CS 410 Project Proposal

Tyler VanderLey

1.

I am completing this project by myself, which makes me the captain of my team by default. My NetID is tyler14.

2.

My free topic is to investigate the progression of fan sentiment over the course of a college football season according to post-game threads on reddit using sentiment analysis techniques. On reddit, there is a subreddit dedicated to college football discussion (https://www.reddit.com/r/cfb), and after each game during the season, a post-game thread is created for fans to talk about the results. Often, after a big win or exciting game, there will be a large volume of comments posted in the thread in which users enthusiastically recount highlights and discuss implications for the rest of the season. Conversely, for games that were frustrating losses for a team, the comment section will frequently reflect a rather negative outlook. Hence, because of these wide swings in emotion that I've observed from reading these threads on the college football subreddit, I thought they would make a great candidate for performing sentiment analysis. In particular, I plan to collect multiple of these threads for teams over the course of a season so that I can analyze how fan sentiment may change over time as a team either improves or declines.

The task at hand for my project will first be to collect a subset of these comment threads from reddit, which will then serve as the dataset on which sentiment analysis can be performed. In the end application, I currently plan on having a pre-collected set for a small number of teams that a user can select, but I will also allow a user to provide links to their own set of reddit threads for a team of their choosing. If a user provides their own, I will scrape the comments from reddit using PRAW, a Python package that allows for access to the Reddit API. After the set of threads is determined, sentiment analysis will be performed on the comments. I intend to allow a user of the application to choose whether to filter comments based on the commentor's flair (a flair highlights which team the user is a fan of), so that they can either drill down into the sentiment of fans for a specific team, or more broadly assess the sentiment of college football

fans collectively. For now, the exact sentiment analysis methods are yet to be determined. After performing sentiment analysis on each separate thread, the application will report back its assessment for each, and display how sentiment progressed for each consecutive game. For now, I plan to report these results via a simple user interface, likely to be developed in Flask.

In general, the expected outcome for this task is that the sentiment progression report generated by the application should roughly correspond with how well a team does over the course of a season. For example, the Illinois football team (at the time of writing this report) is having an unexpectedly good season, so I would expect the application to report most of Illinois's post-game threads to have a predominantly positive sentiment. Conversely, a team like Oklahoma is having a somewhat of a disappointing season relative to their preseason expectations, so most of their reddit threads should see more negative sentiment. Thus, my main strategy for evaluating the application's effectiveness will be to assess whether the sentiment reports roughly match the actual performance of a team that can be objectively seen via wins and losses.

As mentioned above, the tools I currently plan on using PRAW and Flask as my main tools for accessing the data and displaying results that a user can interpret. For now, I plan to use the Python NLTK library to perform sentiment analysis. All other code will be written in Python as well (besides a bit of additional HTML and CSS that may be needed in the Flask application).

I believe this task is interesting because over the course of a college football season, any individual team can see massive swings in fan opinion, as between weeks fans may go from feeling elated to having strong doubts about their team's future direction. Thus, it seems like a prime example of a domain to analyze from a scientific and algorithmic perspective, as I will be highly curious to see how sentiment analysis may report how fans felt over time about certain teams based on reddit post-game threads. Additionally, I think this topic could have real-world applicability. For instance, this application could be highly useful for a football team's marketing department, as they may want to assess engagement and opinion online to determine how fans are feeling about the team.

3.

As discussed in the previous section, I plan to use the Python programming language. The external libraries I intend to use (PRAW, Flask, NLTK) are also from Python.

4.

I believe this project will indeed take at least 20 hours. To justify this, below is a list of anticipated tasks and a rough estimate of how long each will take:

- Obtain a pre-collected set of post-game threads that a user can analyze: 2 hours
- Enable user-selected threads to be scraped from reddit using PRAW: 5 hours
- Enable users to filter comments based on users' flair/team preference: 2 hours
- Experiment with various sentiment analysis approaches and implement approach in code: 5 hours
- Validate finalized approach using actual game scores: 2 hours
- Configure application to produce report of sentiment analysis progression across all provided reddit threads: 2 hours
- Develop a front-end UI in Flask that can receive user input and report output visually: 4 hours

Thus, I currently estimate that my project will take about 22 hours to complete on my own, which meets the workload requirements.