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

**Final Report**

*Ayush Kapoor- 11638482*

*Sannidhya Malpani - 11638049*

*9/1/2020*

[Pandas](http://pandas.pydata.org/) will make our data easy to look at and work with.  
[CountVectorizer](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html), part of scikit-learn, will take care of our NLP tasks.  
[LogisticRegression](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html), also part of scikit-learn, will train and test our predictive models.

Many variables, but the layout is pretty straight-forward

Label value of DJIA will be 1 if stock stayed the same or rose , and 0 if it fell that day.

As mentioned previously, scikit-learn is going to take care of all of our preprocessing needs.  
The tool we'll be using is CountVectorizer, which takes a single list of strings as input, and produces word counts for each one

As mentioned previously, scikit-learn is going to take care of all of our preprocessing needs.  
The tool we'll be using is CountVectorizer, which takes a single list of strings as input, and produces word counts for each one

• analyze an interesting dataset and build a predictive model for it

It looks that the news in Reddit are predominantly negative.