|  |
| --- |
|  |

Table of Contents

[Figures 3](#_Toc7997649)

1. [Introduction 4](#_Toc7997650)

[1.1 Overview 4](#_Toc7997651)

[1.2 Prelude to Astronomically immense Data 4](#_Toc7997655)

1.3 Sources of big data [5](#_Toc7997656)

[1.4 Why we need big data? 6](#_Toc7997657)

[1.5 What is big data analytics? 6](#_Toc7997658)

[2. overview technologies and platform used 7](#_Toc7997659)

[2.1 Twitter 7](#_Toc7997660)

[2.2 apache spark 7](#_Toc7997661)

[2.2.1 Spark streaming 8](#_Toc7997662)

[2.2.2 spark sql 9](#_Toc7997663)

[2.2.3 spark ML lib(Machine learning libraries) 9](#_Toc7997664)

[3. python sentiment Analysis 10](#_Toc7997667)

[3.1 sentimement analysis 10](#_Toc7997668)

[3.2 flow diagram 15](#_Toc7997669)

4. IMPLEMENTATION [15](#_Toc7997670)

[4.1 streaming twitter data using spark 15](#_Toc7997667)

[4.2 Understanding data 16](#_Toc7997668)

[4.3 Pre-processing data 19](#_Toc7997669)

4.4 Developed model [20](#_Toc7997670)

[5. Execution 21](#_Toc7997668)

[5.1 Building project 21](#_Toc7997669)

5.2 submitting job [23](#_Toc7997670)

[6. USE-CASES AND results 24](#_Toc7997668)

7. Future work [25](#_Toc7997670)

[8. conclusion 25](#_Toc7997669)

references [25](#_Toc7997670)

Figures

[Figure 1: V’s of Big Data 4](file:////Users/pearl/Desktop/Uni%20Work/ADBMS/project/AlmeidaPearl_Report.docx" \l "_Toc7996633)

[Figure 2: Spark Architecture 8](file:////Users/pearl/Desktop/Uni%20Work/ADBMS/project/AlmeidaPearl_Report.docx#_Toc7996633)

[Figure 3: Spark Streaming 9](#_Toc7996635)

[Figure 4: Sentiment analysis flow 10](#_Toc7996636)

[Figure 5: Flow Diagram 15](#_Toc7996639)

[Figure 6: Spark Streaming 16](#_Toc7996640)

[Figure 7: Basic workflow model 20](#_Toc7996641)

[Figure 8:The result graph 24](#_Toc7996642)

# Introduction

1.1 Overview

In today's world, there is an abundance of informative data available on convivial media. This data has been a potent resource for data miners to get deep insights. Amassing this data and analyzing it might give us a plethora of subsidiary information which can avail solve an abundance of our day to day quandaries. Taking this concept into consideration, I have harnessed an implement that accumulates data from Twitter and analyses it to engender keenly intellective information. Suppose we utilize this implement according to a consumer/buyer perspective it can be utilized in the following way. This app amasses the Tweets that relate to brands, products or topics set as filter and gives us a review as to which is the best for a given geographical location, what are the sentiments associated with it at a particular location, which brand, product or topic is most trending, what the trend looks akin to on different dates, how popular it is and how many adherents it has and thus availing the consumer to make sapient culls. Consider buying a gaming console as our target mission, we opted to analyze gaming consoles that belong to three brands verbally express Xbox, PlayStation, Nintendo. This implement is not just constrained to gaming consoles but can be acclimated to get reviews and insights for any product or any trending topic for example phones, cars, electronic contrivances. The results can be visualized it sundry graphs like a heat map, bar graph, trending graph, density graph, etc.

1.2 Prelude to Astronomically immense Data

Today, we live in the digital world. Because of the growing digitization and innovations in technologies, the amount of data (structured or unstructured) and data amassment have withal tremendously incremented. The data are deposited in databases that grow substantially and become arduous to amass, manage, share, examine and visualize via traditional database software implements. This is the reason for the emergence of astronomically immense data and a recent area of strategic investment for IT organizations.

Immensely colossal Data can be defined as an amassment of data which may be structured or unstructured and so astronomically immense and perplexed that it may become involute to process it utilizing simple systems or traditional data processing applications

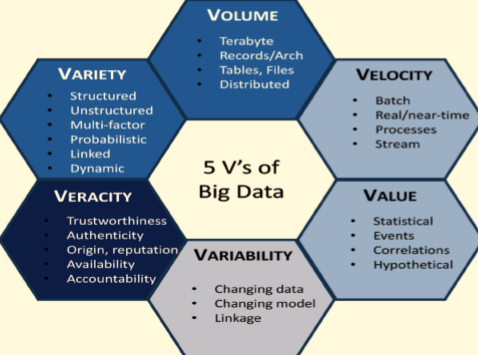


Figure 1 V’s of Big Data

The characteristics of Immensely colossal Data are the following:

Ø Volume

Volume refers to the huge amount of data transformed every second which is not possible for traditional systems to manage. For example, consider a convivial networking website like Facebook where people upload pictures, send messages, comments and click the Like button billion times on a quotidian substructure.

Ø Velocity

The celerity to engender incipient data and kinetically circumvent are verbally expressed as Velocity. For example, the messages or videos on gregarious media that’s going viral within a minute and expeditiously detect suspicious activity on credit card transactions. Systems designed to handle sizably voluminous data can handle and process millions of rows per second, which makes it more facile to get desired insights out of this data.

Ø Variety

The data like structured or unstructured, videos, graphs, images, text messages, tables, audios, etc., are referred to as Variety. These different forms of data can be accumulated via sundry sources like convivial media, sensors, mobile contrivances, regime datasets, utilizer-predicated, etc. This data can have many dimensions and withal sundry forms.

Ø Veracity

Veracity refers to the reliability or disorderliness of the data. The quality and precision are not facile to manage with many forms of sizably voluminous data. For example, the post on Twitter or Facebook with many hashtags, typos, and informal verbalization. The system has many worker nodes and threads running to consummate the desired task which works in synchronization to maintain the reliability of data.

Ø Value

Value refers to the business cost which is obtained through managing astronomically immense data. For example, managing prodigiously and sizably voluminous unstructured data from blogs or a gregarious media stream. The value of data depends on how precise are the insights obtained from the data, the processing time, how deep the mining is done and how it is visualized.

1.3 Sources of Immensely colossal Data

With the evolution and innovation of technology, the amount of engendered data is withal incremented. Withal, the data needs to be amassed in a variety of formats. Sources of Immensely colossal Data can be generally relegated into six different categories:

Ø Enterprise Data

There is a humongous amount of data spread across the sundry businesses, companies, and institutions in different formats kenned as Enterprise Data. The formats include emails, word documents, spreadsheets, presentations, HTML pages, pdf files, XMLs, flat files, legacy formats, JSON, CSV, tables, graphs, etc.

Ø Archives

In today's world, all insignificant data are being achieved by enterprises. As hardware getting cheaper day-by-day, the enterprises can pile up the data. This kind of data includes scanned documents and agreements, records of ex-employees, project documents, and all banking records.

Ø Transactional Data

There are countless applications like web applications, mobile applications, CRM systems, banking systems, accommodation providing platforms, etc., in every organization which involves different kinds of transactions. So, the relational databases can be utilized as a backend to reinforce these applications.

Ø Public Data

The publicly available data like Wikipedia, data from weather department, data published by research institutes, open data-sets provided by the regime and other types of data which are facilely available and accessible to the public at no cost. This type of data is called Public Data.

Ø Social Media

Gregarious networks like Facebook, Twitter, etc., engender astronomical amounts of data every day which are unstructured data which can be text, images, video, and others. This data can be cleansed, authoritatively mandated and transformed into structured data to make it subsidiary data.

Ø Activity Generated

The data being engendered by machines are referred to as Activity engendered data. The inception of these kinds of data emanates from satellites, medical contrivances, sensor data, industrial machinery, surveillance videos, cell phone towers, etc. This data may be present in Log files, log tables, Json format and can be acclimated to get future prognostications of the system.

1.4 Why we need Big Data?

mundane system hardware and software are not capable to deal with very substantial amount of sundry types of data that are engendered and amassed at such a high celerity. Sizably voluminous data is the term for astronomically immense and perplexed data sets that it becomes arduous for traditional data warehousing to store, analyze, manage, process and work on them and visualize. The insights gained by processing sizably voluminous data can be consequential for a business, can avail consumers, can be acclimated to get predictive models for natural calamities and evade them, and can be acclimated to soothsay behavioral patterns and trends. That is the reason for astronomically immense data and its analysis is the main focus of modern science and business.

1.5 What is Big Data Analytics?

“Big data analytics is the process of examining immensely colossal data sets containing a variety of data types, like sizably voluminous data, to denude patterns, unknown correlations, market trends, customer predilections, and other subsidiary business information.” Utilizing Immensely colossal Data Analysis consequential insights can be gained. For this project, we utilize Astronomically Immense Data Analytic for getting insights from twitter data.

# OVERVIEW TECHNOLOGIES AND PLATFPRM USED

2.1 Twitter

Twitter [5] is a gregarious networking platform used to apportion diminutive bits of information across the world. It is the most popular microblogging website in today’s world. The 140 characters’ messages/posts are called tweet and people can follow each other to receive other person's tweets. They can tag others utilizing @ symbol. Like all gregarious media forum, these tweets can include URLs, photographs, and hashtags. A person can apportion another person's tweet (re-tweet) if he is following the later. Hashtag (#) is utilized as a trendsetter in twitter and used to facilely look for information on a particular topic while amassing data from twitter. Twitter has more than a billion registered utilizer accounts and around 317 million monthly active Twitter users. It contains massive amounts of data and included users from all fields like movie stars, brand reviews, news, sportsmen, prevalent people, politicians, etc. thus giving us the ecumenical perspective of people around the world. Thus accumulating the expressions of people from different walks of life, can give us a clear picture of a particular topic.

2.2 Apache Spark

Spark is a framework that provides cluster computing platform to perform sundry tasks like data analytics, machine learning, data streaming, database management, parallel computing, graph operations, etc. Spark can run as a standalone cluster as well as in cluster mode.

Spark uses RDD (resilient distributed dataset) to distribute items over the cluster. RDD’s are read-only datasets and are handled by spark for the purport of maintaining its fault-tolerant deportment. Aside from RDD’s spark withal has DataFrames which are table-like structures utilized by the spark to store data in a table format, for manipulation of data in dataframes spark withal provides SQL libraries and functions. Spark has Spark SQL library which can be acclimated to perform SQL queries on data frames. Apache Spark fortifies Java, Scala, Python and R. Spark Core is the main engine that manages input-output operations, scheduling, recollection management, networking interfaces and dataset (RDD predicated).

The spark structure can be understood from the diagram below:

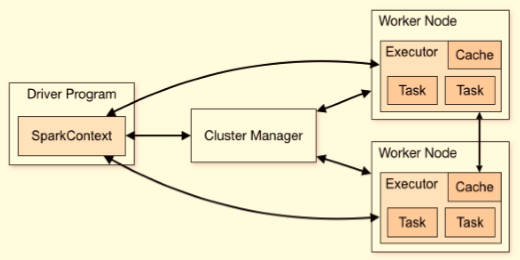


Figure 2: Spark Architecture

Ø Driver Program

The driver program manages the allocation of tasks to the worker nodes. It is additionally called as Master. The driver does the task of heedfully auricularly discerning worker nodes for possible incoming messages. It additionally keeps the task isolated so there is no data leak between the tasks. It prioritizes the task of managing the jobs submitted to the spark cluster.

Ø Executor

Executor is the process initiated for the execution of an application. Each application running on the cluster has its own executors. The executor is responsible for data keeping the data and doing input-output operations between the applications.

Ø Worker Node

Each worker node has an executor who can perform many tasks. Every worker has a cache recollection allocated which is configurable. Every executor isolates the task so there is no recollection leak between the multiple tasks submitted. The only way to apportion recollection between two tasks is to indite it to external recollection.

Ø Cluster Manager

Cluster manager distributes the task to sundry worker nodes. It is like the intermediator between the driver program and worker nodes which manages the cluster when it is distributed. The cluster manager handles the request when either od driver or worker node is diligent. This makes spark capable of fault-tolerant. The cluster manager fortifies sundry applications and packages handlers.

Ø Task

Task is a unit of work assigned by an executor.

A spark cluster can be configured with various parameter settings. The number of executors, worker nodes, memory allocation, cache memory, worker cores, executor cores everything can be configured according to the system requirements. Apache spark provides various other functionalities like:

2.2.1 Spark Streaming

Spark Streaming is an integrated on the package to the core Spark API which is scalable, high throughput and withal fault-tolerant. It provides the functionality of processing live data streams. For streaming, Apache spark can have flume, HDFS, apache Kafka, twitter, kinesis data sources. This data can be then cleaned and structured in spark itself and used to do further processing.



Figure 3: Spark Streaming

2.2.2 Spark SQL

Spark core has an SQL extension that fortifies more optimization on datasets (RDD) and is in a structured format to retrieve data utilizing the SQL queries. Spark SQL provides the most convenient way to perform several transitions on the data. Spark SQL uses data frames for data manipulations.

2.2.3 Spark ML Lib (Machine learning libraries)

Spark ML Lib provides a wide range of advanced machine learning libraries. Spark has two kinds of libraries that perform operations on RDDs and Data Frames. Spark can process a substantial amount of data and perform advanced machine learning algorithms on it. Spark fortifies clustering, relegation, abbreviation, regression, etc.

3 Python Sentiment Analysis

3.1 Sentiment Analysis

In Natural Language Processing there is a concept kenned as Sentiment Analysis. Given a movie review or a tweet, it can be automatically relegated in categories.

These categories can be utilizer defined (positive, negative) or whichever classes you want.

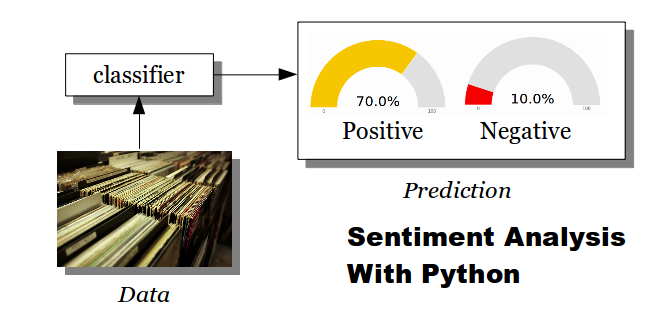
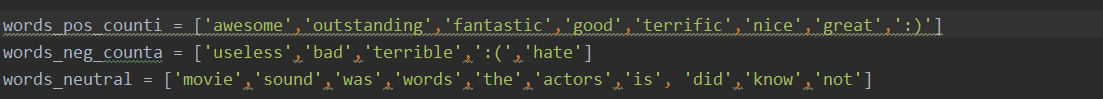


Figure 4: Sentiment Analysis flow

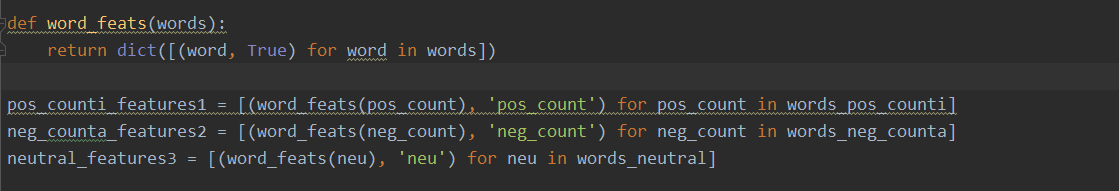
Classification is done using several steps: training and prediction.

The training phase needs to have training data, this is example data in which we define examples. The classifier will use the training data to make predictions.

We start by defining 3 classes: positive, negative and neutral.  
Each of these is defined by a vocabulary:



Every word is changed into a feature using a simplified bag of words model:



Our training set is then the sum of these three feature sets:

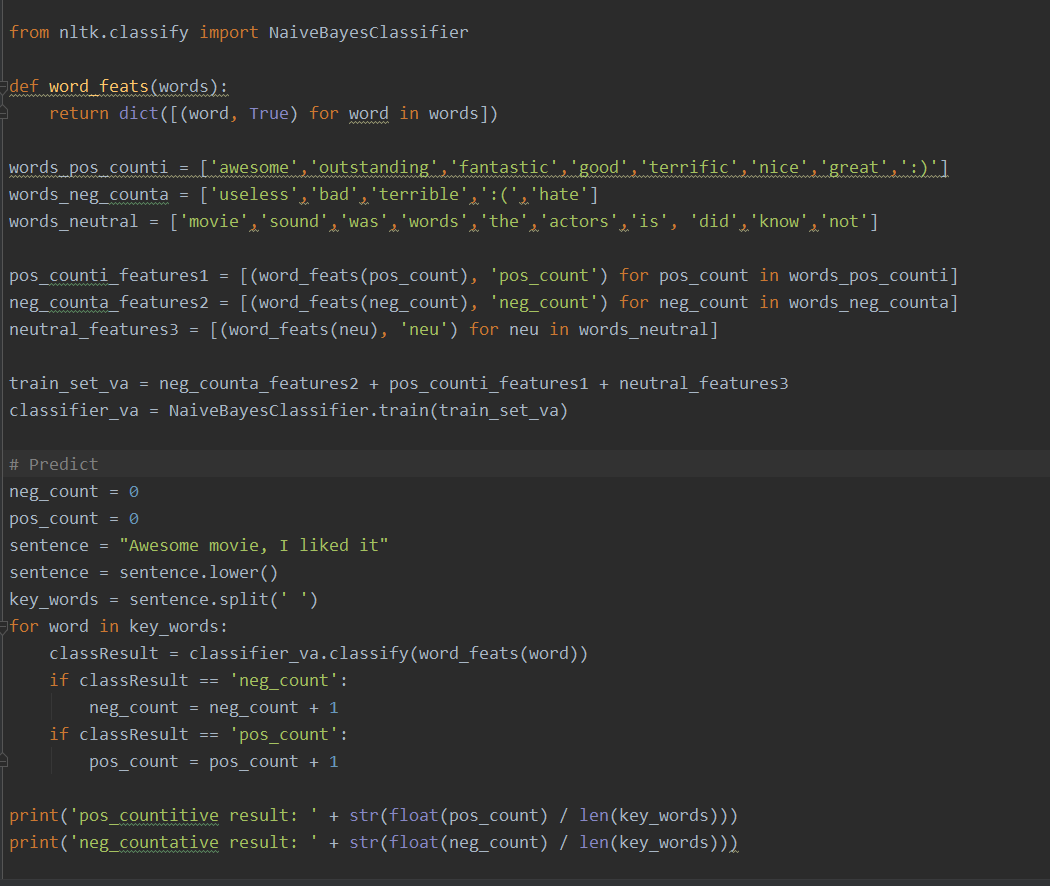


I train the classifier:

  
And make the predictions

The better your training data is, the more accurate your predictions.

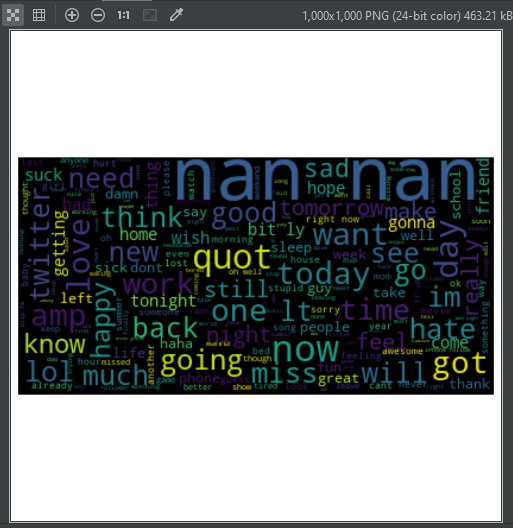
I showed the code below:

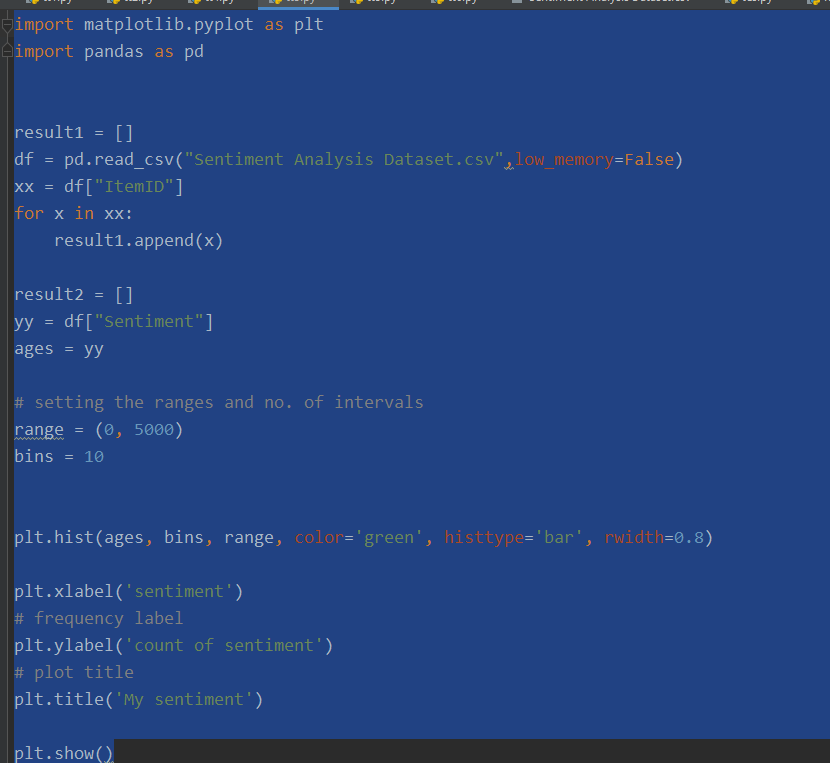


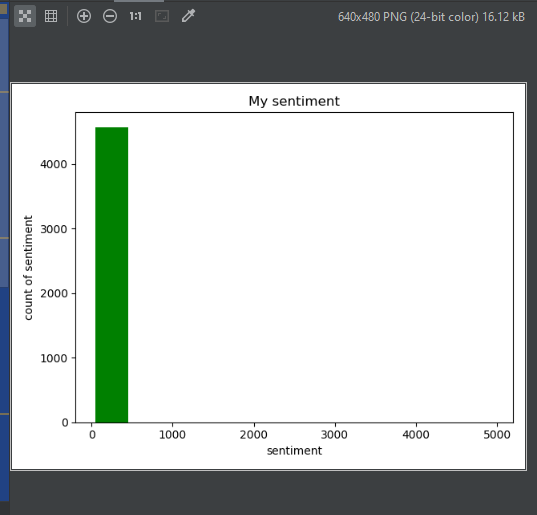
The results:



import tweepy  
import numpy as np  
import pandas as pd  
import cufflinks as cf  
cf.go\_offline()  
cf.set\_config\_file(offline=False, world\_readable=True)  
import matplotlib.pyplot as plt  
  
from wordcloud import WordCloud, STOPWORDS  
  
  
df = pd.read\_csv("Sentiment Analysis Dataset.csv", low\_memory=False)  
  
all\_tweets = ' '.join(str(tweet) for tweet in df['SentimentText'])  
wordcloud = WordCloud(stopwords=STOPWORDS).generate(all\_tweets)  
  
plt.figure(figsize = (10,10))  
plt.imshow(wordcloud, interpolation='bilinear')  
plt.axis("off")  
plt.show()







3.2 Flow Diagram

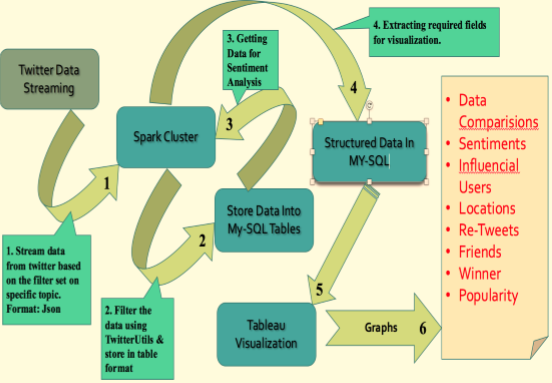


Figure 5: Flow Diagram

Chapter 4

IMPLEMENTATION

4.1 Streaming Twitter Data Using Spark

In my project, Twitter Sentiment Analysis Application is developed using Pyspark which is a combination of Apache Spark and Python. This application fetches Twitter data in a live stream and classifies tweets into positive and negative categories. For the sentiment classification of tweets, the machine learning model (Voting Mechanism) has been developed. Spark’s ability to perform well on iterative algorithms makes it ideal for implementing machine learning techniques as, at their vast majority, machine learning algorithms are based on iterative jobs. Further, live visualization of results is done using Flask and Chart.js technology. Visualization gives the ability to combine data in order to create new insight. I used the python 2.7 version.

We utilize spark streaming to stream live spark data. To load Twitter data into Apache Spark twitter provides an interface to developers that can be habituated to access twitter data. Visit the Twitter applications site to register your application “https://apps.twitter.com/” . I have noted twitter tokens into TwitterKeys.txt which are needed to initialize spark streaming context.

The streaming data is captured into batches. The interval for downloading batches can be set, I have set the interval to be 3 seconds. Consequently, every 3 seconds an incipient batch of data is streamed and captured into batches.

We set the filter which contains a set of keywords to filter the tweets. Only tweets with those keywords present in it would be streamed.

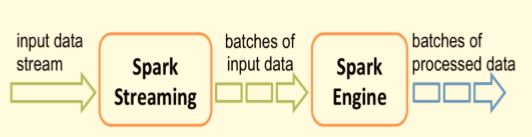
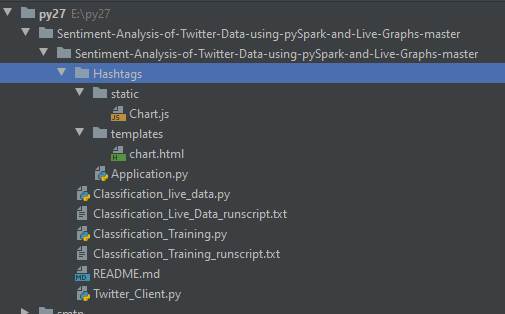


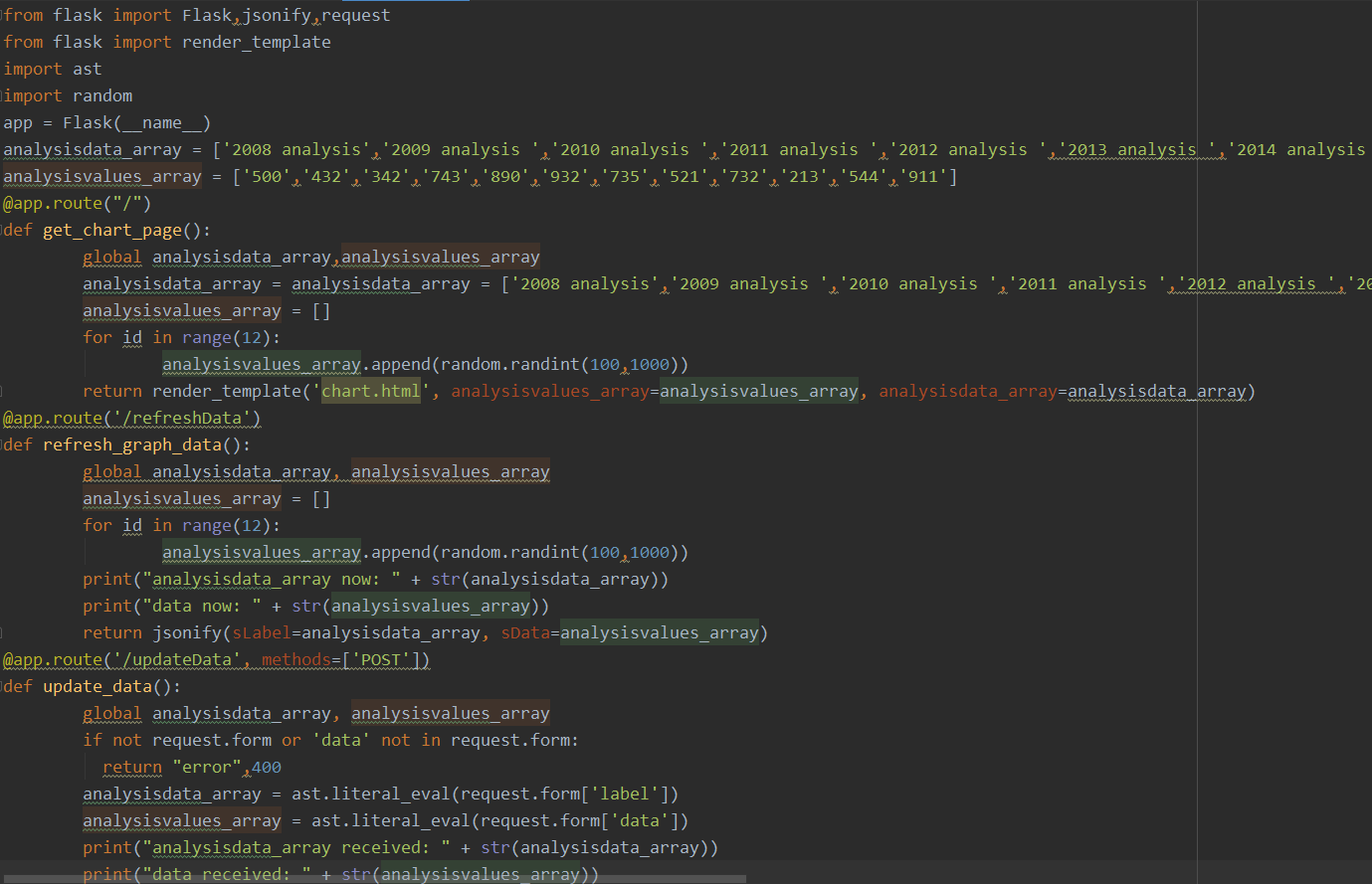
Figure 6: Spark Streaming

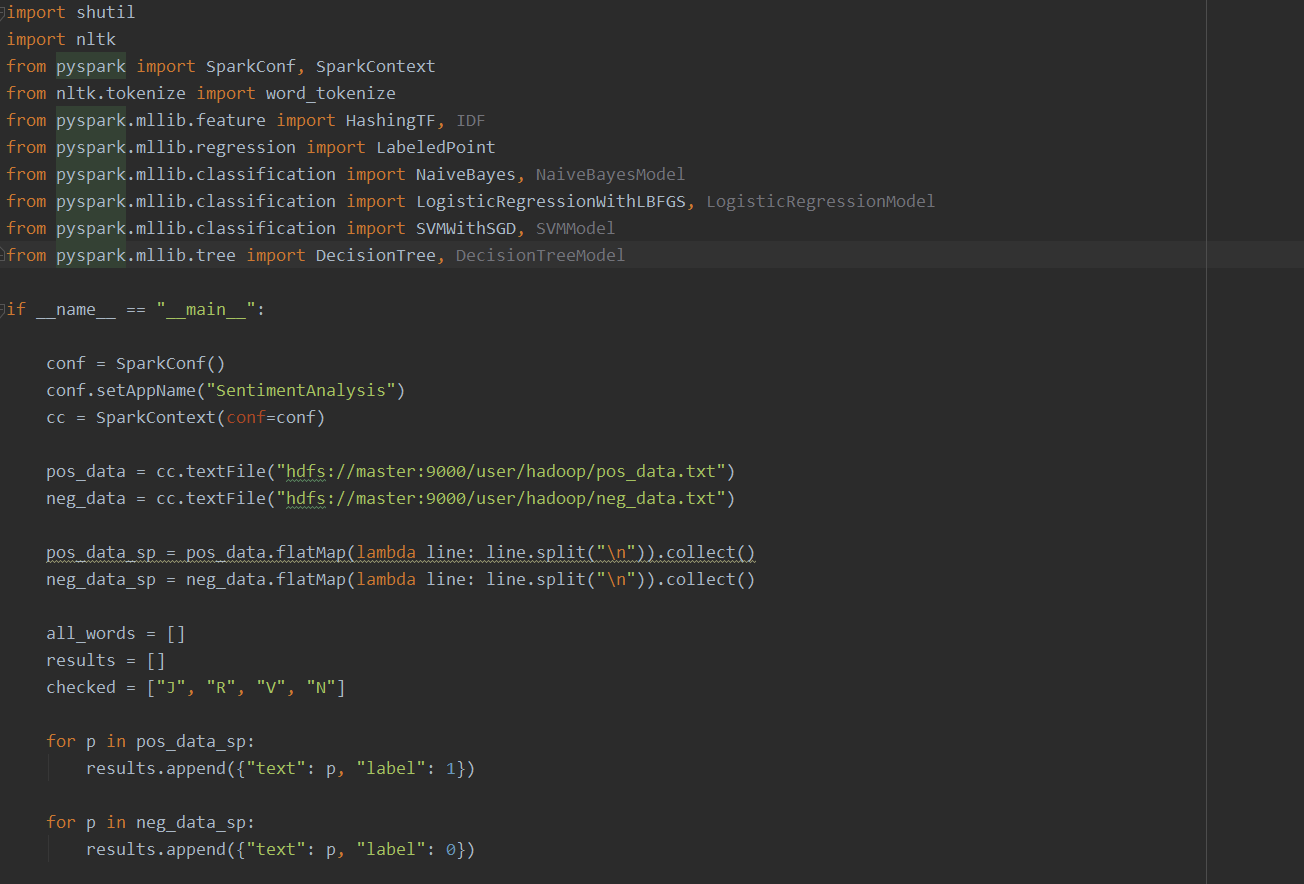
4.2 Understanding Data

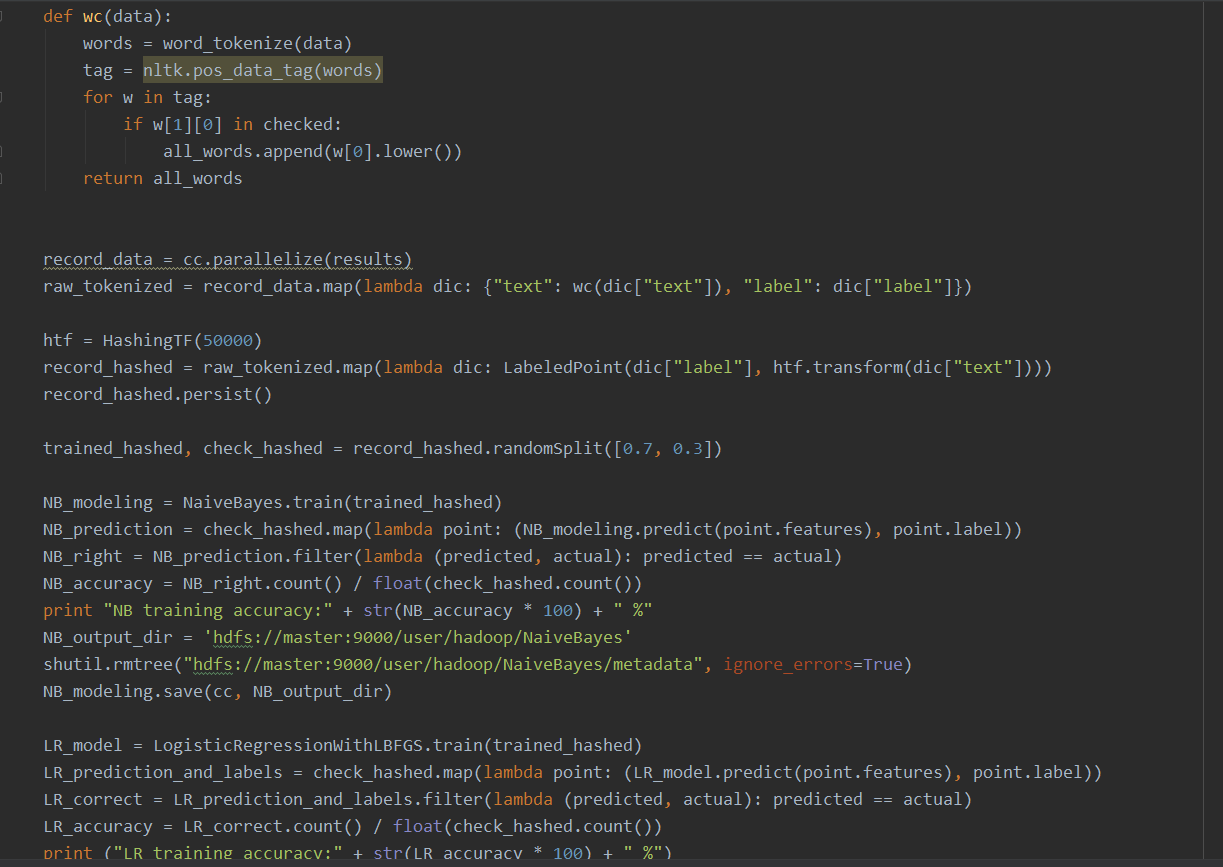
The far most and initial stage in data mining is to perceive and apperceive the quandary and to note down the objectives [14]. We have equipped ourselves with domain cognizance to apperceive the quandary complication, it will tremendously amend data mining efficacy and potency. We have exhaustively understood the sources and types of cognate data and we have amassed, accumulated some utilizable data. We are fixating on finding the right insights from data. The data we receive from twitter streaming is in JSON format. This is program structure.



Program structure







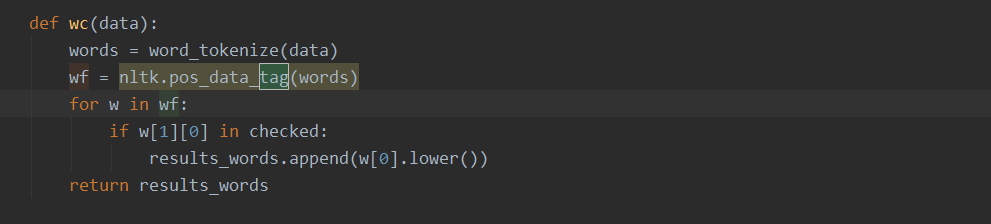
4.3 Pre- Processing Data

The amassed data were strepitous, missing subsidiary info and inconsistently erratic [14]. We have virtually culminated in the Data preparation process. In this process, we have to check if there are vacuous values or inconsistencies in the data. The data should be in a consistent state to be analyzed. To amend the efficiency of our analysis the data should be in a simple format. Data mining is done on this data so to get efficient results the data must be processed by abstracting redundancies. The data is made consequential by deriving information from it like deriving dates out of months, days and years. For example, Deriving the age of the tweet by its date.

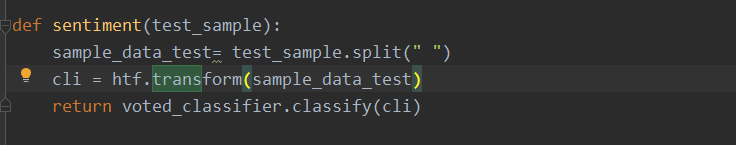
A schema of organized data can be optically discerned as follows

4.4 Algorithm and Sentiment Analysis

Sentiment analysis is performed utilizing Stanford’s Natural Language Processing library [15]. It takes text input and sends this library to get the sentiment in reciprocation. This library constructs a tree-like structure out of the plain text passed to it, this structured is engendered after cleaning the data and abstracting all the cessation words.



Calling function Sentiment sentiment() with parameters as plain text from the tweet. The returned value is preserved into the map so that the status id tagged with its sentiment and can be used further. This is the code for filtering the



After we filter the code and get the plain text we pass it to the function which creates a

tree out of it for getting the sentiment score. The function is defined as follows:

The score is counted by taking calculating the average i.e. by dividing the sum of score by the size of sentence as each sub tree and each word has a score associated to it.

The score is then used to get the weighted sentiment out of the status and then mapped to the sentiment predicated on the score. Here we have verbally expressed that 0 score is very negative, 1 is associated with negative, 2 is neutral, 3 is positive and 4 is very positive.

4.4 Developed Model

The model should be reviewed and checked before utilizing it for obtaining results. The results obtained after mining the data should be analyzed meticulously and interpreted by experts in order to perform efficient data analysis.

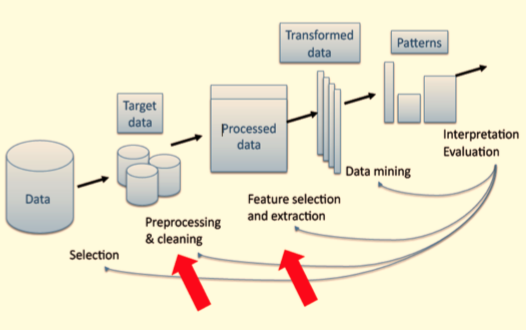


Figure 10: Basic Workflow Model

The above figure describes the steps involved in our project. Fundamental flow diagram of how we utilize our raw data to extract germane information. The extracted information is preserved into case classes and then preserved into data frames. The data frame is engendered utilizing the structure we require.

Constructing DataFrames from schema string. The data types of each column need to be designated while engendering the schema for data frame.

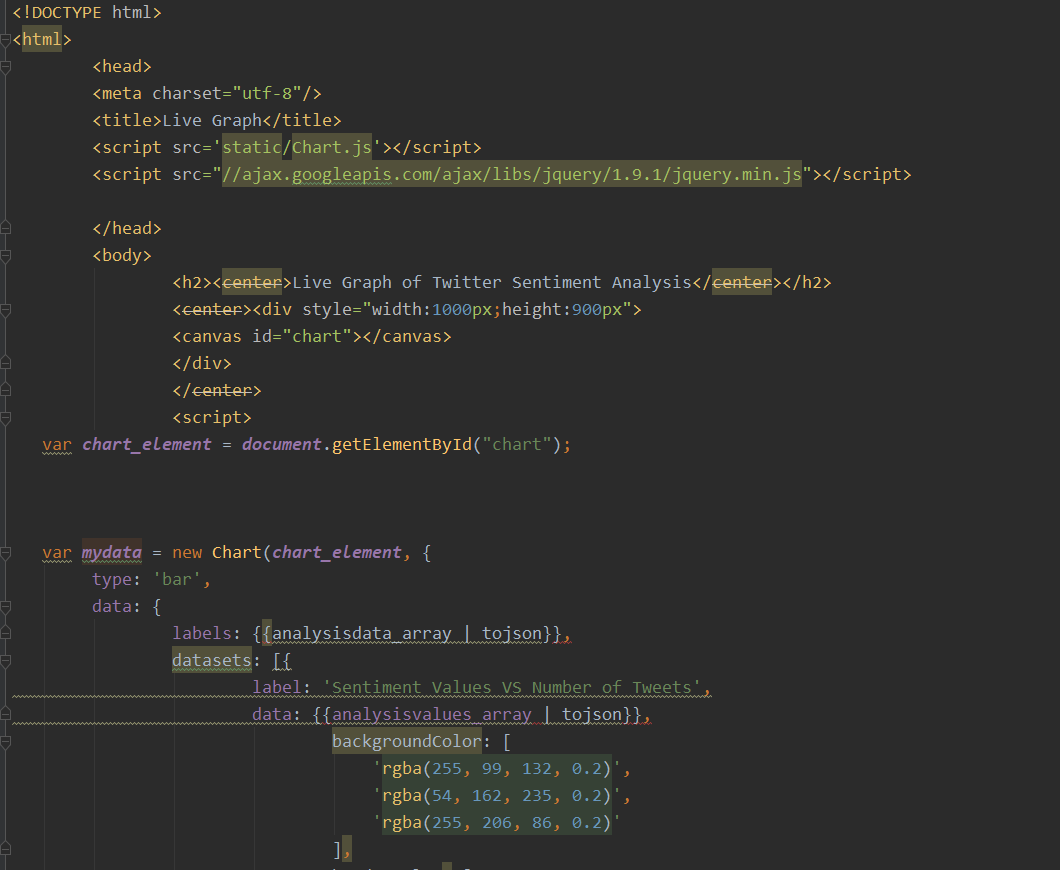
A dataframe looks akin to a table. When we utilize show () method on a dataframe it looks homogeneous too.

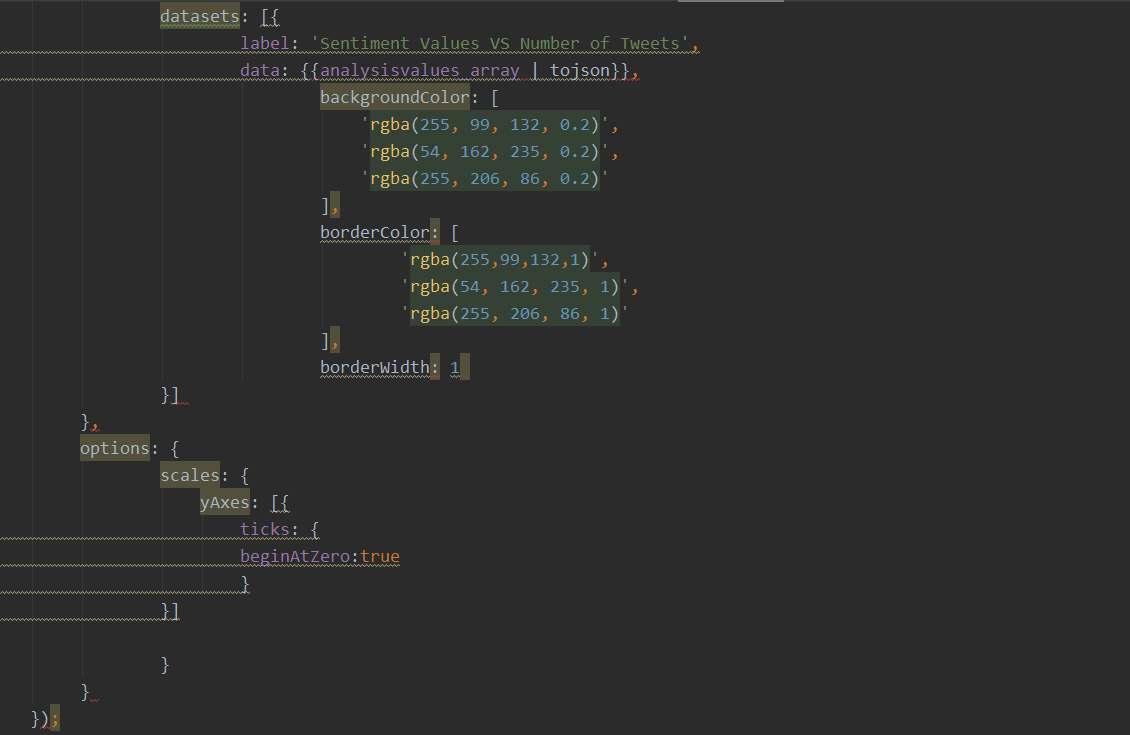
Chapter 5

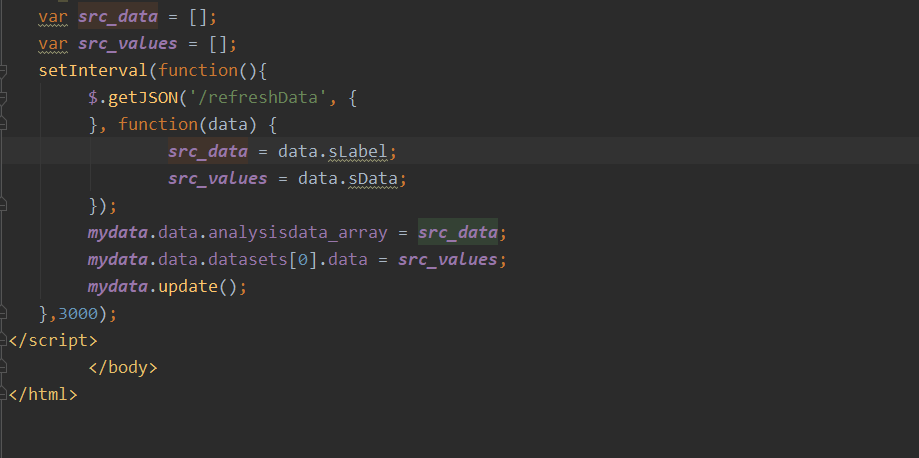
EXECUTION

5.1 Building Project

Launching the project on the spark cluster needs Application.py. Application.py file is the entry in the my application. And then I am showing using the cahrt.html file.





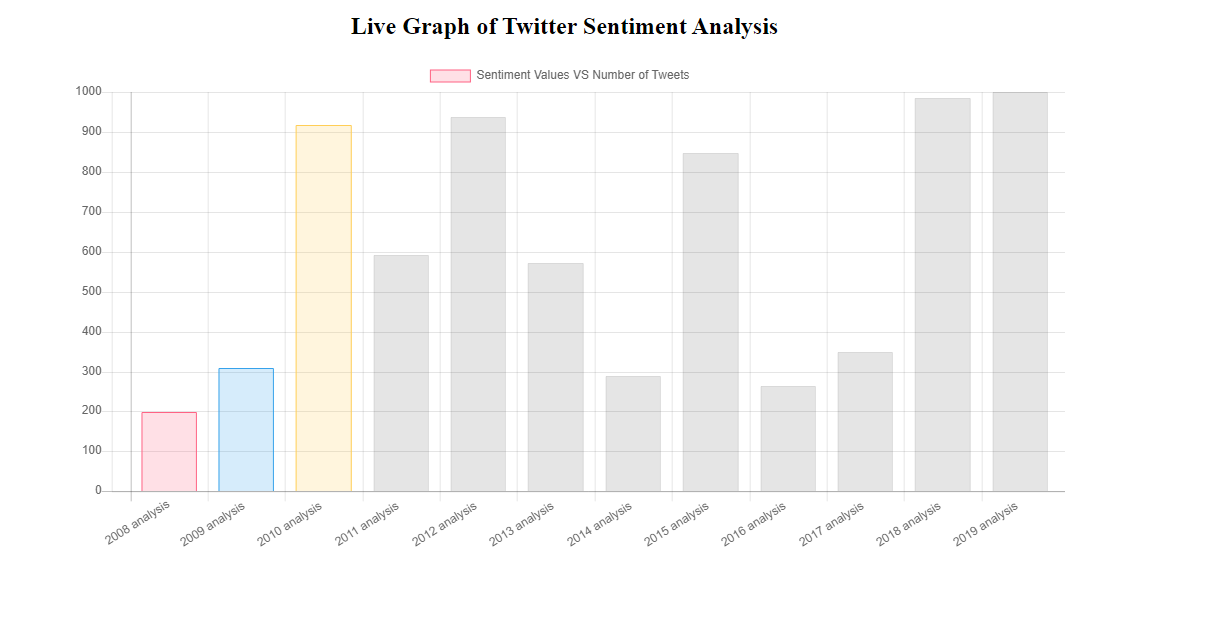


5.2 Submitting Job

Spark cluster needs to be started before submitting any job to the cluster.  For sentiment classification of tweets, machine learning model (Voting Mechanism) has been developed. Spark’s ability to perform well on iterative algorithms makes it ideal for implementing machine learning techniques as, at their vast majority, machine learning algorithms are based on iterative jobs. Further, live visualization of results is done using Flask and Chart.js technology.

Chapter 6

USE-CASES AND RESULTS



The resulting graph

Chapter 7

FUTURE WORK

The world is currently moving towards mobile computing. Mobile applications have become a consequential factor in the prosperity of a product. It’s additionally very convenient to access mobile applications on the Go. In the future, a mobile application could be developed which would give instant reviews as anon as you set the filers as you operate. The task for mobile computing would be the optimization of the cluster on such a diminutive contrivance with minimal hardware support and visualization.

Chapter 8

CONCLUSION

I have learned a plethora of technologies while developing this project. The problems I faced made me think deeper and out of the box. The vibrant and multidimensional visualization would avail consumers cull a better product and avail the benefits of authentic-time review system predicated on location, public view, and trend. It will additionally benefit businesses to get feedback predicated on it there would be immensely colossal scope for development of their product or accommodation. During the course of implementation, I have gained in-depth erudition in the field of immensely colossal data analytics. I got to learn concepts like cluster computing on Apache spark which is a very puissant platform and a trending platform for cluster computing, a language like Scala which is object-oriented as well as functional language provides a lot more flexibility for inscribing optimized code in more diminutive number of lines and has great capability with the Spark cluster. Implements like SBT and Tableau are utilized widely across technological companies.

At last but not least is I have learned how to surmount quandaries and develop a product that would avail people in their lives. The experience has made me technically adept and developed my celebrating power around the quandaries.

# references

[1] Forbes. 12 Big Data Definitions. [Online]. Available:

http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats

yours/#6490f23421a9. Accessed in September 2014. [2] Wikipedia. Big Data. [Online]. Available:

https://en.wikipedia.org/wiki/Big\_data. Accessed in November 2013. [3] Research Gate. The five V’s of Big Data. [Online]. Available:

https://www.researchgate.net/figure/281404634\_fig1\_Figure-1-The-five-V's-of

Big-Data-Adapted-from-IBM-big-data-platform-Bringing-big. Accessed in

September 2015. [4] Slide Share. Big Data Characteristics. [Online]. Available:

http://www.slideshare.net/venturehire/what-is-big-data-and-its-characteristics.

Accessed in July 2013. [5] Wikipedia. Twitter. [Online]. Available: https://en.wikipedia.org/wiki/Twitter.

Accessed in November 2012. [6] Spark. Spark Overview. [Online]. Available: http://spark.apache.org/docs/2.0.1/.

Accessed in October 2013. [7] Spark. Cluster Mode Overview. [Online]. Available:

http://spark.apache.org/docs/latest/cluster-overview.html. Accessed in October

2013.

systems.readthedocs.io/en/latest/datamining.html. Accessed in March 2013.

[8] Spark. Spark Streaming Programming Guide. [Online]. Available:

http://spark.apache.org/docs/2.0.1/streaming-programming-guide.html. Accessed

in October 2013. [9] Spark. Spark SQL, DataFrames and Datasets Guide. [Online]. Available:

http://spark.apache.org/docs/2.0.1/sql-programming-guide.html. Accessed in

October 2013. [10] Spark. Spark SQL, DataFrames and Datasets Guide. [Online]. Available:

http://spark.apache.org/docs/2.0.1/sql-programming-guide.html. Accessed in

October 2013. [11] Wikipedia. SBT. [Online]. Available:

https://en.wikipedia.org/wiki/SBT\_(software). Accessed in September 2013. [12] Wikipedia. Tableau. [Online]. Available:

https://spring.io/guides/gs/register-twitter-app/. Accessed in May 2014. [14] Big Sky. The Data Analysis Process. [Online]. Available:

http://www.bigskyassociates.com/blog/bid/372186/The-Data-Analysis-Process

5-Steps-To-Better-Decision-Making. Accessed in March 2013. [15] Stanford CoreNLP. A Suite of core NLP Tools. [Online]. Available:

http://stanfordnlp.github.io/CoreNLP/. Accessed in August 2014. [16] Recommender Systems. Data Mining. [Online]. Available: http://recommender

https://en.wikipedia.org/wiki/Tableau\_Software. Accessed in April 2013. [13] Spring by Pivotal. Registering an application with Twitter. [Online]. Available: