CS 410 Project Proposal Group 268

1. NetID's: Timmy Sullivan (tps7) – Captain

2. My free topic is I am building a text analyzer for NFL (football) play by play data. The problem this will solve is basic NFL statistical data for players can be scraped online directly, however, more advanced statistical data isn't readily available. What is readily available is the NFL play by play data. The play by play data contains all the information needed to do advanced analytics, but the specific information cannot be easily pulled from NFL play by play data. This is where my project would come in as I would be able to quickly parse the play by play data and pull out the important elements for necessary analysis. This task is important/interesting because it allows me to do advanced analytics on NFL data.
Example play by play data: (14:52 - 1st) (Shotgun) R.White up the middle to TB 33 for 6 yards (T.Bernard)
All of the information in the play by play data can be used to do advanced analysis however, it can be hard identifying important parts of each script of play by play data. Also, there are many edge cases as well for example on plays where penalties occur most of the time the play doesn't count, but the penalty could be declined which means the play would count.

Approach: NFL play by play data can be found online and can be downloaded. Once I have the data I will need to come up with a way to accurately pull important information from the play by play data. Almost as important as getting the data, I have to come up with a way to verify that the data gained from the text is correct. Lastly, I would have to use the data that I pulled out of the play by play data to compute advanced statistics.
Expected: My expected outcome is that I will be able to provide the answer to questions like "Which player on team X gained the most first downs in a game/period of weeks/whole season", or "Which receiver was targeted the most when there was less than two minutes left in the half on team x in a game/period of weeks/season."
Evaluation: I should be able to test my data by comparing the parsed data to some easily obtained data. For example, I can compare yards gained for players calculated from my data and the actual recorded yards gained for a player. I can also try to find some advanced statistics online and compare my data to those as well to ensure correctness.
Tools, Systems, and Datasets: Right now the only thing I am using is a dataset with NFL play by play data for an entire season. I may need more tools/systems as I work on the project.

3. I plan to do all of this in Python

4. I think this project will easily take me 20 hours. My goal was to come up with a project that is small enough that it can be completed by one person (only one person on team), but can easily be expanded on. Right now I am not sure how long each step of the project will take, but if I do everything goes shorter than expected I can expand on the project by doing things like scrapping the data myself or computing more advanced statistics. For reference, there are an average of 153 plays per game, which means there would be over 40,000 plays in a season. I plan to do this for at least one season's worth of data. A rough outline of the main tasks and estimated time is below:

- Obtain data (1 hour)

- Write text parser (~7 hours)
- Write tests for the parser and ensure accuracy (~12 hours)
- Calculate advanced stats (2-X hours depending on the number of calculations)
- Do write ups and reports for the project (~5 hours)