# Stock Market Movement Prediction

**Byungsu Jung**
Department of Mathematics
bjung@uw.edu

**Seoyoung Park**
Department of ACMS(Data Sci&Stat)
syp1017@uw.edu

## 1 Summary of research questions.

**Question Summary**
In a brief sentence, our research is trying several NLP techniques to predict the stock market and get words that perform as important features.

The aim of our study is to roughly predict stock market price movement based on top 25 News headlines for each day. Considering this task involves NLP, we will implemented several NLP feature extraction techniques such as N-gram, TF-IDF and others. For each of feature extraction techniques, we will implement different machine learning models to evaluate which feature extraction showed highest prediction. After successfully identifying best feature extraction technique and machine learning model, we will further provide report on "Why selected feature extraction technique and machine learning model showed best performance", and we will propose future improvements,

## 2 Motivation and Background

The core context of our research questions is “Does new headlines have a measurable significant effect on movement of Stock Price?”. Within scope of this question, we are especially interested in verifying if “News Headlines” does have impact on the movement of stock price, what machine learning model shows best prediction. Also, we are interest in finding out the word or words that have most effect on such analysis.

With the result of this research through computing, we can verify the effect of news on stock price. With an answer to this question, we can understand at least few aspect that could effect the stock price and we can use this information to continuously look for other aspects that could affect stock price. Furthermore, we can possibly build a model that could be used in fields such as financial engineering and so on.

## 3 Data set

We will use dataset from the **Kaggle**. This dataset consist of three different CSV files with each containing different type of data. The following is URL to our dataset.

`https://www.kaggle.com/aaron7sun/stocknews#DJIA_table.csv`

**News data**: Crawled historical news headlines from Reddit WorldNews Channel (`https://www.reddit.com/r/worldnews?hl`), which are ranked by Reddit users’ votes. Only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)

**Stock data**: The Dow Jones Industrial Average, or simply the Dow, is a stock market index that indicates the value of 30 large, publicly owned companies based in the United States, and how they have traded in the stock market during various periods of time. (Range: 2008-08-08 to

2016-07-01)

- **RedditNews.csv**: The first column is the "date", and the second column is the "news headlines".

- **DJIA table.csv**: Gained from Yahoo
  - "Date": YYYY-MM-DD format
  - "Open": Opening weighted average stock value in USD
  - "High": All day high in USD
  - "Low": All day low in USD
  - "Close": Closing weighted average stock value in USD
  - "Volume": Number of trades
  - "Adj Close": Adjusted closing prices - adjusted for both dividends and splits - in USD

- **Combined News DJIA.csv**: The first column is "Date", the second is "Label", and the others are news headlines("Top 1" to "Top 25"). "Label" indicates whether the DJIA Adj Close value rose or not - Binary classification. If it rose or stayed the same, it's 1, and if it decreased, it's 0.

## 4 Methodology

In a brief sentence, our goal is to predict stock price movement base on news headlines. The following is the steps that we are going to perform to get to the result. These following steps include algorithms, analysis and metric we will be using to predict stock price movement.

1. As mentioned, the dataset contains daily top 25 new headlines in a form of separated string. In order to use all these headlines to be used to predict single day stock movement, we are going to combine them and clean the data. Cleaning data includes steps like making all the character to lower cases and removing all the special characters.
2. With cleaned dataset, we are going to vectorize the string by using both bi-gram and Tf-Idf. We chose to use bi-gram to avoid ignoring effect of sequence of words.
3. With Tf-Idf setup to be used for analysis, we are going to use multiple classifiers such as Gradient-boosting, Ada-boost, k-NN, Stochastic Gradient Descent Classifier and more. We will also evaluate each model to find out which classifier best classify up and down movement of stock price.
4. Furthermore, we will use the highest predicting classifier to perform hyper-parameter tuning to possibly determine best hyper parameter and see why this hyper-parameter shows best result.

In a conclusion, we will propose further research topic to consider based on our conclusion.

## 5 Work Plan

- Import data (< 1 hour)
- Feature extraction (1-2 hour)
- Perform several NLP techniques and get accuracy score (4-6 hour)
- Make some plots to compare different techniques' accuracy scores (1 hour)

We are going to meet on Tuesday and Thursday after 5:30pm to see what we have done so far, discuss what to do for the next step, and work together. Other than those days, we are going to divide responsibilities. Also, we are going to use Git to manage changes.