New York Times Dataset Analysis – Project Report

- Sri Ram Pulipaka
- Karthik Bangera

Objective:

The main objective of this project is to understand the progression of international relations between the United States and the rest of the world over the years by performing sentiment analysis on The New York Times dataset. A sub objective is to verify the reliability of this sentiment by performing a sentiment analysis on various businesses and validating the generated sentiment against their stock performance over a given period of time.

Datasets:

a. The New York Times dataset downloaded using their 'Article Search API'. This dataset contains articles from 1851 to today, retrieving headlines, abstracts and links to associated multimedia.
b. Historical stock data downloaded from Yahoo Finance from 1980 to today, retrieving name of the company, open and closing prices and trade volume.

Sentiment Analysis Tool:

a. The sentiment analysis tool to be used in this project is the Stanford CoreNLP, a deep learning framework which can take raw text input and indicate sentiment by analyzing the sentence as a whole and assigning a score.
b. Sentiment value 1 indicates Negative sentiment.
c. Sentiment value 2 indicates Neutral sentiment.
d. Sentiment value 3 indicates Positive sentiment.

Description:

Part I: We retrieve all the New York Times articles which mention a given country (or set of countries) and perform sentiment analysis on all those articles. The analysis is performed on a month-by-month basis for every year and the results would be plotted on a graph showing the progression of Negative, Positive and Neutral sentiments.

Ex: (Country → China) would analyze all the news articles which mention China and plot the sentiment on a graph for 2007 - 2014.

Part II: This part is to check if the sentiment trends produced by StanfordCoreNLP for a company would follow the actual stock performance of that company. This is performed by correlating the stock values and the sentiments of a company.

Ex: (company → Microsoft) would calculate sentiment and stock trend of Microsoft and correlate them.

Data Loading:

The collected articles are stored in MongoDB - a document-oriented database. MongoEngine is a document object mapper is used for working with MongoDB and python.

The steps in data storage are:

1. The data downloaded consists of news articles related to various countries and a few companies. The data model for the news articles is built using mongo engine by specifying a class and the name of the class is the name of the collection (**articles**) that will be stored in mongoDB. The file models.py & models_comapnies.py, specifies the class articles that depict the model in which the data is stored in mongoDB for countries and companies.

2. We built another collection named **parsed_articles** to store information needed from the news articles to perform sentiment analysis.

   - The parsed articles collection for countries is built by extracting the glocations (geographical location) from the news article keywords. The glocation and the rank specify the relevance of the article to that particular country and we have stored article abstract or snippet with rank 1 and rank 2. If the news article abstract is not empty and greater than 45 characters then the abstract of the article is stored in mongoDB else the snippet of that article is stored.

   - The parsed articles collection for companies is built in a similar manner using the keyword 'organizations' instead of 'glocations'. We pick up article abstract or snippet with rank 1 or rank 2 and based on the same conditions above either pick the abstract or the snippet to store in MongoDB.

   - A collection named Stock is built that has information about the company stock prices on day to day basis; along with company code, volume traded, stock high price, stock low price, day opening price & day closing price.

<u>Data Processing for Countries:</u>

Hadoop is utilized to pick parsed news articles from MongoDB and use the Standford CoreNLP to analyze the sentiment of the articles for countries. Hadoop is used to perform the distributed processing of the parsed articles. The sentiment has been calculated for articles for each country from 2005.

Sentiment of each article is calculated and the no. of articles for each sentiment is stored by month. For example, below is how the data looks like for the month of October 2014.

Negative articles for 2014-10: 40
Positive articles for 2014-10: 10
Neutral articles for 2014-10: 20

Thus, a spike in the no. of negative articles for any set of given months implies that the sentiment coming out of a country for that period has been negative. This would help us understand how the negative sentiment has progressed over the months in a given year. The same with positive and neutral sentiments.

<u>Data Processing for Companies:</u>

The data retrieval for companies is the same as for the countries. But as the sentiments have to be correlated with the company's stock values, the processing part is different. The sentiment has been calculated for every article and stored in MongoDB.

The closing stock prices have also been retrieved for each day. As these values have to be correlated with the sentiment values and the range of sentiment values is (1, 2, 3), the stock values had to be coded. The stock value of any day was compared with the stock value of the previous day. If the stock decreased, that day was assigned code '1'. If the stock was the same as the previous day, that day was assigned code '2'. If the stock value increased, that day was assigned code '3'. This enabled us to directly correlate the stock and sentiment values for each day. The processed data was again stored in MongoDB.

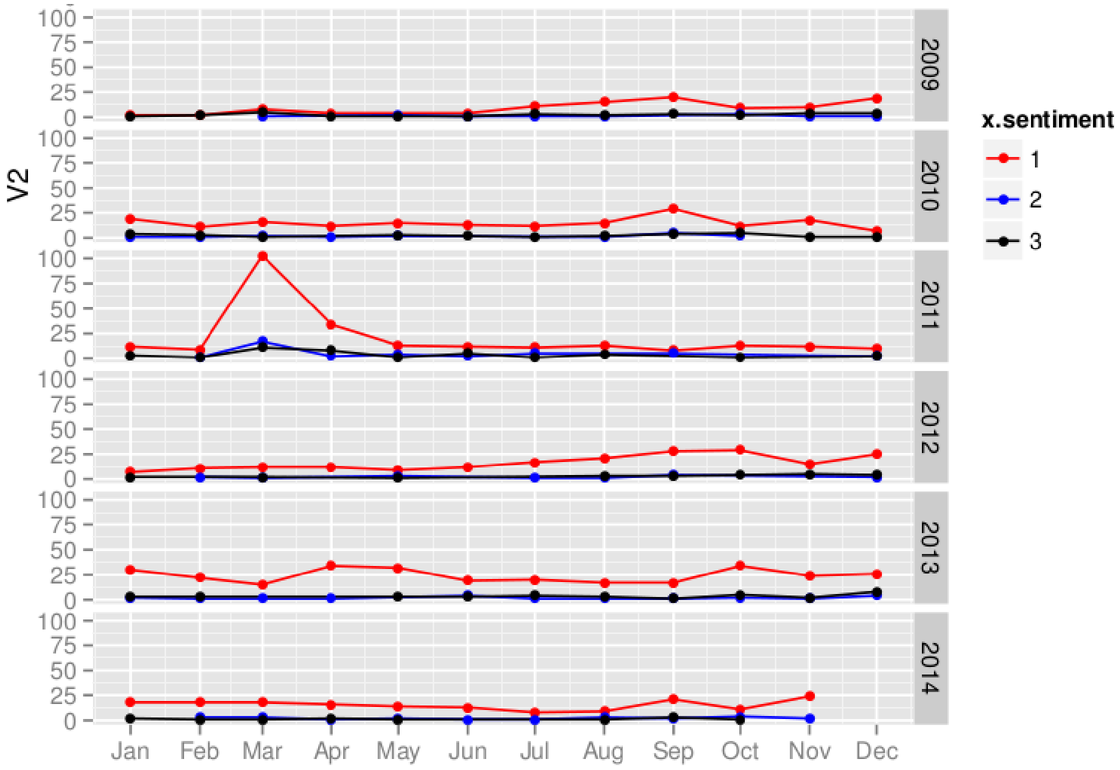
Data Visualization:

Visualizations were generated using R. For a given country, the count of articles for each sentiment has been plotted by month as a line graph using R to show the trend of Negative sentiments, Positive sentiments and Neutral sentiments over months in any given year. For a given company, the correlated stock and sentiment values have been plotted as a line graph for each day over a year.

Results:

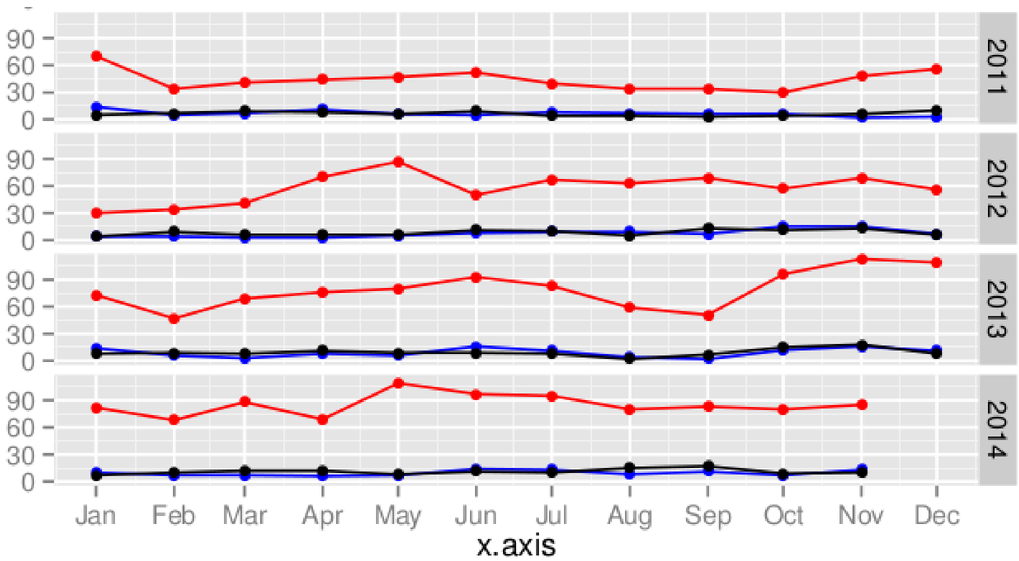
Below is the graph generated for Japan showing the no. of articles classified as Negative, Positive and Neutral. The text of the NYTimes articles having 'Japan' as the main 'glocation' was fed as input to the CoreNLP. It generated sentiments for each article.

Sentiment value 1, marked by a 'red' line, indicates Negative sentiment.
Sentiment value 2, marked by a 'blue' line, indicates Neutral sentiment.
Sentiment value 3, marked by a 'black' line, indicates Positive sentiment.
X-axis shows the months and the Y-axis indicates the no. of articles.

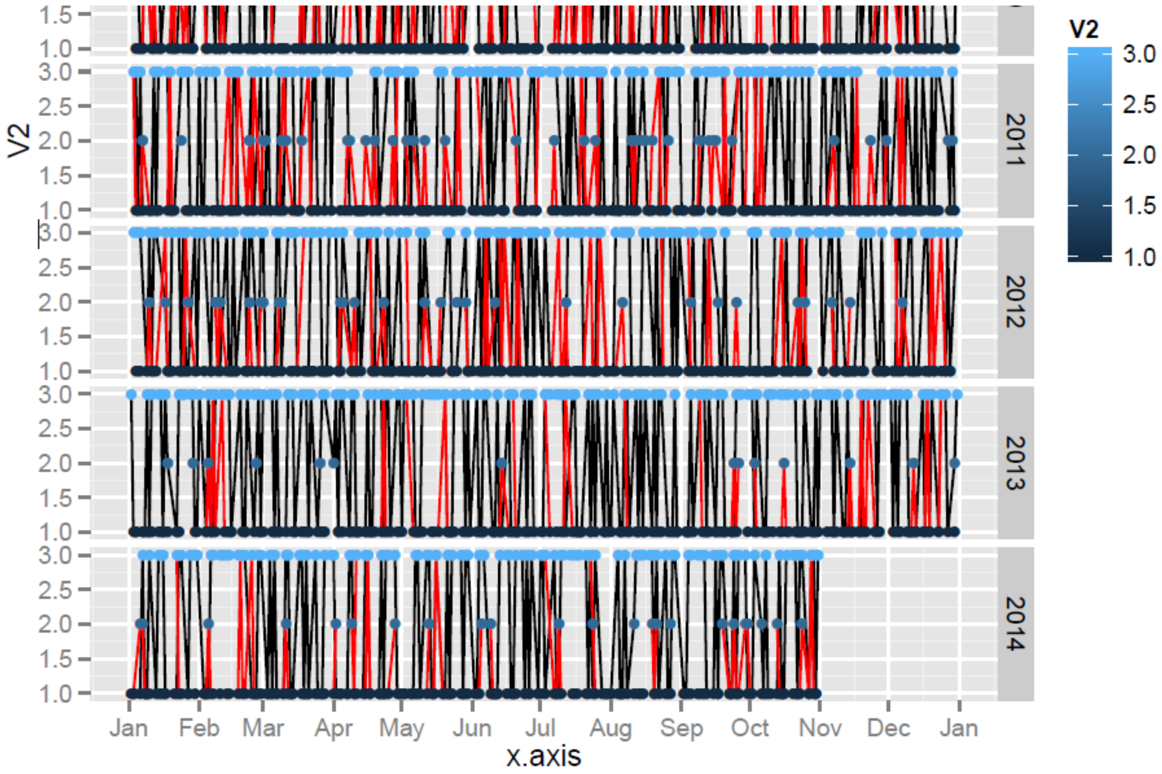


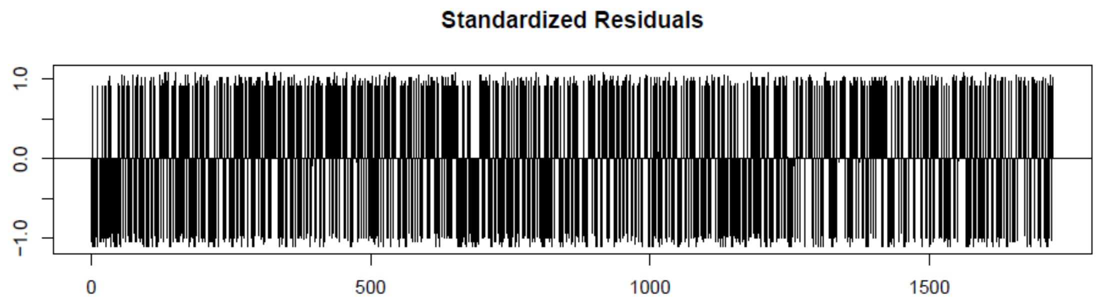
In the above graph, we can observe a large negative spike in Mar-2011. That was the exact period of 'Fukushima Nuclear Disaster' in Japan. This spike indicates that the sentiments of the articles during that time was negative. In general, it can be observed that the no. of negative articles coming out of Japan is quite low over the years and has also been fairly constant. So, the trend could be assumed to be stable over time which, in terms of 'International Relations', could be termed 'favorable'.

Below is the graph for China. It can be observed that the no. of negative articles continuously increased as time progressed.



Below is the graph showing the sentiment of news for each day vs stock and then the correlation plot between sentiments generated for Microsoft and its actual stock performance. Black lines indicate stock values and the red lines indicate sentiment values.





The standardized residuals above have not decreased at any point and the results indicate a correlation of approx. 30% between the stock and sentiment values, which implies that the sentiment values generated for NYTimes articles on Microsoft do not really reflect the Stock performance of Microsoft and hence is not a good indicator of stocks.

Inferences:

From the results above, we can conclude that while the CoreNLP sentiment on NYTimes articles gives an acceptable trend of sentiments of the news coming out of a country, it's not really a reflector of the stock performance.

Countries on which analysis has been performed:

India, China, Japan, Afghanistan, Taiwan, Venezuela, Canada, Australia, Nigeria, Iran, Iraq, Russia, Brazil, Thailand, Cuba, France, Germany, Italy, Israel.

Improvements:

This project could be improved in various ways:

1. Training the StanfordCoreNLP models with large amounts of NYTimes data would give better results. This requires huge manual work though to build the training datasets.
2. Though the news has been cleaned to some extent to pick only the articles relevant to a country or company, this cleaning could be improved so that we pick the only those articles which are very relevant.
3. Also better statistical models could be utilized to make more sense out of the data.

Technologies Used:

1. Stanford CoreNLP – to perform sentiment analysis of the news articles collected.
2. Python wrapper - to download data from New York Times using their 'Article Search API'.
3. MongoEngine (a Python Object-Document mapper for MongoDB) – to load datasets into MongoDB.
4. Mongo-Hadoop connector – a plugin for Hadoop that provides the ability to use MongoDB as an input source and/or an output destination.
5. Hadoop – to perform data cleansing and sentiment analysis.
6. R - To plot the graph.

Objectives Achieved:

1. Using cloud environments like OpenStack and FutureGrid systems.
2. Using a management tool like Cloudmesh.
3. The concepts of big data like Data Procurement, Cleansing, Storing and Analysis.
4. Configuring and deploying Hadoop and MongoDB across clusters.
5. Working with Hadoop and MongoDB.
6. Internals of MongoDB and building data models.
7. Using other tools like MongoEngine, R etc.
8. A brief introduction to the field of Natural Language Processing.

Team Members:

Sriram Pulipaka (sripulip@indiana.edu), Karthik Bangera (kbangera@indiana.edu)

References:

1. I590 Class Lectures by Prof. Geoffrey Fox - http://bigdataopensourceprojects.soic.indiana.edu/
2. J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Commun. ACM, vol. 51, pp. 107-113, 2008.
3. Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.
4. Hadoop: The Definitive Guide, 3rd Edition – Tom White
5. MongoDB: The Definitive Guide – Kristina Chodorow
6. Stanford CoreNLP. Available: http://nlp.stanford.edu/software/corenlp.shtml
7. Apache Hadoop. Available: http://hadoop.apache.org/
8. MongoDB. Available: http://www.mongodb.org/
9. The R project for statistical computing. Available: http://www.r-project.org/
10. MongoEngine. Available: http://mongoengine.org/
11. The New York Times API. Available: http://developer.nytimes.com/
12. Yahoo Finance. Available: http://finance.yahoo.com/