CS 410 Project Proposal

Members/Captain: Elias Hadgu

NetID: eliash2

Topic: r/soccer scraper and analyzer

Problem: Reddit can be an intimidating website for newer users due to the amount of text in and frequency of posts. Someone who is new to the website or someone who just wants to know information about soccer quickly could benefit from a condensed analysis of the subreddit. This project relates to the theme of the class because it involves processing large amounts of text, doing analysis on it, and returning relevant information to a user.

Tools: I will be using PRAW (Python Reddit API Wrapper) to access reddit posts and comments as well as the BM25 retrieval function to find relevant reddit posts/comments based on a user's query.

Testing: I will know if my implementation works because the information that will be returned to the user (Best player of the month, Fan favorite of the month, Goal of the month, etc..) are things that I would know because I follow the subreddit and the sport closely. If a player that was injured wins player of the month or if a goal that was scored last month wins goal of the month, I will know there was a problem. I could also try making a small test dataset with my own data to see if what I want to happen is actually occurring.

Programming Language: Python for backend and JS for frontend

Workload (Overestimates but I included more that 20 hours if it takes less time than I think):

- Reading Documentation on/learning PRAW - 2hr
- Create UI for user to input and view response - 10hr
    - I am pretty inexperienced in Js so this might take me longer than normal
    - Hopefully the UI will be able to show the video of the goal of the month or a picture of the fan favorite of the month instead of just linking to the reddit post
- Goal of the Day/Month/Year - 1hr
- Fan favorite of the Day/Month/Year (how to define favorite) - 3hr
- Best Player of the Day/Month/Year (how to define best) - 5hr
- Most Popular Team of the Day/Month/Year - 1hr
- Top User of the Day/Month/Year - 1hr
- Clip of the Day/Month/Year (non-goal) - 2hr
- Top Reporter of the Day/Month/Year - 1hr