Tim Powell

CS 410 Project Proposal

Topic: Text mining of recent sports news

1. This project will be completed by myself therefore I will be the captain. My NetId is timp3.
2. My free topic is Text mining of recent sports news. My application will fetch text data from an online source, clean the data, classify the data, and use the classifications to make predictions. The source of the text data will be posts from one of my favorite websites, reddit. The source could be expanded to any website whose API provides easy access to its webpage contents. I'll use common approaches to clean the data such as removing stop words. Once cleaned, I'll further analyze the data in order to classify each post as positive or negative news. With these classifications I'll make predictions about the performance of players/teams in their upcoming games/seasons. This topic is interesting to me because I would like to study the correlation between team/player performance and the public opinion. I would like to see if the general opinion from outside the actual team reflects the performance of the team/players. I'll also look at what team/player is being mentioned most frequently and determine if this factors into the performance prediction. The tools involved include VS Code so that I can write and debug the source code. The data set involved will be the text data that is pulled from reddit. To be more specific, I'll be looking at subreddit forums that are specific to certain sports/leagues/teams/etc. If you are unfamiliar with reddit, it's a large collection of different forums. Each forum has its own topic and members. The expected outcome is an

application that will use a general public opinion regarding a team/player to predict their future performance. I’ll evaluate my work in real time by taking the predictions and comparing them to future performances. I’ll also evaluate the application by using old data sets. This will include old text data as input and old performance results that I can compare to the output of the application.

3. I plan to use python to create this application.
4. In order to complete this project I’ll spend 2 hours determining how to retrieve and format the data from reddit, I’ll spend 4 hours determining how to clean all of the data, I’ll spend 6 hours determining how to classify the data which will include trying different approaches, I’ll spend 4 hours testing/debugging the application with an old data set, and I’ll spend the final 4 hours optimizing the application and expanding it to accept other text data sources.