Title: Analysis of Stock Based on Text Mining

Team name: 7 Geeks

Team members: Dailin Shen, Yuhan Sun, Qinling Li, Guanzhong You, Minghao Dai, Weihan Li, Weikang Zhu

Data source: WSJ corpus which contains 4 months worth of Dow Jones Newswires data. The content is comprised of raw text data with annotated entities, which all comes packaged in XML format.

Description of our project:

Our group focus on the hidden information extracted from among 24169 stories in 6 months. By analyzing publications’ performances and the entities of each story, we find out the entities with highest frequency are Apple Inc, Google, Twitter. Specifically, we use Python to read the raw data and then convert it into csv format. We concentrate our analysis on the news published in 2015 which is representative in the dataset. Since Apple launched Apple Watch in March 9 and held Worldwide Developer Conference in June, we notice that the news releases in March and June are much more than the others. Sentiment analysis is our next step. We count the emotion words in each story, and find out certain types of emotion that appear most, which may represent the entire emotion of the story. For example, if TYPE: POSITIVE has the highest frequency, we mark the story “POSITIVE”. After that, we improve the method of Native Bayes Classifier by using compression matrix to predict the following day’s stock return based on current news stories.