**CptS 475/575: Data Science Fall 2020**

**Project Proposal**

*Ayush Kapoor- 11638482*

*Sannidhya Malpani - 11638049*

*9/1/2020*

For our Data Science project, our team has chosen to analyze three interesting datasets retrieved from Kaggle (reference is given below). These datasets are in csv format and are explained in brief below:

* News data: This dataset includes crawled historical news headlines from Reddit WorldNews Channel, (ranked by reddit users’ votes) with a range from 2008-06-08 to 2016-07-01. In this dataset, there are two columns (“dates” and “news headlines”). All news is ranked from top to bottom based on how hot they are. Therefore, there are 25 6 lines for each date.
* Stock data: This dataset is the stock data of Dow Jones Industrial Average (DJIA) downloaded directly from Yahoo Finance
* Combined Dataset: A combined dataset is also provided which contains the first two datasets presented in three columns. (“Date”, “Label” and “news headlines ranging from Top1 to Top25”)

Our project will involve using 8 years of daily news headlines from the datasets provided above, in order to predict stock market movement. We will be taking the data from 2008-08-08 to 2014-12-31 as Training Set. Our testing dataset will be the following two years of data (2016-2018).

References

[1] Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [Date You Retrieved This Data] from https://www.kaggle.com/aaron7sun/stocknews.