Project Deliverable 1

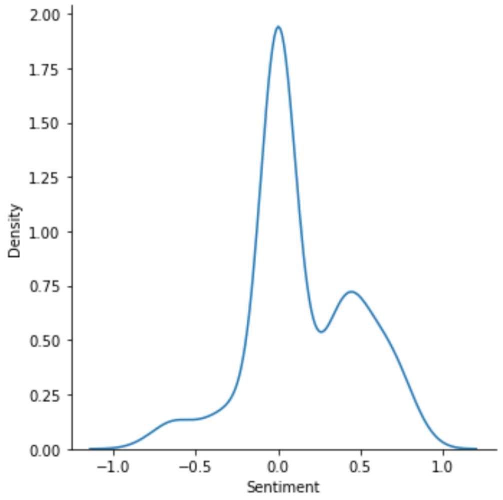
Group Members: Conor Walsh, Yash Bengali, Patrick Kuzdzal, Carlos Lopez

---

Collect and pre-process a preliminary batch of data:
We tried parsing tweets using the twitter api but we ran into issues with the limit of how many tweets we could parse at once. We were able to get about 28k tweets. We ended up finding a link to a kaggle competition which contained tweets relating to stocks about the top companies including Apple, Microsoft, Tesla, ect. This data is only tweets that contain the stock ticker symbol $APPL for example instead of #APPL. We filtered the tweets to only contain tweets that were Apple related and removed the tweets that were from other companies. We in total now have 1,425,013 apple tweets that we got from the kaggle dataset.

As a side note, we have since fixed the issue that occurred when collecting tweets using the twitter api, and are now currently gathering more. Our projected total is around 300,000.

Preliminary Analysis:
For our preliminary analysis we wanted to look at the sentiment of the data for the first day and the graph below shows the density. Much of the data was rated at 0 or near 0 which makes sense because a lot of the tweets were written in a neutral way. However, there is still enough data for each day such that we can get the mean of the day and do analysis with that.



Key Question - Can we find a correlation between our current data and the stock price for 1 day.

We ran the SentimentIntensityAnalyzer for NLTK on the jan 1st 2015 days and got an average sentiment of 0.13. The stock price on this day went from 27.85 to 29.29. Therefore, for a sample size of 1, it does appear like a correlation may be possible. This is only one day's data of course and we will work on analyzing more days in the future as well as using better statistical methods such as r^2, but as a proof of concept for n=1, this project seems pretty promising and scalable.

Refined Scope/Limitations:

Our project scope is still the same as before which is to see if there is a correlation between the sentiment of apple tweets and the stock price. For this, we will be looking at tweets and stock prices from January 2015 up until now.

Modifications to Original Proposal:

The only modification made was that tweets were scraped using the keyword '$AAPL,' rather than any other hashtags. This ensures that we get only stock relevant tweets and information, and not, for example, a person tweeting about the actual fruit.