Capstone Project 2 Proposal

Market prediction is an interesting topic to individuals and investment companies, and there are dozens of related competitions and datasets on Kaggle. To practice NLP, I'm going to use the past ten years' daily news headlines from Reddit to predict Dow Jones Industrial Average (DJIA) up and down through binary classification.

This capstone project contains three parts of data science studies: natural language processing, feature engineering, and training and prediction. The preprocess of NLP will include lower case, removing stop words, token, stemming and lemmatization; and after preprocess, sentiment analysis will be applied to each news headline. Feature engineering section will select the best features, including news feature and time series features, to train the machine learning model and predict market.

The public and free dataset used in this study is here:
https://www.kaggle.com/aaron7sun/stocknews

The preliminary plan of this project:

1. Download the past ten years' news headlines on Reddit and Dow Jones Industrial Average (DJIA) market datasets from Kaggle.
2. Data cleaning: filter out the news during weekends when market is closed.
3. Run NLP on the news: preprocess including lower case, removing stop words, token, stemming and lemmatization, and sentiment analyze after preprocessing.
4. Select the best features to train the model and predict.
5. Test features on different machine learning models.