Group 19

UCSB Course Recommendation System

Phuc Lu, Yuchen Zhu, Calder Glass, Jiaxin Su

03/17/2025

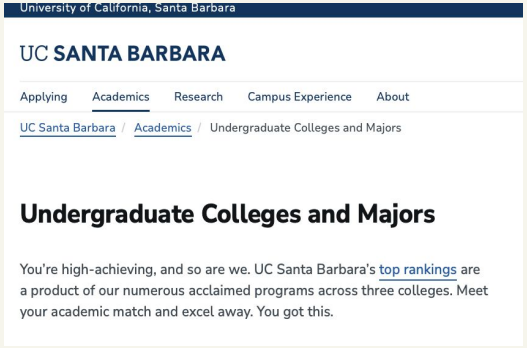
PSTAT 127 - ADV STATS MODELS							Units:4.0 Grading:Optional
44180	TR	11:00 AM-12:15 PM	ILP, 1302	MEIRING W	Space 28	Max 100	Course Info / Final
44206	W	10:00 AM-10:50 AM	Psychology East, 1806	XIONG MOYA	Space 6	Max 25	Add
44214	W	11:00 AM-11:50 AM	Phelps Hall, 1529	XIONG MOYA	Space 4	Max 25	Add
44222	W	12:00 PM-12:50 PM	Phelps Hall, 1529	WANG YAQIAO	Space 12	Max 25	Add
59659	W	1:00 PM-1:50 PM	Phelps Hall, 1529	WANG YAQIAO	Space 6	Max 25	Add
PSTAT 130 - SAS BASE PROGRAM							Units:4.0 Grading:Optional
44230	R T	9:30 AM-10:45 AM 9:30 AM-10:45 AM	ILP, 2101 Remote instruction	SWENSON J S	Space 4	Max 125	Course Info / Final
44248	T	11:00 AM-11:50 AM	Phelps Hall, 1525	JIN Z	Space 3	Max 25	Add
44255	T	12:00 PM-12:50 PM	Phelps Hall, 1525	JIN Z	Space Full	Max 25	Add
44263	T	1:00 PM-1:50 PM	Phelps Hall, 1513	ZHU YUDONG	Space Full	Max 25	Add
44271	T	2:00 PM-2:50 PM	Phelps Hall, 1513	FENG E	Space Full	Max 25	Add
44289	T	3:00 PM-3:50 PM	Phelps Hall, 1513	FENG E	Space 1	Max 25	Add
PSTAT 131 - STAT MACHINE LEARN							Units:4.0 Grading:Optional
44347	TR	5:00 PM-6:15 PM	LSB, 1001	BARACALDO LAN	Space 16	Max 80	Course Info / Final
44354	M	10:00 AM-10:50 AM	Phelps Hall, 1525	LIU ROBIN	Space 3	Max 20	Add
44362	M	11:00 AM-11:50 AM	Phelps Hall, 1525	LIU ROBIN	Space 4	Max 20	Add
44370	M	12:00 PM-12:50 PM	Phelps Hall, 1525	AUNG T	Space 6	Max 20	Add
44388	M	1:00 PM-1:50 PM	Phelps Hall, 1525	AUNG T	Space 3	Max 20	Add
PSTAT 134 - STAT DATA SCIENCE							Units:4.0 Grading:Optional

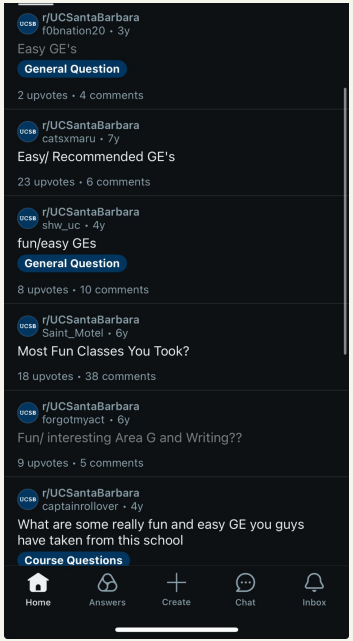
Problem

- UCSB offers a liberal arts education
  - Vast variety to courses in different subjects.

- No official way for students to get tailored list of courses that match their interest outside of their major.
  - besides friends or Reddit

- **Goal**: Help with course searching by creating a course recommendation system that recommend UCSB courses to students based on their interests.

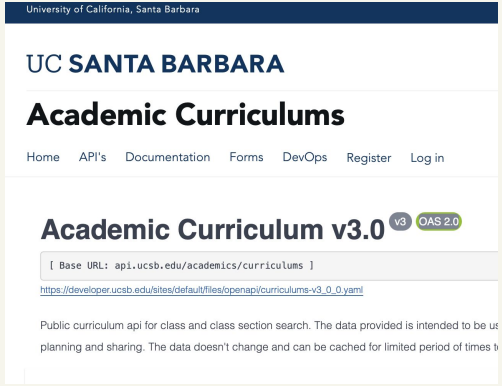




API and Courses Data

- Academic Curriculum v3.0 API + key

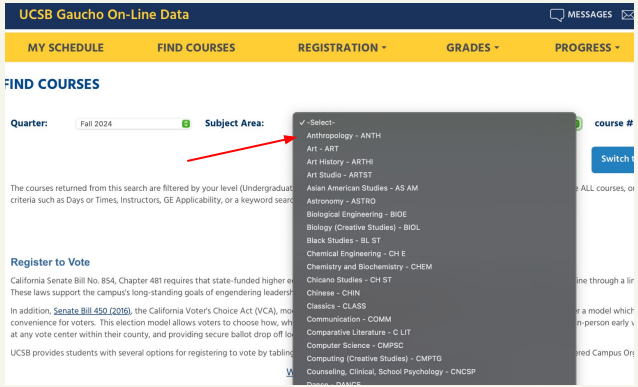
  - Example Url:
    `https://api.ucsb.edu/academics/curriculums/v3/classes /search?quarter=20244&subjectCode=PSTAT&pageNumber=1&pageSize=500&includeClassSections=true`



- Fall 2024 Course Data
  - the quarter, courseId, title, department code, course description, the subject area, course level, which college, number of units, the instruction type, whether the course is online.

- Stored as a large dataframe .csv file.

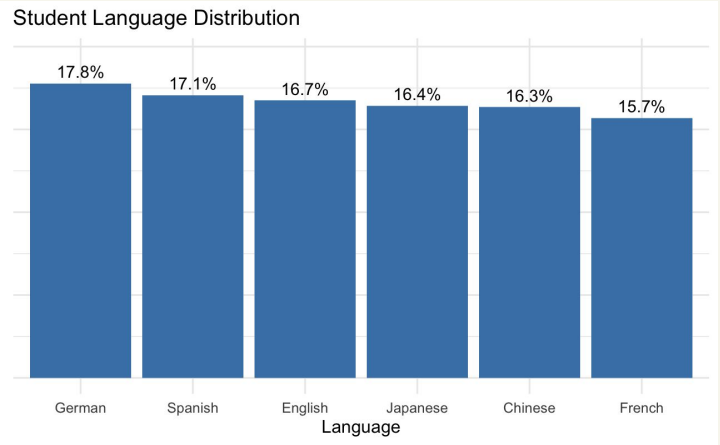
quarter	courseId	title	dept_code	description	subject_area
20244	PSTAT 5A	UNDERSTANDING DATA	STATS	Introduction to data science. Concepts of statistical th…	PSTAT
20244	PSTAT 5H	STATISTICS	STATS	Introduction to a variety of statistics and applied prob…	PSTAT
20244	PSTAT 5LS	STAT LIFE SCIENCES	STATS	An introduction to statistics for students interested in…	PSTAT
20244	PSTAT 8	TRANS DS PROB STAT	STATS	Introduction to techniques of mathematical logic, pro…	PSTAT
20244	PSTAT 10	DATA SCIENCE PRINC	STATS	Fundamentals of programming for data science using…	PSTAT
20244	PSTAT 100	DS_CONC&ANLS	STATS	Overview of data science key concepts and the use of …	PSTAT
20244	PSTAT 115	BAYES DATA ANALYSIS	STATS	An introduction to the Bayesian approach to statistica…	PSTAT
20244	PSTAT 120A	PROB & STATISTICS	STATS	Concepts of probability; random variables; combinato…	PSTAT

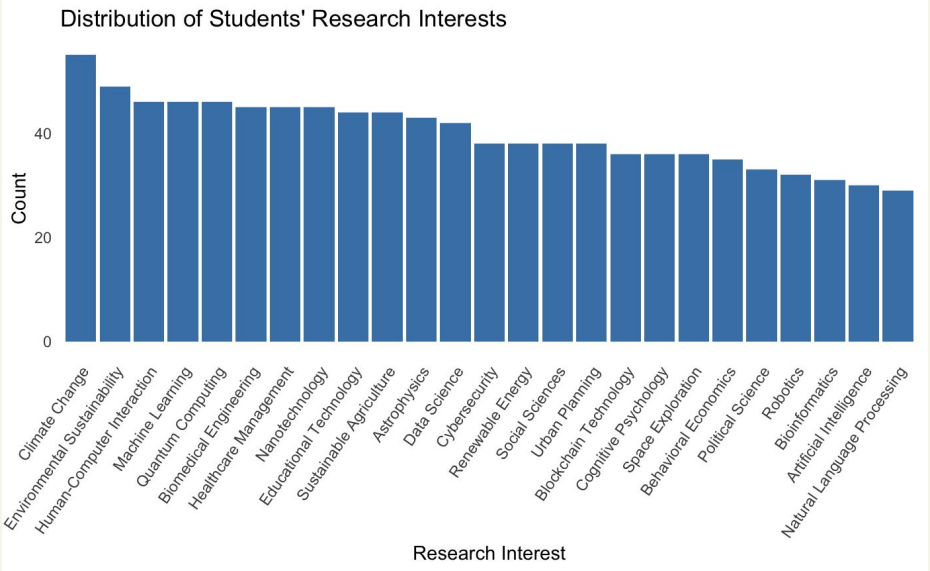


Student Data

- Kaggle Dataset: “Student Extracurriculars Info”
  - Include information about students’ academic interest, major, etc

StudentID: A unique identifier for each student.
Name: The name of the student.
AcademicInterest: The field of study or academic interest of the student.
ExtracurricularActivities: Extracurricular activities the student is involved in.
Skills: Skills possessed by the student.
Location: The city or location where the student is currently based.
YearOfStudy: The year of study for the student (e.g., Freshman, Junior, Senior, Graduate).
Major: The major or field of study that the student is pursuing.
GPA: The Grade Point Average of the student.
Languages: Languages spoken or known by the student.
ClubMemberships: Memberships in various clubs or organizations.
ResearchInterests: Research interests or specialization of the student.





Methodology- Recommender System

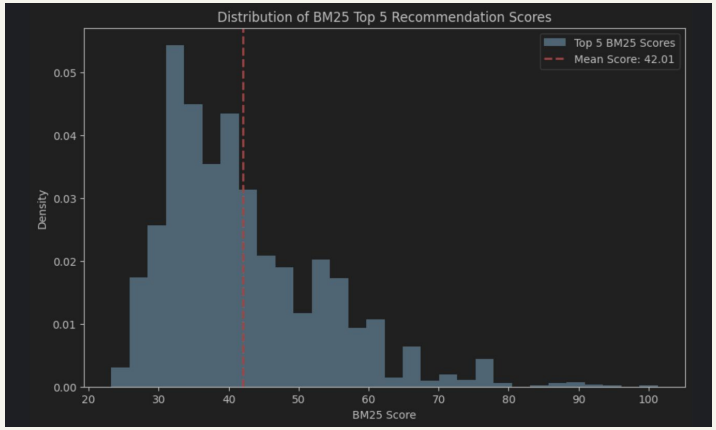
- **Content -based Filtering Methods**
  - No historical data for each student's course selection behaviors
  - Recommend based on students' academic and extracurricular interests, and the course description
- **Merge Students information into one merged column**
  - academic interest, extracurricular activities, skills, major, club memberships, research interests, and the languages students speak.
  - Give different weight to each type of students' interests to give more emphasis on important information
- **Tokenization of Course Description into Words**
- **Similarity Algorithm – BM25**
  - Recommend the top 5 courses that have the highest BM25 scores

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Results

Model performance

- BM25 Score Distribution of Top 5 Recommendations
- Histogram Analysis
- Mean Score ~42



Example Input

```
Student Information:
--------------------
Name: Student 620
AcademicInterest: Computer Science
ExtracurricularActivities: Coding Club
Skills: Programming, Public Speaking
YearOfStudy: Sophomore
Major: History
GPA: 3.06
Languages: French, Japanese, Spanish, Chinese
ClubMemberships: Coding Club, Music Club, Volunteer Group, Art Club, Sports Team
ResearchInterests: Cognitive Psychology
```

Example output

```
Top 5 Recommendations:
  Course ID: CMPSC     9  , Title: INTERMEDIATE PYTHON, Score: 41.5192
  Course ID: FR      111  , Title: GREAT FR SPEECHES, Score: 35.4095
  Course ID: PSY     108  , Title: COGNITIVE PSYCHOLOG, Score: 35.2420
  Course ID: CMPSC     8  , Title: INTRO TO COMP SCI, Score: 35.1722
  Course ID: CMPSC    99  , Title: INDEPENDENT STUDIES, Score: 34.3286
```

Conclusion & Discussion

- Worked with an API and Kaggle to data.
- Build a course recommendation system.

- **Weakness**:
  - No access to student data consistent to the courses attributes & minimal student information.
  - Leads to naive recommendation system.

- **Future Direction:**
  - Improvement:
    - Implement a better way to collect student information
    - Good start would be previous coursework.
  - Apply pretrained model
    - Pretrained Embeddings model: deeper semantic capture
    - Apply API from LLM such as Chat-GPT:
      - Advanced NLP reasoning to analyze students' interests
      - **Hybrid Pipeline**: Shortlist ~30 courses from generated result, refine with LLM.