



# PULSE OF NEW YORK TIMES

## (May 2017)

### Big Data Management and Analysis

Alexey Kalinin, Enrique González, Priyanshi Singh, Shalmali Kulkarni, Sunny Kulkarni

#### A. INTRODUCTION

Social media is a new digital press of 21<sup>st</sup> century not only challenging traditional media in spreading information and influencing minds, but also offering additional powerful opportunities for sharing moments, experiences, reactions and feelings. Analysis of social media is one of the ways to understand the complex urban context using the human opinions and sentiments expressed through the social media.

Sentiment analysis has received a significant attention from researchers from every field in the last decade. Analysis of emotions and opinions of news articles could be different from that of other text types. There are various platforms to record opinions about news articles such as comments, sharing of the articles, personal blogs, etc and it has created new opportunities and challenges for researchers to track down the attitude of other people. The simple definition of sentiment analysis is to determine whether a piece of text is positive, negative or neutral. The most use case scenario of this analysis is to understand people's emotions about a particular topic. This emotion identification can then be used for diverse purposes.

#### B. GOALS & OBJECTIVES

The aim of the project is to perform Sentiment Analysis / Opinion Mining for the New York Times articles. In this project, tweets were used as one of the metrics to define emotions on the trending NY Times articles. The popularity of an article is defined in many ways such as number of views, number of searches using Google, number of emailed links of an article, number of shares on social media, etc. The popularity of the articles presented by NY Times and its sentiment on social media platform - Twitter is studied in this project. Twitter is selected here, as it is the most popular microblogging platform. Twitter contains a very large number of short messages, referred to as tweets, created by the users of this microblogging platform. The contents of the messages vary from personal thoughts to public statements making it an ideal source for accumulating a vast amount of opinions towards a wide diversity of topics.

#### **Project Objective -**

The first objective was to understand New York Times trends on Twitter across the U.S. The project initially started with a focus on tweets from New York City only but later nationwide tweets were used



considering significant readership of New York Times all across the country.

The second objective of the project was to reveal spatial clusters for the tweeted subjects and understand if a pattern exists for twitter activity related to the different topics e.g. politics, business, entertainment, sports. However, considering the time frame used for data collection across the United States, from the end of April till the beginning of May, only few trending topics were captured. The sentiment analysis of these popular articles allowed us to extract people's emotions and reactions to popular NY Times articles across the U.S. and visualize them.

### C. DATASETS

As the aim of the project is the classification of the emotions of Americans towards the trending news, especially NY Times articles, understand their reactions in order to investigate what relevant topics matters to them. As stated before, the project focused on the sentiment reaction to these trending articles and their spatial distribution of the twitter activity across the country. To perform analysis we collected geocoded tweets from Twitter and the list of most popular articles from NY Times. We also used Sentiment Word List from the NLTK package to assign sentiment score to the tweets. All these sources and package are openly available to public and the API source code is easily accessible. The following dataset was collected and used for analysis.

#### *a. Twitter*

Twitter produces a huge amount of data on a daily basis. The data thus generated can be

accessed using the Twitter search API to collect the geo-tagged and time stamped data. The data collected was for a time intervals of one week to get approximately 3GB of data across USA.

The data collection time frame spanned from 28th April 2017 to 3rd May 2017.

For this timeframe, 3.23 Gb of Twitter feed was collected. It is comprised of 600,000 tweets for US region. The number of geo-coded tweets in this collected feed is approximately 111,000. This data is considered for further restriction and analysis.

#### *b. NY Times Articles*

The primary data to be collected is the NY Times articles which will be link to the twitter reactions. NY Times articles are available from 1851 to present with more than 13 million articles in total. This project aims to use only the most popular articles of the day for analyzing the tweets. This data is fetched through the NY Times API. The NY Times API provides 10 most popular article for each day based on the number of views received for the article.

#### *c. Vader Lexicon Wordlist*

There are multiple publicly available resources for sentiment analysis from Stanford University, Princeton University, etc. The VADER Lexicon wordlist from the Natural Language Toolkit (NLTK) is used for the project. It has 9,000 lexical feature list including positive, negative, neutral words list, emoticons, sentiment-related acronyms and initialisms (e.g. LOL) and commonly used slang with sentiment value. The VADER sentiment lexicon is sensitive



in both the polarity and the intensity of sentiments expressed in social media contexts, and it is also generally applicable to sentiment analysis in other domains.

## D. METHODOLOGY

Sentiment Analysis on Twitter Data is a challenging problem due to the its nature, diversity and volume of the data. In order to address the nature of Big Data, some pre-processing steps were introduced for achieving better results in Sentiment Analysis.



The project uses Apache Spark as it is a fast and general engine for large-scale data processing. The speed, generality and ease of use makes Apache Spark the choice for the project. The data-preprocessing phase would include the count of tweets for each of the popular NY Times articles along with its location and time.

### *a. Data Preprocessing*

The Twitter raw data is in 'nested dictionaries' json format with almost 35 categories divided into multiple subcategories e.g. the 'entities' has subcategories like text, media, symbols, hash tags, etc. The schema for Twitter data is elaborate and has different types of data from text, images, gifs, etc. The initial phase of data cleaning was to understand the schema and filter the categories required for the analysis.

The data collected from Twitter was merged to NY Times articles based on the 'url' field

from the Twitter json e.g. '<https://www.nytimes.com/2017/05/10/us/politics/comey-russia-investigation-fbi.html?smid=tw-nytimes&smtyp=cur>'. It was observed that majority of tweets also had the NY Times article embedded within the tweet text as they might be tweeted from the NY Times website or mobile application.

All the tweets containing 'www.nytimes.com' or '#NYTimes' and '@NYTimes' from the text field were selected for further sentiment analysis.

The next step was to filter the tweets that contained geolocation information. This process reduced the data size considerably as geolocation is not available as a default feature. It is available only at the discretion of the user and requires to be turned on for the user's account in Twitter settings section..

### *b. Sentiment Score*

The basic idea behind sentiment analysis is to apply simple statistical measures of text - such as word length, difficulty, and density - that a computer can learn to associate with things people actually care about, like personal preferences giving the pulse of the city, state, country, etc.

A comment is 'positive' if it has a normalized score (score / number of words) of 0.1 or higher and negative if the score is - 0.1 or lower. Between -0.1 and 0.1 a comment is considered as 'neutral'.

The *compound score* is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme

positive). This is the most useful metric given a single dimensional measure of sentiment for a given sentence. It is a 'normalized, weighted composite score'.<sup>1</sup>

### c. Natural Language ToolKit (NLTK)

NLTK is a lexicon-based approach for sentiment analysis which relies on a fixed and pre-assigned sentiment polarities of words. NLTK processes text in below steps -  
-Word Tokenizer (text prep)

In this step, the text is prepared for analysis by tokenizing the words in the tweet. Sent\_tokenize is used from the NLTK package which uses an instance of PunktSentenceTokenizer to separate each word, remove all the punctuation marks, split the emoticons from the text and also change all the words to lowercase.

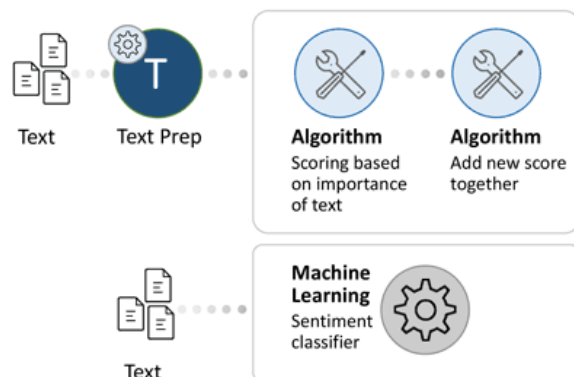


Figure 1: Text Analysis Process<sup>2</sup>

### -Sentiment Scoring

Once the words are ready to be analyzed, the next step is assigning scores to these words.

The VADER lexicon word list is used to score these tokenized words. The polarity score is calculated for each tweet.

Further this could be used to model a machine learning classifier and predictive modeling.

### d. Apache Spark - Big Data Platform

The complete analysis is performed using Spark data frame. The data cleaning, filtering and the sentiment analysis is all performed using a single script. This huge amount of data was processed quickly using the Hadoop ecosystem as opposed to local machine, which would require longer time to perform the same analysis. Spark provides a platform to scale the project from a small subset of data to the final 3 GB data without any changes to the existing script. The biggest advantage was performing exploratory queries on large amount of data.

## E. RESULTS

The sentiment analysis scores for these tweets are -

Positive - 38,131 tweets

Negative - 10,753 tweets

There are more positive tweets than negative, but the overall sentiment across US is neutral.

<sup>1</sup> Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

<sup>2</sup> Sentiment Analysis  
<https://www.lexalytics.com/technology/sentiment>

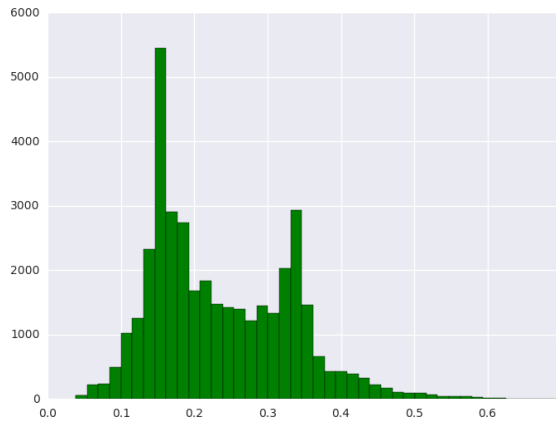


Figure 2: Positive Sentiment Distribution

It can be observed from the positive sentiment score distribution that the average positive score is 0.24, which is not a strong positive sentiment. The maximum number of positive tweets is about 0.15 score and the next peak is about 0.35.

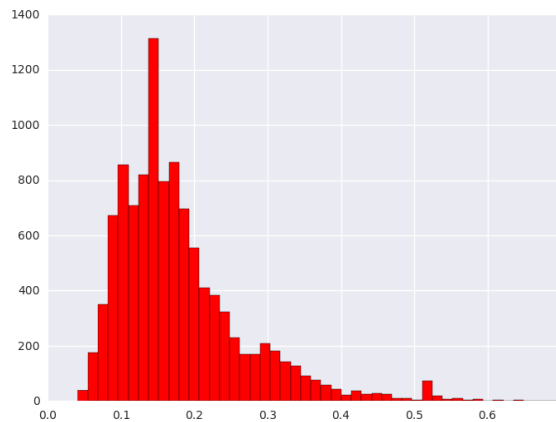


Figure 3: Negative Sentiment Distribution

The negative tweets have a low average of 0.18 indicating a soft negation towards the NY Times articles. There are very few (only couple of hundreds) strong negative sentiments.

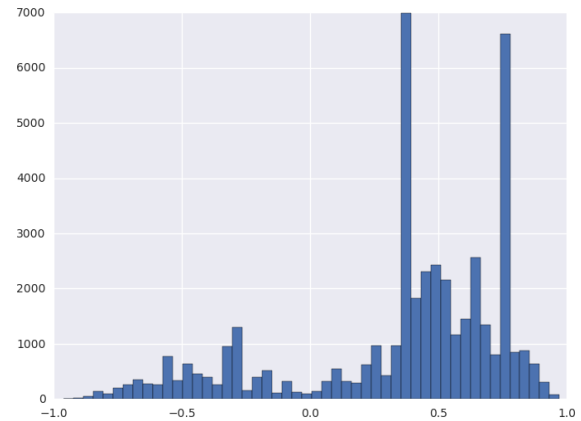


Figure 4: Compound Score

As discussed earlier the compound score is *normalized weighted score* ranging from -1 to +1. The figure shows the complete range of score, which is skewed towards positive scores indicating more positive sentiment for the tweets.

## Visualization - Sentiments across USA

As part of the descriptive analysis, the tweets are mapped into three categories of sentiment analysis to start exploring the different sentiments aroused by the news across the country.

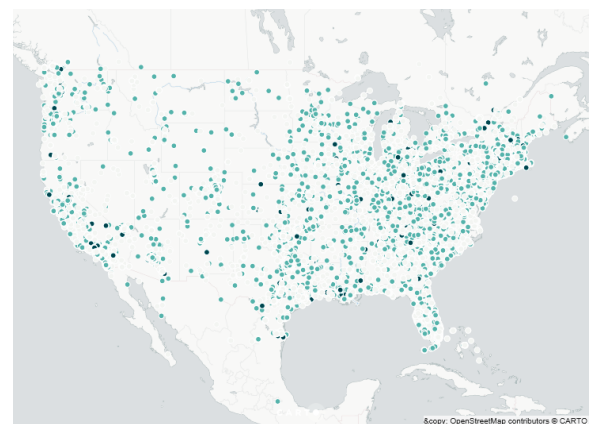


Figure 5: Map of Positive Sentiments

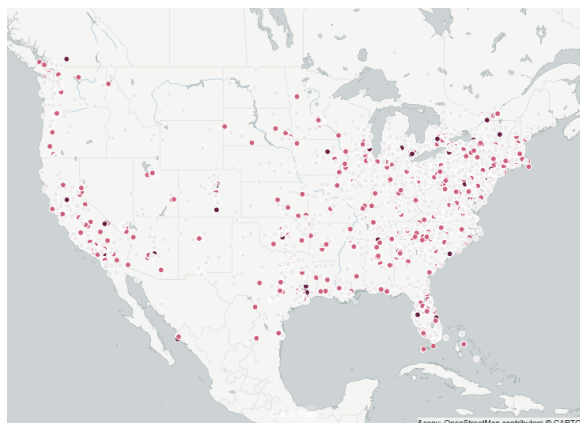


Figure 6: Map of Negative Sentiments

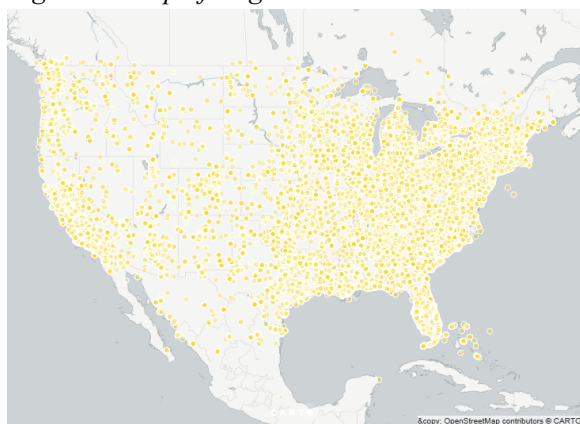


Figure 7: Map of Neutral Sentiments

Comparing the three sentiment visualizations (Figure 5, 6 & 7) it can be inferred that the most frequent sentiment is the neutral, which is spread uniformly across the country in its highest value. It can be clearly seen that mostly the tweets contains more neutral words than strong or negative ones.

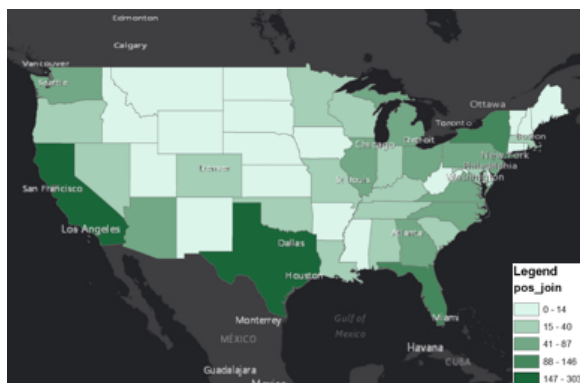


Figure 8: State level positive sentiments

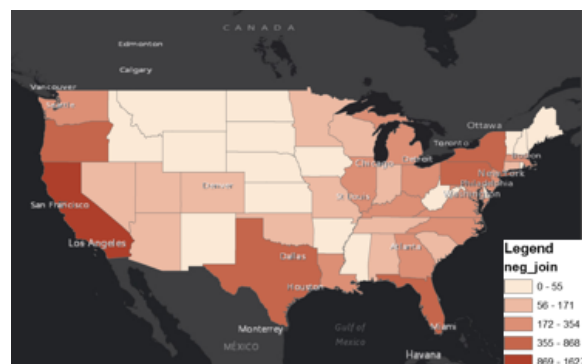


Figure 9: State level negative sentiments

The state level analysis of the sentiment is shown in above maps (Figure 8 & 9) the highest positive as well as negative tweets are from Texas, California and New York. This might be due to more readership from these states or more Twitter users from these states. The top three states have more negative tweets than positive ones. Thus, indicating that the predominant sentiment from these states is negative for the trending New York Times articles.

## Trending Topics -

Most tweeted topics were analyzed to better understand the sentiments in the Twitter feed and their location. For the analysis period of 28th April to 3rd May 2017, news articles similar to below two articles influenced the tweets -

“On Donald Trump’s First 100 Days” and  
“The Cost of Barack Obama’s Speech”

Current political situation in US is strongly divided in opinions. As seen in the above Visualization for Sentiments across USA, a pattern is seen for the positive and negative tweets across the different regions of US.

Also, we have to note here that the tweets would be a biased representation as the data





is limited for a short span of time and only geocoded tweets are considered here.

## Visualization - Spatial Clustering

Clustering is an effective unsupervised technique for preliminary analysis, as the clusters are formed by similarity of characteristics between features, grouping those characteristics. Twitter allows its users to geolocate the tweets making it relevant to analyze the spatial patterns. The filtered dataset has approximately 100,000 geolocated records, which were grouped, based on similar sentiment scores. K-means clustering technique was performed using spatial features and sentiment score. K-mean method used the ‘*euclidean distance*’ metric to give clusters. In the beginning, a small number of clusters were chosen, from 3 to 5, yielding the result that tweets with the same sentiment are also located in the similar geographic area i.e. sentiments are spatially correlated. The result of these initial clusters were the following visualizations:

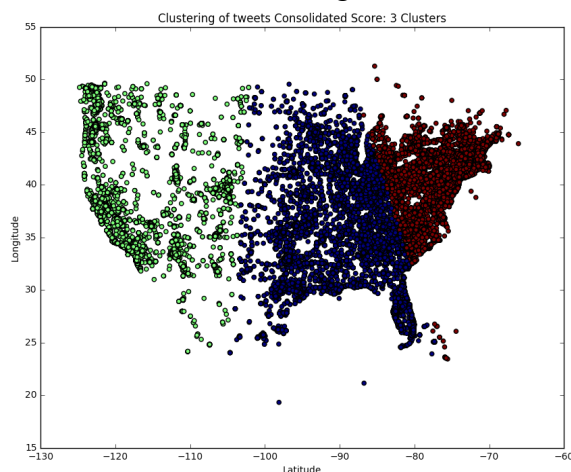


Figure 10: Spatial Clustering- Three clusters

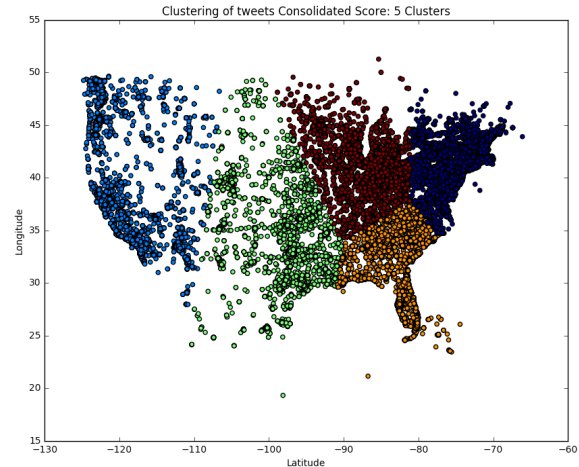


Figure 11: Spatial Clustering - 5 clusters

The first interesting result - due to the intrinsic characteristic is the first data set (New York Times News) - is that there are more responds in the east coast and closer areas than in the middle of the country where is a gap. Moreover, it can be seen how the west coast is very similar to the east coast. The second interesting result is the reaction to the news aka sentiment analysis; it can be analyzed that North part of the country reacts similarly and some parts of the middle react in the same way as part of the North and the West. Visualization has been performed using CartoDB tool.

## F. BIG DATA CHALLENGES

During the course of project there were many challenges relating from large volume of data to unstructured format of the data. Below are some of the areas of difficulties faced and the solutions adopted to manage them.

*Scalability* - The original idea was to identify specific articles in New York Times that are tweeted in high volume in the New York City. Our preliminary analysis showed that New York Times has high readership across United States. Since the processing

time required in Spark Pipeline and in the Hadoop Cluster system was very low (few minutes) for the New York City twitter feed volume, the analysis was expanded to United States region.

**Data Volume & Quality** - Twitter provides API to collect feed from current point onwards. The only restriction is no ability to collect it for the historical period. Since twitter is the go-to medium for many social media followers, the twitter feed gets large in a very short span of time. We applied below restrictions to limit the volume, requisite for our analysis-

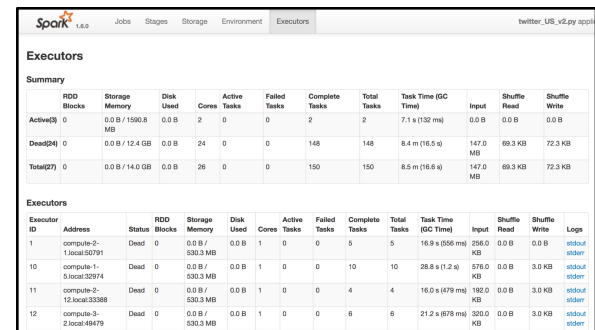
- Limited the feed for a specific date range, from 28th April to 3rd May 2017.
- Limit to only geo-tagged tweets, as we need to map the sentiments against the location.
- Filter the tweets connected to the NY Times articles using the web URL as connecting link or the article and NY Times reference in hash tag area.

**Data Structure** - Twitter feed is received in a highly nested JSON file. It is a rich quality data source with information such as hash tags, url, emotional icons, animated image (gif) files as well as embedded videos. It has information on count and details for retweets, liked, followed, comments, geo-location, etc. This data is a hybrid of structured and unstructured data. Information for each tweet spans across a page in the nested JSON format structure. Understanding this structure and extracting the relevant fields for analysis is a time consuming activity and required careful analysis.

## Processing Time

Using capabilities of Spark Pipeline and Hadoop Cluster, the job required less than 2 minutes and 30 seconds to process 3.23 Gb of data. (Refer Image 1: Spark Job Processing Time and Image2: Spark Job Log) This job was executed from NYU Dumbo system. Input and Output data resides on NYU Hadoop Cluster system.

Spark Pipeline first filters the data to relevant tweets and then NLTK library is used to calculate positive, negative, neutral and cumulative score for each tweet. The output file comprises of these 4 scores along with details of each tweet such as its location, text information, user id, etc.



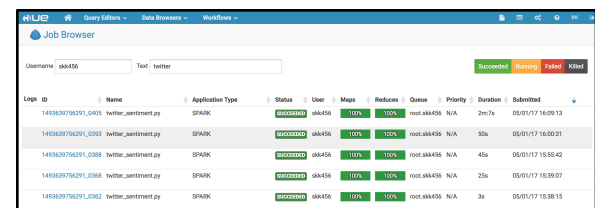
Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active[2]	0	0.0 B / 1590.6 MB	0.0 B	2	0	0	2	2	7.1 s (132 ms)	0.0 B	0.0 B	0.0 B
Dead[24]	0	0.0 B / 12.4 GB	0.0 B	24	0	0	148	148	6.4 m (16.5 s)	147.0 MB	69.3 KB	72.3 KB
Total[27]	0	0.0 B / 14.0 GB	0.0 B	26	0	0	150	150	6.5 m (16.6 s)	147.0 MB	69.3 KB	72.3 KB

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
1	compute-2-1.local:50791	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	5	5	18.9 s (356 ms)	256.0 KB	0.0 B	0.0 B	stdout stderr
10	compute-1-3.local:50784	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	10	10	28.8 s (1.2 s)	576.0 KB	0.0 B	3.0 KB	stdout stderr
11	compute-2-12.local:53388	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	4	4	16.0 s (579 ms)	192.0 KB	0.0 B	3.0 KB	stdout stderr
12	compute-2-12.local:49479	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	6	6	21.2 s (879 ms)	320.0 KB	0.0 B	3.0 KB	stdout stderr

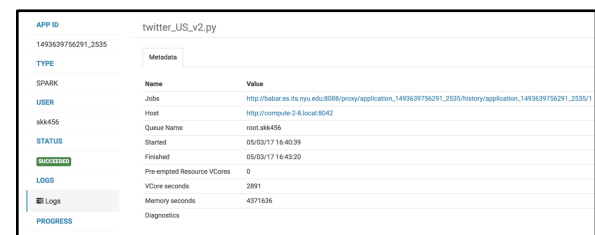
Figure 12: Spark job log



Job Browser

Log ID	Name	Application Type	Status	User	Maps	Reducers	Queue	Priority	Duration	Submitted
1493639756291_0403	twitter_sentiment.py	SPARK	Success	ak456	100%	100%	root.aka456	N/A	2m7s	05/01/17 16:09:13
1493639756291_0503	twitter_sentiment.py	SPARK	Success	ak456	100%	100%	root.aka456	N/A	50s	05/01/17 16:00:31
1493639756291_0308	twitter_sentiment.py	SPARK	Success	ak456	100%	100%	root.aka456	N/A	45s	05/01/17 15:58:42
1493639756291_0308	twitter_sentiment.py	SPARK	Success	ak456	100%	100%	root.aka456	N/A	25s	05/01/17 15:39:07
1493639756291_0302	twitter_sentiment.py	SPARK	Success	ak456	100%	100%	root.aka456	N/A	3s	05/01/17 15:38:15

Figure 13: Spark job-processing time



twitter\_US\_v2.py

NAME	VALUE
TYPE	SPARK
USER	ak456
STATUS	Success
LOGS	Log
PROGRESS	Progress
Job	http://barbar.es.ltu.nyu.edu:8088/preview/application_1493639756291_2335/History/application_1493639756291_2335/1
Host	http://compute-2-8.local:8042
Queue Name	root.aka456
Started	05/01/17 16:40:39
Finished	05/01/17 16:42:20
Pre-empted Resource VCores	0
VCores seconds	2891
Memory seconds	4271636
Diagnostics	

Figure 14: Spark job





## G. WAY FORWARD

The logical future direction for this project would be to -

- a. Spark Streaming - Currently, data is downloaded using Twitter and NY Times API to local machines and transferred to Hadoop system using UNIX commands. This raw data is an input to the python spark script executed on the NYU Dumbo cluster. Using Spark Streaming, the live twitter feed would be an input to the script giving run time output of scores for the tweets recently posted. The visualization would then have most recent data giving a live display of sentiments. A filter capability to key in specific NY Times articles would be useful for the application to gauge the people's sentiments and reaction to the latest trending articles on important topics. Also, NYU Dumbo and Hadoop systems are secluded inside Firewall and setup architecture would need evaluation to check Dumbo systems access to live twitter stream.
- b. Spark Machine Learning (ML) Libraries - Explore and apply Spark ML libraries to enable prediction of sentiments for different regions of the US before publishing a new article. Build & train models using sentiment scores and reactions for past news articles as input. This could be an effective engine to anticipate how people would react to a written segment.
- c. This application would also be useful to writers to understand the reader's responses and possible with improvement insights. It is also useful for editors to compare different writing

styles and evaluate reader participation in sharing viewpoints for trending articles. As news media inherently should take neutral standpoint when reporting news, presenting only facts and leaving it upto the reader to make his viewpoint. This viewpoint is loosely received in terms of unstructured data as feedback. Analysis of this feedback data is crucial to understand the thought process of the common man and woman.



## H. REFERENCES

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

NYT most popular API

<https://developer.nytimes.com/>

Twitter API

<https://dev.twitter.com/rest/public/search>