Twitter Market

By: Ben Walczak

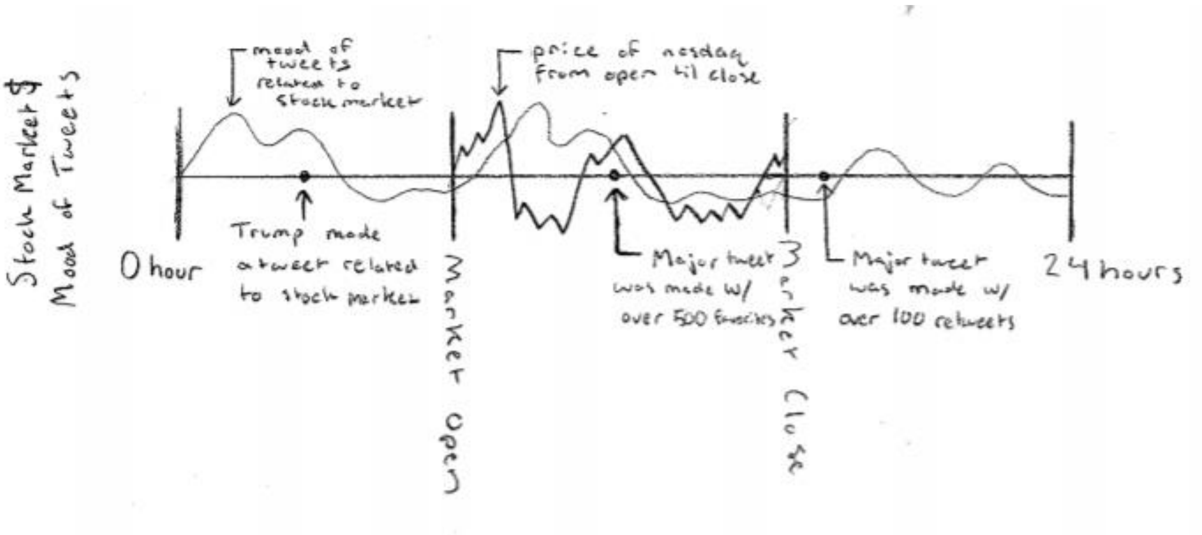
https://github.com/ben-walczak/TwitterMarket

Abstract:

If you have ever taken an intro to economics class, you would probably know that the expectation of the buyers and sellers in a market can greatly affect the market price. Apply this to the stock market; there are many stock traders, constantly buying and selling. Perhaps a trader could have seen a bit of news on twitter that may have influenced them to think differently about a stock and caused them to trade that stock away or buy more stocks. This is the expectation of the buyer or seller; the thought that a stock could potentially rise or fall in the future. This phenomenon is known as behavioral economics. We can take this idea and apply it to twitter. Buyers, sellers, news outlets, and other influential twitter accounts will often express how they feel about the current state of the market over twitter. Inevitably, these tweets will reach others and affect how those potential buyers and sellers feel about the market as well. Using this knowledge, we can track many of these tweets related to the stock market. After analyzing each tweet using sentiment analysis, we can judge the polarity of how negative or positive a tweet is to determine the current mood of the market, hopefully using that to predict the current stock price.

Introduction:

The goal of this project is to compare the mood of recent tweets related to the market compared to the stock market prices throughout the day. The primary language that will be used will be R, but that may change in the future to incorporate more machine learning techniques. The plan is to first retrieve tweets related to the stock market and the stock prices of nasdaq throughout a given day. Then, additional work will be done on tweets to clean them and apply sentiment analysis to determine an algorithm to judge the polarity of each tweet. A time series graph will be designed to display and compare the mood of tweets and the stock price throughout a given day. The final model should look similar to the following:



- Additional curves will be added to the final model. These curves include the mood of favorited tweets and the mood of retweeted tweets.
- Polarity of each tweet will be determined based on the individual tweet, which will affect the overall mood.
- Influential tweets with many favorites or retweets will be plotted on the time series to show how these tweets may affect the stock market price.

Research questions related to the final model and project:

- How does the mood of tweets throughout a day relate to the stock market price?
- How do tweets with favorites or retweets relate to the stock market price?
- How does the mood of tweets relate to the highs and the lows of the stock market price throughout the day?
- How do influential figures or tweets with many favorites or retweets relate to the stock market price?

One data set will be the nasdaq market price plotted throughout the day. Important metrics of this data set include the opening price, closing price, the low of the day, and the high of the day. The only metrics needed though, will be stock price and its relative time. All of the previously mentioned metrics can be retrieved using the two given metrics.

The second data set will be tweets throughout the day. Each tweet contains 16 metrics, but we only need a few. The few metrics we do need are important to the final model and the research questions. These metrics include text, time of tweet, whether it was favorited, favorite count, whether it was retweeted, and retweet count.

Retrieval of data:

The tweets were retrieved through a few steps. To be able to search for tweets, a twitter account was first created. After setting up the twitter account (@NotARobot1010), the twitter account was authorized to be used by a developer (@https://apps.twitter.com/). Then, proper credentials were entered in R, so the computer could access the twitter account. Finally, searchTwitter() method, part of the OAuth R package, was used passing the key word “nasdaq”, the maximum number of tweets to be retrieved, the language, as well as the range of dates the tweets were to be retrieved from. It is important to have OAuth and twitter R packages to be loaded in order for the commands to work.

Data cleaning/wrangling/processing:

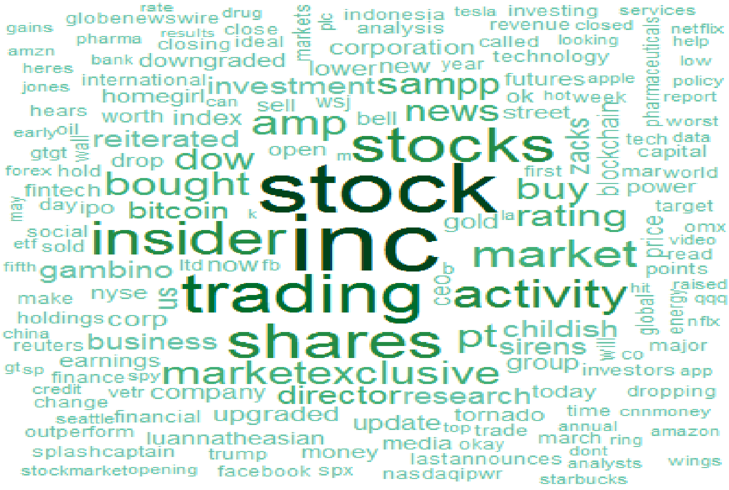
After tweets were initially retrieved, they were organized into a data frame for easier manipulation. Column names could be checked at this point to check metrics of the tweets here. The data frame is then reduced from 16 metrics down to 6 metrics, which were "text", "favorited", "favoriteCount", "created", "retweetCount", and "retweeted". This data frame will be used in the future for the time series graph but to analyze text it must be manipulated further. The text of the data frame will then be converted to a single metric of text. The text will have all URL’s removed in addition to punctuation and stopwords. Many stopwords were considered, such as articles, conjunctions, etc. Additional stopwords were eventually added after exploratory analysis. These additional stopwords include “via”, “rt”, and “like”. The text was then converted into a term document matrix, which allows for future manipulation for exploratory analysis.

Exploratory analysis:

For both forms of exploratory analysis, the word “nasdaq” was removed from the text, so visualization of frequent terms was not skewed by that term.

Word Cloud:

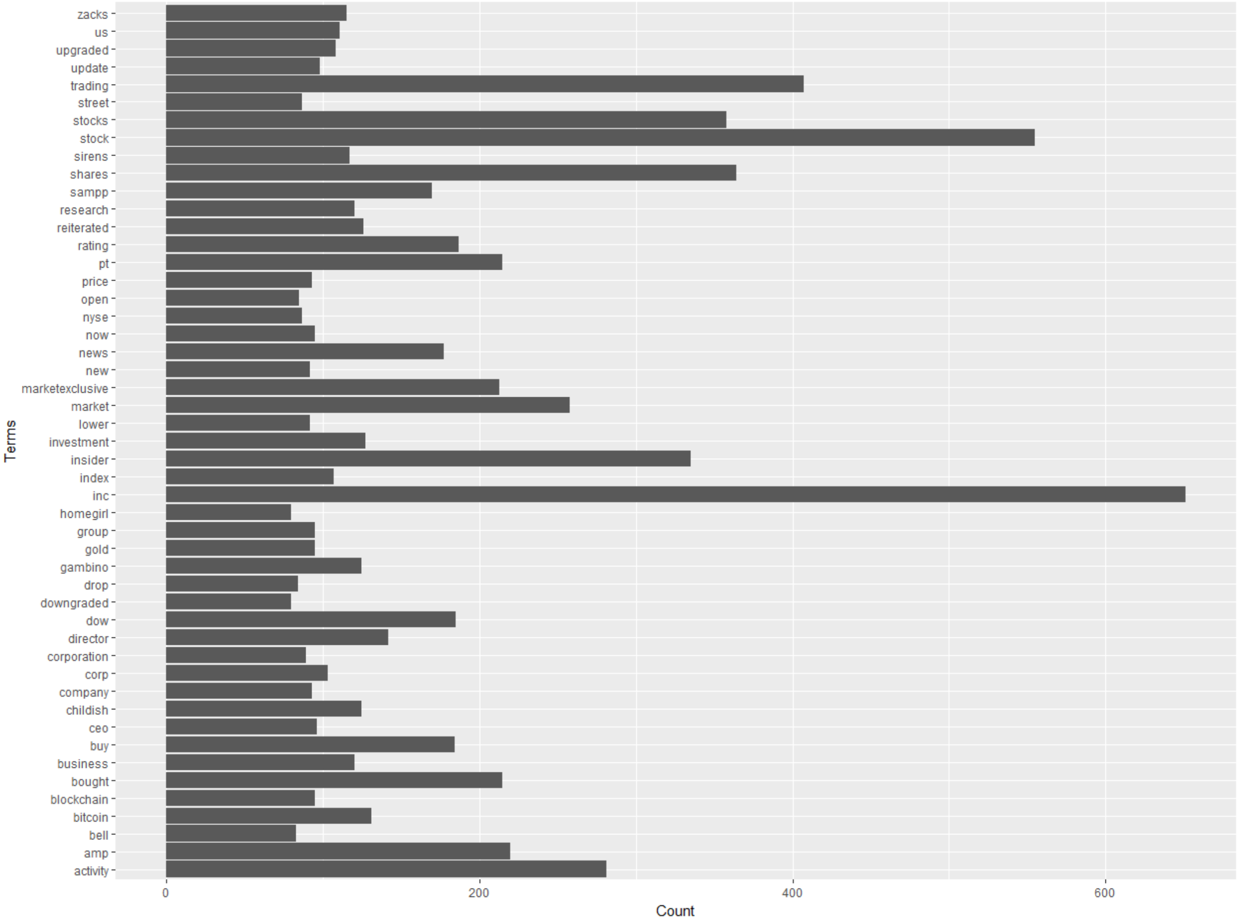
Infrequent terms were removed from the term document matrix to prevent the word cloud from being too large. Next, the term document matrix was converted to another matrix. Then, that matrix was used to calculate the word frequency. Finally, the word cloud was set up.



- Important stocks and stock collections mentioned: s&p500, dow jones, nyse, gold
- Influential companies mentioned: Tesla, Netflix, and Facebook
- Significant figures mentioned: Trump, ceo
- Important currencies mentioned: bitcoin

Bar graph of frequent terms:

The term document matrix was retrieved from the text a second time, so all terms would exist within the matrix. Sparse terms were then removed from the matrix, which was a different amount than before. Term frequency was calculated using the term document matrix, and then reduced to only terms that were greater than or equal to 15. A data frame was set up for term frequency to be passed through ggplot. Then a bar graph was created displaying the frequency of the most common words.



- Grouping important terms from before and displaying them on a bar graph may also yield interesting results