Big Messy Data Project Proposal

Tongcheng Li (TL486),Yinglei Wang (YW287)

September 23, 2016

1 Dataset Selection

We will use daily news for stock selection data on Kaggle datasets (https://www.kaggle.com/aaron7sun/stocknews).

This dataset consists of two parts:

1. Dow Jones Index.
2. Top 25 daily news headlines in Reddit.

2 Potential Questions to Investigate

1. First, we want to look into if there exist the correlation of how "big" the news are and the volume on the Dow Jones.
   Intuitively, this should make sense because some unexpected important news will lead the market into horror, which will lead to the increase of volume.
2. Second, we want to see if there is any correlation with company specific news related to companies in Dow Jones.
   To achieve the second objective, if the Reddit news is not enough, we will look into alternative data sources for news.
3. Third, we will see what other features (volumes, volatility etc.) of other index or assets are predictive of news (especially the sector specific ones.) This step might involve more gathering of alternative financial data (which can be big and messy).