Major Recommender Project

DATA 322

Tamu Suttarwala

Introduction

Cal Poly Humboldt's admissions department is seeking an interactive and innovative tool that can recommend a suitable major to incoming students. Such a tool could potentially improve admission rates to the college, reduce drop-out rates, minimize switching of majors, and improve overall satisfaction with higher education.

My focus in building such a tool is to provide a slightly longer-term satisfaction. For this project I aim to design a major recommender that helps narrow down the best match based on career objectives. Specifically, based on the job roles associated with a major, as well as companies, agencies, or institutions associated with it. This way, the tool helps elucidate the user's preferred career path. Additionally, it can encourage them to continue on their suggested academic journey, by enlightening them of the career waiting for them after graduating.

Methodology

Beginning with Professor Overholser's code as reference, many changes were made in various stages of the machine learning tool. A comprehensive list, with a brief explanation, of the stages and the changes is as follows:

1. Data collection: For a list of occupational roles and companies to work for, I downloaded all the pdf files available on the Career Exploration webpage. This was done using the Python packages *selenium* and *BeautifulSoup*. Extensive aid was provided by *ChatGPT*. The code is linked here.
2. Data processing: Text was extracted from these pdfs and cleaned for further use. The text in the pdfs was unfortunately not in an easily extractable format, therefore plenty of formatting code was needed over and above the basic extraction performed using the *PyMuPDF* package. A brief description of the cleaning performed is listed below:
    a. Removing non-ASCII characters (square-shaped characters)
    b. Splitting illegitimate CamelCase words that were a result of improper pdf formatting, for e.g. splitting "Policy AnalystData Specialist" to separate lines: "Policy Analyst" and "Data Specialist"

c. The splitting mentioned above had to be done carefully in order to avoid acronyms from splitting, as well as splitting two or more acronyms if they are found to have whitespaces absent between them
d. Long lines with multiple sentences were split into separate lines with a sentence each. This was done for clarity needed when reviewing the text files for errors
e. Exceptions also had to be incorporated for legitimate CamelCase words such as "BioTech" or "AmeriCorps"
f. Excess white-spaces were removed, and blank lines were eliminated
g. Known phrases with irrelevant text were removed from the dataset. For e.g. lines containing the phrase "Career Guide:" were removed, since they are present in all pdf files, and do not contribute to the clustering objective.

It was found that the pdf for Journalism was outdated and did not have the information needed. So, this file is removed and the program does not appear in the results. Extensive aid was provided by *ChatGPT*. The code is linked here.

3. Data Analysis: Finally, the files are matched against each other for similarity in lines – if the text in a line from file A matches (or is a subset of) a line from file B, it is counted as a match. The higher the number of line-matches between two files, the higher the similarity. This measure is used to create a distance (or dissimilarity) matrix, which when fed to the linkage function from the *scipy* package, creates a dendrogram. Extensive aid was provided by *ChatGPT*. The code is linked here.
4. Interaction: This step has not been completed, but here I describe how I would develop it. I would begin by adding a second layer of linkage between the majors and their descriptions, which would help an unknowledgeable user to interact with the tool. An interactive questionnaire can be used, where the user chooses *k* of *n* words they relate with, and as a feedback mechanism, they could be shown some of the job roles and/or companies they could work for. If they like the examples from the feedback, the questionnaire would adapt to narrow user choices further, if not then we would try choices from an alternative cluster.

Evaluation Plan

This being an unsupervised learning tool, it is difficult to gauge the accuracy or effectiveness through simple means. Additionally, there is no historical data available that could be tailored to evaluate such a metric either. Regardless, the results do seem to adhere closely to the logic I intended to use. I also sampled a few major-pairs to check if the matches make numerical sense, and they seem to be consistent.

The closest branches seem to have a predictable pattern, for e.g. Theatre Arts and Dance Studies, or Wildlife and Fisheries. It is at the higher levels, when looking at over 4 major clusters that we see unexpected groupings happen. This was my intention; finding unlikely connections between majors based on the workplace one might end up many years after they graduate. For e.g. A Math graduate might be working alongside a Social Work graduate, which makes the question, "If I'm going to end up at my dream workplace X in the dream job role Y either way, should I start entertaining the idea that I'd rather major in B instead of A?". Established notions of higher education and professional journeys can be challenged using this tool.

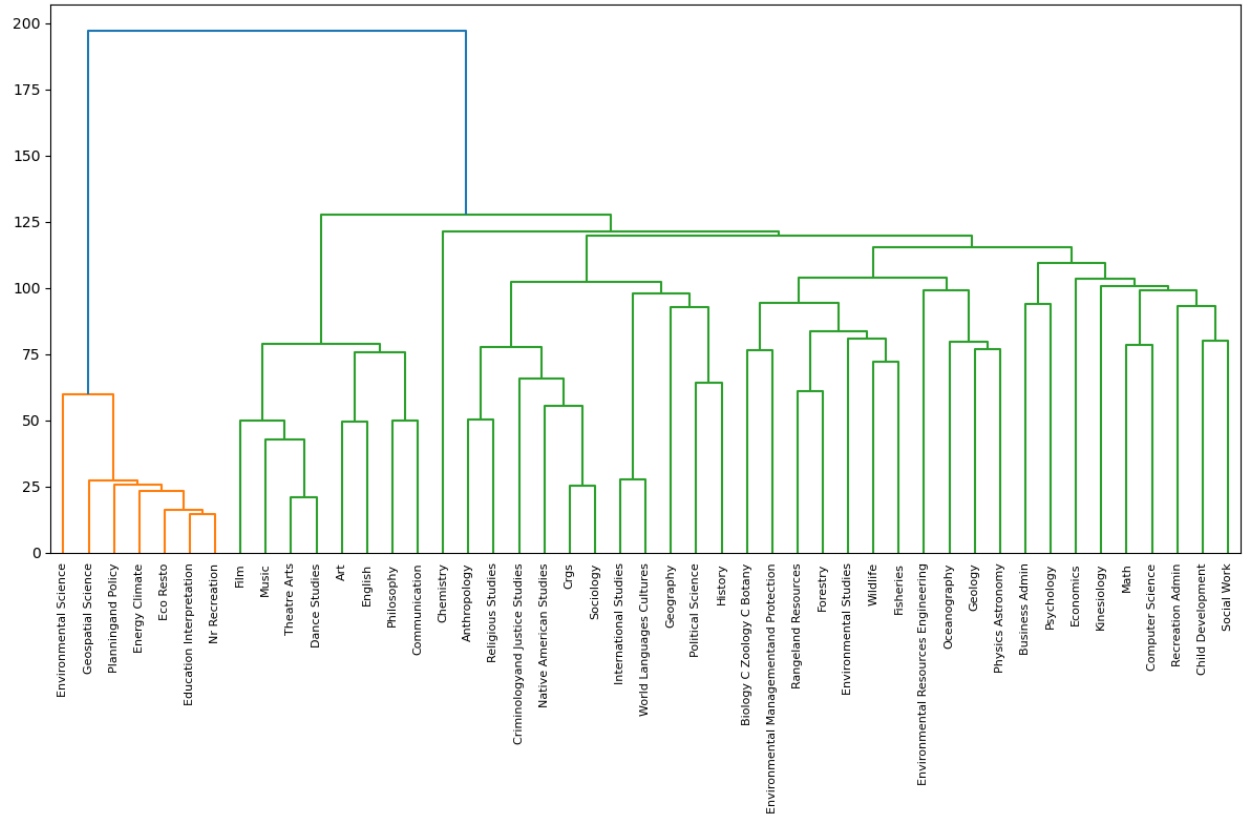
When compared to Professor Overholer's clustering (setting aside the fact that I had access to limited data, limiting me to clustering 46 majors instead of all the 100+ plotted in her tool), as mentioned above some of the results are along expected lines and match her work, while others are completely different. For instance, Economics is not close to Math or Computer Science, it is instead clustered with History, Geography, and Political Science. This is illustrative of the fact that the methods and tools that economists use to do their work closely match with the tools mathematicians and computer engineers use. However, the work that economists do tends to intersect with what historians and political scientists do and the domains of society that they affect, which is something my algorithm exposes.

For a more thorough evaluation however, a user study would be required. A randomly selected experiment group would use this tool, and every year their satisfaction with major selection would be surveyed against that of a randomized control group. It would also be important to record how significantly the tool changed people's minds. Someone who dreamed of being a geologist, using the tool to conclude studying geology, and satisfied with their geology major is probably not strong evidence in favor of the tool's efficacy.

Results & Discussion

The final cluster that includes all the majors is shown in the figure below. The features mentioned above can be noted here. While the unsupervised learning process itself turned out to not be very challenging, the data collection and processing stages required meticulous attention to detail. It made me realize that the most sophisticated machine learning methods are ineffective without a clean, complete, and accurate data set.

I am also happy with the diversity of clustering that the results show – a mix of predictability and unpredictability.



Conclusion

In many ways my modification with the input data is a definite improvement. The only drawback is the lack of more data. At the core of my approach is the belief that higher education is about more than just the tests and the assignments, it's about employability and professional acumen at the end of the academic journey.

To that end, more can certainly be done. Other variables that could be considered for the model are job growth rate, job market demand, median salary, and even work-life balance. Truly, with the right kind of data and enough of it, much more can be done to improve the tool.

Code Appendix

Link to pdf downloader

Link to pdf scraper

Link to clustering algorithm