Omar Sadek

MAIS 202: Deliverable 1

i- The dataset I chose to use for my project is from Kaggle and it contains the top 25 daily news headlines from 2008-06-08 to 2016-07-01. The dataset also has a binary column that indicates how the Dow Jones Industrial Average Index performed on that specific day, where it is assigned a value of 0 if its value decreased and a value of 1 if its value rose or stayed the same.

The link to the Kaggle database:
https://www.kaggle.com/aaron7sun/stocknews#Combined_News_DJIA.csv

The link to the Yahoo Finance database:
https://finance.yahoo.com/quote/%5EGSPC/history?period1=1218067200&period2=1467331200&interval=1d&filter=history&frequency=1d

ii- I think this dataset is good because it has already scraped Reddit's WorldNews subreddit and provides the top 25 news headlines of any given day within a specified range in a clean and concise manner. This will save me a lot of time in scraping this data myself.

iii- If I am to use the binary column indicating the Dow's daily performance, this would make this a classification task; however, I will prefer to have this be a regression task, where I could examine the magnitude of the effect of daily news headlines on the US stock market instead. I would then use Yahoo Finance to get the daily prices of the S&P 500 index (I prefer this index to the Dow Jones for many reasons) for the same time period, and then calculate daily returns as a percentage using that dataset and merge that new vector with the existing news headlines dataset, replacing the binary Dow Jones outcome column.

iv- I plan on using this dataset to create a model that will be able to predict how the US stock market performs on a specific day given that day's major news headlines. It is common for the stock market to reflect sentiment and the wording used in news plays a big part in how the markets interpret daily events. Therefore, I believe using an NLP algorithm will be crucial in this project along with some form of regression model.

v- For presenting this project, I plan on building a simple webapp with a single input field where users can copy and paste a news headline and the model would then output a predicted percentage change in the S&P 500 index for that day. I prefer this method over a poster presentation as it would make it more interactive and actually useful, if someone decides to implement it.