Transforming the UCSD Course Search Experience with Semantic Search

Misa Franknedy
mfrankne@ucsd.edu

Tony Ta
ttta@ucsd.edu

Colin Jemmott
cjemmott@ucsd.edu

Abstract

In the fast-paced academic environment at the University of California, San Diego (UCSD), students often face challenges in regards to finding courses that align with their interests and academic needs due to the limitations of WebReg, the current course search system. WebReg's search application relies solely on exact matches of search queries to course titles, limiting the effectiveness of course discovery for students. This project introduces a novel semantic search engine designed specifically for UCSD courses, leveraging state-of-the-art embeddings to capture the context of the search queries and matching them with the most similar courses. Our approach not only addresses the flaws in the current search system, but also provides a more intuitive and efficient course search experience for UCSD students.

Search App: http://ucsd-course-search.westus2.azurecontainer.io:8000/
Website: https://toekneeta.github.io/UCSDCourseSearch/
Code: https://github.com/toekneeta/UCSDCourseSearch

1 Introduction

At UCSD, a university known for its diverse student body with vastly different career goals and interests, the ability to plan their academic path is vital to their success and satisfaction with their education. Due to the large variety of courses offered at UCSD, it is important for students to be able to search through the list of classes that are offered at UCSD to find the ones that match their aspirations. However, the current course search system used at UCSD, known as WebReg, is very restrictive for students, as it only finds exact matches from the search query within the course titles. This leaves less room for broader, more contextual searches, meaning that students end up having a difficult time finding the courses they are looking for. Our project intends to improve on WebReg's search capabilities by creating a new search application that utilizes semantic search to interpret the nuances in what the users are searching for. Our search application aims to understand what the students are looking for and present results that are more pertinent to the students. Our approach demonstrates a marked improvement in search relevance and user satisfaction, indicating significant progress towards the goal of helping students plan their academic journeys in a simple and stress-free manner.

1.1 Description of Data

Our project uses 3 datasets, all of which were web scraped to obtain the necessary data:

1. The **UCSD Course Catalog**, which contains information on every course offered at UCSD, found here.
2. The **UCSD Schedule of Classes for Spring 2024**, which contains which courses are offered in Spring 2024, found here
3. The **CAPE (Course And Professor Evaluations) reviews** of all UCSD courses, found here

The UCSD course catalog is our primary dataset and acts as the set of documents that our search engine will be parsing through. The course catalog contains the different courses offered at UCSD, a detailed description of the topics that are taught in the course, and the prerequisites required for students to be able to take the class. After obtaining all the course catalog data, we found that there were 7,169 courses across 173 departments. 59.4% of the courses found in the course catalog were undergraduate courses, with 19.1% of the undergraduate courses being lower-division and 80.9% being upper division, and 40.6& of the courses were graduate level courses. Our dataset is somewhat limited, however, since not all departments are listed in the course catalog. A majority of the missing departments are from professional programs and are all graduate-level courses.

The UCSD schedule of classes dataset contains the list of courses that will be offered during spring quarter of 2024. We planned to release our search engine and website for public use in time for the spring quarter registration period. Thus, we decided that in our search results, the courses that are being offered for the upcoming quarter should be highlighted to let students know if they will be able to take that class. There are 2,075 total courses

being offered in Spring 2024, with 1,921 of those courses being in departments that are found in the UCSD course catalog. Of those 1,921 courses, 1,730 were found in the course catalog.

The CAPE reviews dataset contains student evaluations of every course offering from Summer 2007 until Spring 2023. The CAPE reviews was replaced with Student Evaluation of Teaching (SET). The SET reviews are less informative than the CAPE reviews, as they remove a lot of useful information regarding the student evaluations of the professor and course compared to CAPE reviews. CAPE reviews included student ratings of the courses and professors, the average grade in each course offering, the average number of hours spent on the course, among other information, that made it very popular compared to SET reviews amongst students. Thus, we thought that providing the CAPE reviews of courses would be more insightful to the students than using the SET reviews.

2 Methods

2.1 Search

In developing our search function, we first experimented with ElasticSearch (Gormley and Tong 2015) to get a baseline as the model uses an inverted index but does not employ semantic search. We then tested two embedding models: FlagModel (Zhang et al. 2023) and Msmarco-distilbert (Bonifacio et al. 2021) to create our own embeddings of the course title and description. Of the two transformers, we found more success with the FlagModel, which was the final model used in our search function.

Our search mechanism begins by processing the user's query along with other search parameters, such as filters and the desired number of results. The initial step involves a filtering process that discards irrelevant entries based on the specified filter criteria. The filtered dataset is then inputted into our search function, which employs cosine similarity to compare the query's embedding representation against the embeddings of both the titles and descriptions of courses. For each course, scores from the title and description are assigned weights and aggregated to calculate a final similarity score, with titles receiving a greater emphasis. The final step involves returning and displaying the results on our website, with the quantity of displayed results being determined by the user's initial selection of the number of outcomes desired.

2.2 Hosting

The website is hosted on Microsoft Azure and is built using Flask, along with conventional HTML, CSS, and JavaScript. The website features a search functionality equipped with filters, presenting the outcomes in a structured box layout that details course code, title, description, prerequisites, and a link to CAPES. Users can refine their searches with filters for upper division, lower division, and graduate-level courses, and have the option to include

or exclude specific departments in their search results. Additionally, users can review the relevance of each result by clicking either a thumbs up or thumbs down button, which is then logged onto an Azure SQL Database. This data is later used as our testing set when evaluating our search models.

2.3 Public Release

Following the completion of the search functionality and the website development, we created a Reddit post on the UCSD subreddit to publicize our search tool for students to use to search for courses. Our post was intentionally timed to be posted a few days before course registration for the spring 2024 quarter was supposed to begin, so that students could use our search tool to look up classes they might be interested in taking. The response from the community was exceptionally positive, and some of the feedback was used to help us improve the user experience. Within the first week following the release, the search tool registered over 1,200 searches, providing a substantial dataset for analysis. From this influx of user interaction, we were able to extract 145 unique queries that would form the basis of our test set. This test set would be used to conduct an evaluation of our search model's performance.

3 Results

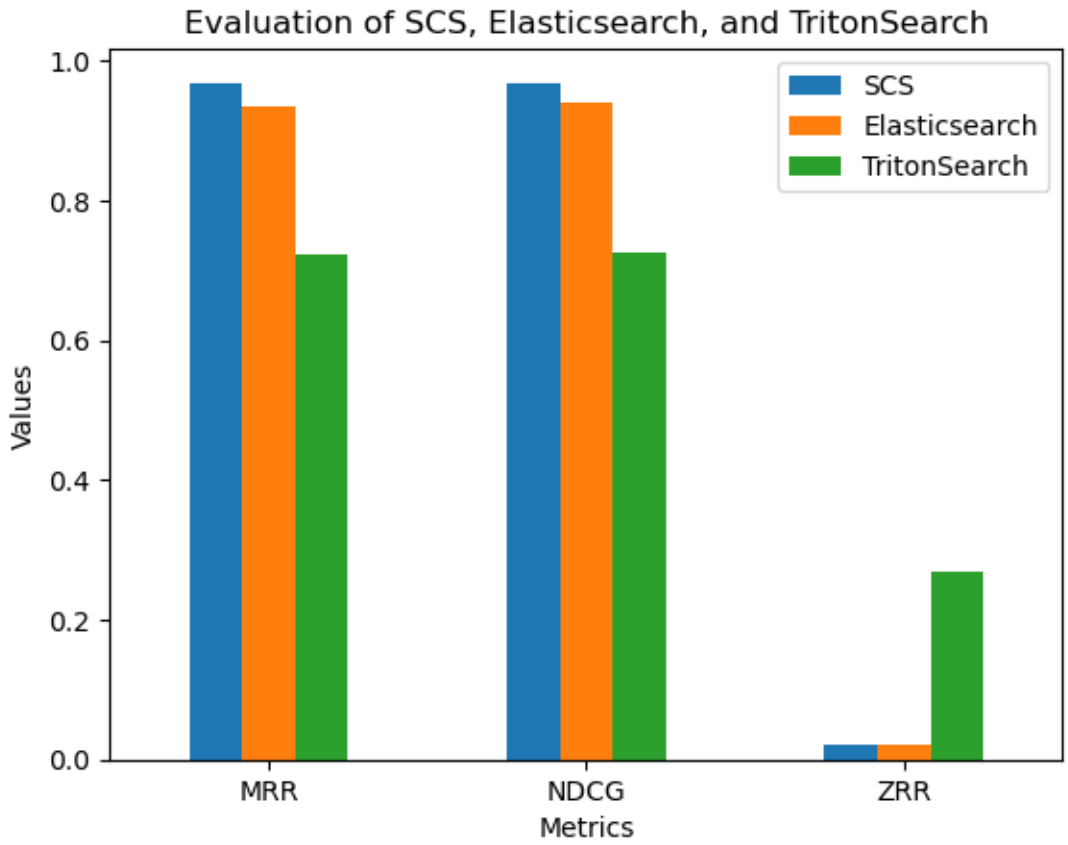
We evaluated our semantic course search (SCS) model along with a baseline Elasticsearch and another student-created course search called TritonSearch (Navani 2023) using three metrics: mean reciprocal rank (MRR), normalized discounted cumulative gain (NDCG), and zero results rate (ZRR).

We chose to use MRR and NDCG as they both take into consideration how many relevant results are returned, as well as the positioning of relevant results. Additionally, we chose to use ZRR to show the limitations of the various search methods. We used a small k=5, assuming that the user will primarily look through the top few results.

The first two metrics were calculated by taking the top 5 results for each of the 145 test queries, along with each result's relevance score (0=not relevant or 1=relevant). ZRR was measured by calculating what percentage of queries returned zero relevant results.

Table 1: Comparison of SCS, Elasticsearch, and TritonSearch

Metric	SCS	Elasticsearch	TritonSearch
MRR@5	**0.9690**	0.9345	0.7241
NDCG@5	**0.9684**	0.9404	0.7265
ZRR	**0.0207**	0.0219	0.2690



4 Discussion

From the table above, our SCS model performed the best out of the three models: it has the highest MRR@5 and NDCG@5 scores, as well as the lowest ZRR score. Looking more closely at the zero response rate, it highlights the recall of the various models, namely how TritonSearch failed to return any relevant results for 27% of the queries, which is a main factor in its low scores in MRR and NDCG. This brittleness in search (seen first and foremost in UCSD's WebReg) was the motivating factor for our project, and we are glad to see our model regularly returns relevant results to the user.

We originally planned to use user feedback as evaluation (e.g. thumbs up/down); however, due to a lack of user interaction with the buttons, we instead made a larger test set for assessment. We recognize that our hand-labeled relevancy may be biased and that 145 test queries may not be enough to be representative of a model's performance, but we hope that the computed scores give insight into the performance of each search engine.

5 Conclusion

The development and deployment of our semantic search tool for UCSD courses represents a significant advancement in the accessibility and efficiency of course discovery for students. This project has not only demonstrated a considerable improvement over the existing UCSD course search tools, but has also notably outperformed both our baseline model and TritonSearch across all three of the measured evaluation metrics: NDCG, MRR, and ZRR. This underscores the effectiveness of the semantic understanding that was integral to the development of our search tool.

Upon its release to UCSD students, the search tool quickly proved its value and relevance, as evidenced by the overwhelmingly positive feedback received from its users. With over 1,200 searches conducted within the first week alone, our search tool has had an immediate impact on the student experience. Overall, our search tool provides students with the means to effectively navigate their course options, thereby highlighting the potential of this tool in helping to improve academic planning at UCSD.

References

Bonifacio, Luiz Henrique, Israel Campiotti, Roberto Lotufo, and Rodrigo Nogueira. 2021. "mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset."

Gormley, Clinton, and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc."

Navani, Aarav. 2023. "TritonSearch." [Link]

Zhang, Peitian, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. "Retrieve Anything To Augment Large Language Models."