Market Prediction Based on Daily News from The New York Times

Capstone Project 2:
Milestone Report 1

August 4, 2019

David Gray

The goal of this project is to train a machine learning model to predict the movement of the stock market based on the news for that day. To motivate this project, a parallel might be drawn between the news that is presented to the public (what the newspaper anticipates the public wants to learn about) and the volume of specific search terms (what the public wants to learn about). It has been stated that the search frequency of terms is correlated with stock movement. This is from the website, Seeking Alpha:

- Past research suggests that the relative change in the volume of Google searches for financial terms such as “debt” or “stocks” can be used to anticipate stock market trends.
- In this analysis, the search term “debt” was used to obtain monthly search volume data from Google Trends.
- The analysis shows that a decrease in search volume typically preceded price increases in the S&P 500 index, and vice versa.
- Switching between ETF (SPY) and ETF (IEF) based on monthly search volume data from 2005 to 2018 would have made a profit of 634% versus 220% for buy-and-hold SPY.

This is some motivation to try to predict the stock market with information deemed important. Reasons to study this question of connecting the news with stock market movement is both to learn the tools of NLP and machine learning model generation, and

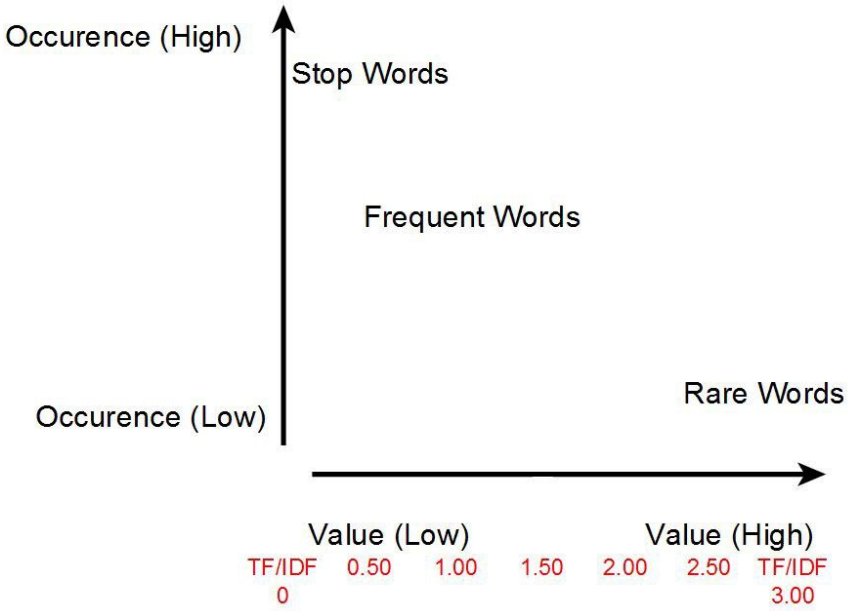
also to see whether financial gain could be made with a model generated such as in this work.

The news collected for training and testing consisted of ten abstracts from national news articles and ten abstracts of international news. These abstracts were joined for analysis. The beginning to end dates were July 20, 2009 to July 19, 2019. Only those news articles were collected on days where the market was open, and the stock market news was accessed from Yahoo Finance. In summary, the approach of this project is to gather daily news from ten years of New York Times' abstracts and connect that with the movement of the S&P 500 (binary up or down movement) in order to predict the stock market on a daily basis. Importantly, the model predicted movement for the day, whether it increased or decreased and not the magnitude of change.

This data set was collected directly from The New York Times using an API that the news organization provides. There are access limits for its use, so a delay was set between every API call of twenty-three seconds to abide by the daily limits.

Standard NLP techniques are employed, which include tokenizing the collection of twenty total articles, removing the stop words, and lemmatizing the remaining words. The resulting data set consists of approximately 40,000 columns (unique words) and 2,500 rows (days of news). A pandas DataFrame is used to store the information of how many counts per word exist for each day of news. So far, both count vectorization

and TF-TDF approaches were used. The count vectorization approach has resulted in a slightly higher accuracy than TF-TDF. Here is a graph summary of using TF-IDF:



From: Towards Data Science

Future plans include:

- Use bigrams in tokenizing words
- Train a machine learning model using approach other than Random Forest, such as Naive Bayes and Support Vector Machine (SVM)
- Collect the daily news into groups of five (five workdays per week) to train based on a longer time scale.

One thing to note is that although the accuracy could be low for predicting whether the stock market increases or decreases on a given day, money could still be made if it predicted increases or decreases when there are higher magnitude changes than other

days. That is, buying stocks at the start of the day would yield higher gains since it rises higher.