Cayden Dunn
Data 510 hw2

Proposal:
Twitter text sentiment analysis. The question I am trying to answer is “Does twitter amplify outrage or misinformation more so than other tweets”. We have all seen headlines like this ‘Tweets containing falsehoods were 70% more likely to be retweeted than truthful tweets’. I want to see if I can actually show that this is true through web scraping and sentiment analysis. I wanted to do some work with machine learning for this project but I needed to find a way to include gathering, cleaning and viz’ing some data. I think this idea of a sentiment analysis model fits the bill. While I have never sought out an NLP project I dont think this experience will at all hurt me and could ignite a new interest.

The Data:
I plan to scrape popular yet polarizing twitter pages as well as pages that are factual and shouldn't be posting outrageous content as a control(ex. Trump, aoc, elon_jet_tracker, weather_tracker). I will need to control for the metric of how many followers a particular person has and find a way to normalize this data so it does not skew my model. I plan to use libs like nltk, numpy, pandas, twitterscraper(or some lib that is built for scraping twitter), re, spacy, tensor, and perhaps some other ML NLP libs as I have never done any NLP before.

Data Cleaning:
As far as raw data cleaning is concerned I plan to scrub and tokenize all the tweets I gather to get them in a uniform format. I also think I will have to do some feature normalization on follower counts and retweet counts so I can compare accounts of different sizes. I anticipate having to create some subscripts that will rip the follower count from each profile of the tweet I am scraping.