Data 620 Final Project Proposal

Summer 2019

Due: 7/14/19

Group 6: Alice Friedman, Stephen Jones, Jeff Littlejohn, Jun Pan

Our Project:

What makes a social media post popular? What elements cause a Tweet to be liked or shared via retweet? Using a corpus of tweets from users loosely connected to a political candidate, we hope to find out. We will start with text analysis like word frequency distributions and sentiment analysis. What are the sentiments of the most liked tweets? Which words occur at irregular frequencies in heavily retweeted tweets?

Next, we will use a portion of our tweet corpus to train models using various algorithms from Python's scikit-learn to predict likes and retweets. We will then use the remainder of the tweets to test the effectiveness of the models.

Time and data permitting, we hope to look at differences in the tweets of verified and non-verified users in attempt to determine how professional pundits or other prominent users may use social media tolls differently than amateur users. This should be in our text analysis portion. Additionally, we hope to use the tweet corpus to build a bot that writes its own tweets that aim for likes and retweets.

For Project 1, Group 6 dove deep into getting data from the Twitter API. We performed social network analysis looking at the followers (and for a sample of followers, the followers of followers) of the Democratic presidential candidate Mike Gravel. In performing this exercise, we learned that working with Twitter's API can be difficult both in the amount of data you're allowed to extract and the time which it takes to do so.

Data Source:

For the Final Project, we are sticking our initial approach of pulling data via the Twitter API from followers of Mike Gravel that we identified in that first project. However, we've updated our code to pull the text of the tweets and metadata such as number of retweets and likes. As of the time of the proposal, we appear to be limited to only pulling the last tweet from a given user at the time of our accessing the API, which is a significant limitation.

We're aware that our sample of tweets is specific to people following a fringe primary candidate and not representative of social media overall or even just "political Twitter" itself. Still, an older political candidate curious about the sentiments and topics prevalent among younger, politically engaged users on social media could find an expanded version of our exercise useful in defining a strategy for voter engagement.

Group Breakdown:

We will leverage code developed during Project 1, which was principally developed by Stephen Jones and Alice Friedman. Jun Pan augmented the code to pull actual tweets from the Twitter API and will be specializing in data visualization for the final project. Stephen Jones will be primarily responsible for text analytics, including word frequency and sentiment analysis. Alice Friedman will be creating the model(s) that predict the number of likes and retweets based on a tweet's contents. Jeff Littlejohn will be working on developing a bot that creates its own political tweets, compiling the final work, and making the video.