Final Project Proposal

DS 710

For your final project, you will collect and analyze data from Twitter to answer a question that interests you.

1. What is your question? You might be interested in a brand, an organization, a meme, a trend, a news story, or a place. What do you want to learn about? Do you have predictions or hypotheses about what you will find?
   1. Did the College Football Playoff (CFP) committee accurately select the contenders of the 2017 CFP? Also, can the accuracy of the top eight (8) teams rankings be identified based on the tweets, and their sentiment, received during and after Selection Sunday (12/4/15), the day that the four (4) football teams were announced?
   2. I believe that with sentiment analysis plus some other text analytic practices, it may be possible to accurately identify “the crowds” displeasure/pleasure of the teams that were selected, or not, after they are announced on live television.
2. What is your audience? Who is interested in the answer to your question?
   1. The audience is broad for this model, from the casual college football fan, to the analysts and reporters from sports media publications (ex. ESPN), all the way up to the commissioner of the NCAA. This audience can use this model, if accurate, to validate and verify that the system in place to select the teams (a committee comprised of 12 members) is a good one or needs to be improved upon.
3. What information will you gather? What features will the tweets you collect have?
   1. The data will come from an api exposed by twitter that will return tweets and other meta-data related to them. Within this data the most important to this model will be the actual text of the data as sentiment can be deduced by it. To narrow the amount of data needed to be collected I will focus on the top eight (8) teams (ranked by the CFP committee) as of the week before Selection Sunday. To do so I will target the twitter handles and their hashtags (if applicable) of all eight (8) teams as well as tweets directly related to the CFP selections signified by the handle @CFBPlayoff and the hashtag #CFBPlayoff. The latter will then be attributed to one of the 8 teams (if one is possible to identified) to be included in their pool of data.
4. How will you collect the data? Will you use the REST or Streaming APIs, or a combination of both?
   1. Again to collect the data I will use the API exposed by twitter. As the selection has already concluded the following few days will be included within the data, so for this purpose I will use the REST API.
5. How will you analyze the data? What features are you looking for? Are there specific hypotheses you plan to test?
   1. I will first attempt to identify the teams that the tweets returned from the @/#CFBPlayoff data pull are attributed to. To do this I will likely use a bag of words technique and use the tweets specific to the teams handles as a baseline to try and identify on. This process will likely leave me with a few tweets that are not identifiable to any one team and will be excluded from the further analysis. Then the negative, positive, and overall sentiment of the tweets will be identified on individual scales and will be used to identify the agreeableness, disagreeableness and overall perception of each teams ranking. The null hypothesis of this model is as follows – The selection committee accurately selected the correct teams to be a part of the College Football Playoffs based on the sentiment of tweets respective of each team. The alternative would be – The selection committee did not accurately select the correct teams based on the sentiment of tweets respective of each team.