README

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.

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. In addition, weekly changes were considered as well as creating bigrams with three models total.

The accuracy of all the models (weekly or daily, unigram or bigram, SVM or random forest or Naive Bayes) approached the accuracy of guessing for every period that the stock market increased. While the model did not predict with high accuracy the movement of the stock market, there were lessons learned in the process of preparing the data and developing the model. There are at least several aspects to consider in potential next steps

- Remove counts of numeric values, such as "2,000" but keep years. This could potentially be creating noise in the model.
- Consider processing a different section of the news for analysis, such as business and the editorial page if possible. To repeat with national or international news, gather top news instead of random stories for processing.
- Prepare all of the code necessary for analysis of the data after it has been downloaded and processed. This is to prevent any delays of trying to load the data at a later time.
- Perform grid search on all the models to find the best parameters with a greater range and number.
- Predict magnitude rather than binary change

Contents of this project folder include:

- Final report
- Final slides
- Jupyter notebook with code