COMP 523: SOFTWARE ENGINEERING LAB

CREDDIT HOURS - Project Proposal

Spring 2023

Justin Lee Do

Abstract

This document is a proposal for a project seeking clients for the Spring 2023 offering of COMP 523 - Software Engineering Lab. The project is an application that will leverage web-scraping and NLP techniques to design a smart system that can gauge insights into the health of the UNC community through an examination of the UNC subreddit. Desired features include: metrics and visualization surrounding contributions from individual users, a recorded history of all activity on the subreddit, and anomaly detection for posts that generate significant activity or extreme sentiment. The scope of the project is flexible (for example, the desired shape and tech stack of the application has not been determined) and I intend to work closely with my team to define requirements and ensure that they have the skills they need to successfully develop this project.

1 Project Summary

As a former moderator of the UNC subreddit and someone who still frequents the subreddit (I was living out-of-state and curious about what was going on at UNC), I have noted that the sub often serves as a good pulse for how the overall student body is doing. It skews towards certain demongraphics (notably non-traditional students and STEM students), but it has a large readership and it has been fascinating to read throughout events such as COVID and the UNC CS admissions process.

I propose a web-scraping application that uses NLP and other heuristics to get a "pulse" of what the heartbeat of the subreddit is and flag potential areas for concern. We should assume the best intentions of everybody in the community, but as an anonymous forum rooted in a real community, the subreddit can have the potential to cause harm in the real world. For example, a student might be dissuaded from visiting CAPS or another professional resource based on advice shared by another user who had a bad experience with university resources. At the same time, the subreddit has the ability to uplift and encourage those in need of encouragement or advice.

Since the community has little in the way of moderation (just a few unpaid moderators), it would be beneficial to have an external tool that can intelligently gauge community participation, highlight positive posts, and flag problematic posts before they get out of hand and result in the spread of misinformation or other real-world consequences.

2 Desired Features

What the project will look like (i.e. the tech stack) has not been determined and may be catered towards the interests and experienced of the team. However, at minimum, I would like the following features to be implemented in a full-stack application:

- A continuously running web-scraper that logs all activity on the subreddit (e.g. new posts, comments, upvotes, downvotes) and leverages it to determine whether something is receiving an abnormal amount of attention or positive/negative sentiment.

- A notification system to notify when an event described in the prior heading occurs.
- Some sort of leaderboard indicating the positivity of a user's contributions (i.e. the number of upvotes and downvotes they have received).
- A way to visualize activity on the subreddit.

The scope for this project may be expanded upon over the course of development, but here are some additional considerations that one might make:

- What tools already exist in this area and what does the existing Reddit API allow us to access? If a good chunk of the features described here already exist, then there's no need to reinvent the wheel and our time would be better spent developing features specific to the UNC community (and by extension, other college communities).
- When might it be appropriate to use machine learning techniques during this project? The community has a relatively large readership, but typically receives no more than a couple dozen of posts per day. Given this, it might be more appropriate to use rule-based algorithms instead.
- What ethical considerations ought to be made when dealing with user data of real community members and how can we be intentional throughout our development process?

3 Tech Stack

As discussed in the previous section, the precise tech stack has not been determined but should be settled on early in development. I am willing to consider the technologies that my team wants to work with, but have a preference towards Amazon Web Services. I am happy to help get team members up to speed on AWS and potentially earn a AWS certificate if interested.

4 Team Composition

My desired team will contain students from diverse backgrounds - ideally there would be representation from several demographic backgrounds, majors outside of computer science, and experience with Reddit and the specific subreddit this project is centered around.

5 Future Work

Upon successful delivery of this project, I am interested in furthering the development of this application, either as a tool for a research project or a commercial piece of software - and I would be particularly keen on taking on students who are interested on staying with the project after the duration of the course.

6 Project Video

A video highlighting most of the information in this document and putting a face to my name can be found here.