FRAMEWORK

Brief discussion or interpretation of overall research process and model performance
Highlight significant findings or models.

1. web_Scaping_10K.ipynb (On Progress)

   For web scraping, there are two kind of open data source we can get, either twits or 10-K(Q). The board idea here is to download 10-K by clk and use tfidf/bert to get sentiment/features from 10-K. Then use stock yearly/quartly returns as label to do supervised learning. The main challenge is how to transfer 10-K into useful features to predict stock price(y). This require some amount of research and I'm still in progress, stay tuned.

2. Sentiment Analysis (Version 1)

   Since the EVENT_SENTIMENT_SCORE is given, thus the first idea to train headline with sentiment scores as labels seems trivial since we don't have newly generated headlines. The second idea is to do word embedding by Word2Vec or BERT to change the headlines into features and then use stock monthly return as labels to train a supervised learning algorithm. Because I hear of BERT the first time and I need some time to learn, so I leave this to implement.

3. Prediction Models (Version 1)

   My methodology is composed of 4 steps.
   (1) I clean the data by using Robust Scaler standardization and filling NA with median.
   (2) I introduce the features, (for news sentiment, I use aggregated mean group by RP_ENTITY_ID and date) whereas we set up our two targets: 1. monthly_return (binary) 2. monthly_return_F1 (binary).
   (3) I define the setup of the three machine learning methods we employ, namely random forest, XGBoost and GRU.
   (4) Finally, evaluate by accuracy score and confusion matrix.

   Detail summary about models can be referred to model.pdf.