1. *What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.*

- Dixon Liang
    - dixonl2@illinois.edu

2. *What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?*

My free topic is to create a sentiment analysis of soccer games from the English Premier League using Twitter. Specifically, I am interested in discovering which players from either team had a good or poor game based on sentiment of tweets. This is important or interesting because based on this analysis, we can use it to come up with a detailed "form" analysis to see which players have been playing well over an extended period. Although specific individuals from Twitter might not be the best pundits of games, I will be using a "wisdom of the crowd" type approach on the quality of the data gathered.

The main task I will be doing is taking tweets from a sample game in the past and categorizing the words in tweets related to certain players for a positive or negative sentiment. After this categorization, I will be able to aggregate and determine which players had a good or poor game.

My planned approach is to use a particular game in the past few weeks as a demo. I would text mine all the tweets related to the game using a filter of the time period around / during the game and then those using a hashtag related to the game. I would then further analyze the tweets that have mentions of specific players and the words in context. Based on the context of the tweets, I would categorize the tweets related to players as either "positive', "neutral", or "negative". Totaling the sentiments for each player during the game should give me a classification for each individual player determining their performance. I will likely to be able to further quantify based on some measure on how many "positive" or "negative" tweets each player has been categorized.

The main tool I will be using is "Tweepy" which is a Python package to read tweets from the Twitter API. The main dataset will be the tweets from the time period around the games that I have chosen, and that I have categorized as relevant.

If I have the time, I would also like to incorporate one of the functions from the course into my project. I will have a better idea through the planning process, but as of now, I would likely treat each tweet as a "document". The most likely adaptation will be to create a likelihood model using the game tweets as the primary data set. An interesting application would be to try to categorize tweets relating to players to specific parts of the game which would be the "topics".

My expected outcome is to produce a report detailing the findings from one or several games. I should have enough data per game to show all the players' performances who were involved in the game. In this report, I will show which players were categorized as having good games or poor games based on the categorizing of tweets. In a further breakdown, by using a likelihood model, I will be also be able to show the topics where a player performed well or poor. As an example, a positive tweet might be relating to a specific player's passes during the game.

I will evaluate my work based on reviewing some "match ratings" by pundit type publications to see if my reviews based on sentiment are in line. Although I anticipate some differences, if this works successfully, there should not be big differences in the way each player performance is viewed against the experts. I will also try to watch the games that I use to review myself if my ratings make basic sense.

3. *Which programming language do you plan to use?*

- Python

4. *Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.*

I will be working alone on this project. I anticipate the breakdown of time spent on the project as follows:

i. Initial Research and Outline (2-4 hours)
ii. Familiarity of Tweepy and Other Tools (2-4 hours)
iii. Text Retrieval and Data Cleaning (2-4 hours)
iv. Initial Implementation of Algorithm (10-20 hours)
v. Testing and Improvements (10-20 hours)
vi. Final Reports, Documentation, and Demo (5-10 hours)