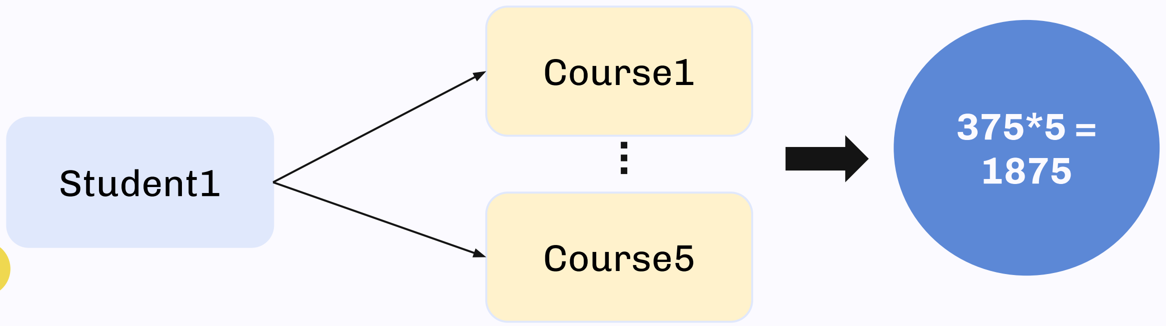
Naive Method: BERT



Naive Method: BERT

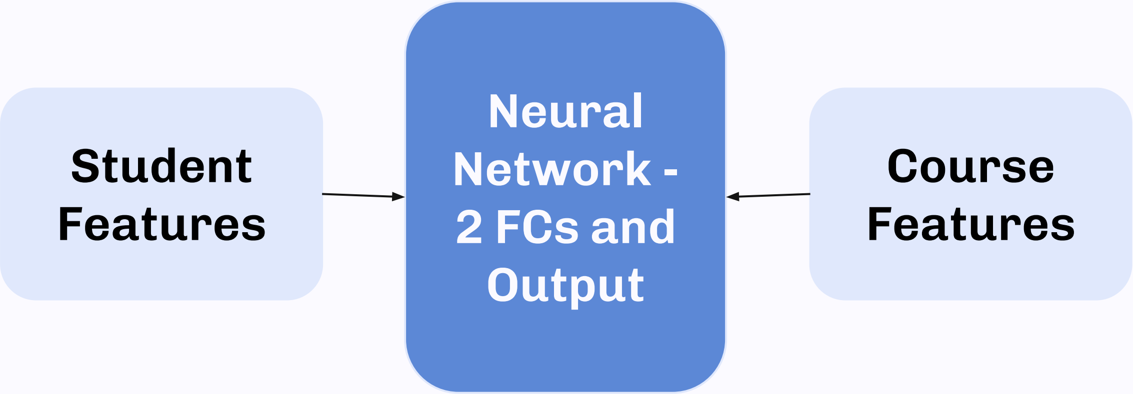
Example -

Biomedical Engineering	Dance	Drawing	Biomedical Research	['Compiler Construction', 'Computer Architecture', 'Writing about Performance', 'Special Topics in Electrical and Computer Engineering', '**Data Science**']
Civil Engineering	Wood Working	Hiking	Structural Engineer	['Physical Chemical Processes in Environmental Engineering', 'Compiler Construction', '**Data Science**', 'Computer Architecture', 'Special Topics in Electrical and Computer Engineering']



Deep Learning Approach: NCF - MLP variant

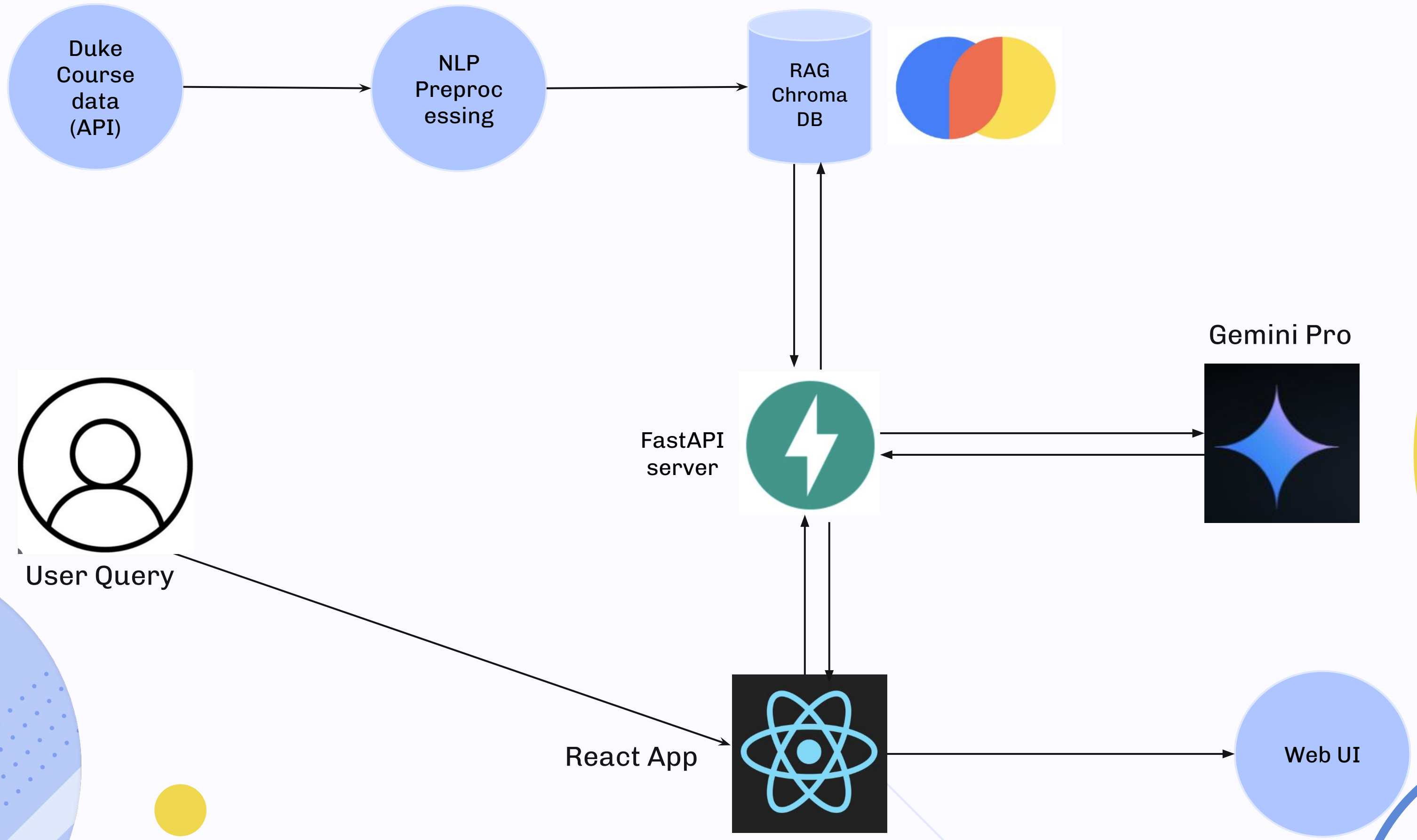
Rationale behind NCF is that MLPs are general function approximators, making them potentially better than a fixed similarity measure like a dot product.



RAG

1
Stop Words & Lemmatization
2
Word embeddings into Chroma DB
3
Gemini Pro with prompt engineering
4
API Exposure for inference
5
React app consumes api

4.
Pipeline



Duke Course data (API)
NLP Preprocessing
RAG Chroma DB
Gemini Pro
FastAPI server
User Query
React App
Web UI

5. Evaluation

Key points

Difficult to evaluate using only labeled training data

- Similar students may differ in their opinions of the same recommendations
- Evaluation on individual classes or whole group of classes

Evaluation

- Validation – We voted
- Test – You vote

Precision

- Increase the number of relevant courses suggested

Validation Results

Model	Precision
TF-IDF (Non-DL)	68%
BERT (Naive)	46%
NCF	84%
RAG	82%

6.

Demo

Voting Round

Test Results

Number of Acceptable Recommendations	Number of Votes
5	
4	
<=3	

Discussion

- Cold Start Problem -Addressing data scarcity by leveraging language models to generate synthetic data.
- We changed our minds a lot on user-user, item-item, and user-item similarity selection
  - There are many ways to solve the need for recommendation
- Evaluation may need to be done separately, such as done today. Hard to quantify results.
- Pulls from Duke API, so if a new class is added, we will have that data readily available.
- Simple can be good enough - TF-IDF vs BERT

Future Scope

- Expand course corpus for both graduate and undergraduate programs.
- Develop user profiles and track course histories for personalized recommendations.
- Implement rating system for course feedback.
- Integrate with Dukehub for direct course registration.

Thank you!