ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

SOCIAL NETWORKSBASED DISEASE ANALYSIS USING HADOOP

Anbarasi M*, SaleemDurai M. A
School of Computer Science and Engineering, VIT University, Vellore, Tamilnadu, India.
Email: manbarasi@vit.ac.in

Received on 25-07-2016 *Accepted on 28-08-2016*

Abstract

The problem the world is facing today is the large amount of data i.e. the data is increasing each and every day in an enormous amount. The growth is so large that the data which has been found over last year i.e. approximately 5 zeta bytes can be seen today within two or three days. The world not only facing the problem of storing of this massive amount of data but also the methods to access and analyse this large amount of data in specified amount of time. So to overcome this problem a concept has been evolved over past few years in technological field i.e. “Big Data”. The most powerful platform responsible for this massive growth of data is “Social Networks”. A large amount of data is generated on Social Networks in a day and that can be structured, unstructured and semi-structured and is one of the most widely used platforms for everyone. There are various implementations to solve the Big Data problems and among them one of the most popular and widely used implementation is using the Hadoop and Map Reduce programming. This paper aims to analyse social media (Twitter,etc.) big data to identify the widespread of certain keywords related to different diseases at various locations such as India, TamilNadu and World. It will be used to identify the number of people who are using these keywords in a social network classified by the location of the people. It is going to fetch social media data set using Social Networks API and data set is stored in HDFS and a Map Reduce program is to be written to obtain the number of times a disease keyword is used by a certain number of people on their geographic location and finally visualize the results using graph. Also, classification techniques such as Naïve Bayesian classification can be applied to the categorical data.

Keywords: Hadoop, API, Map Reduce, Big Data, Social Network (Twitter, Facebook), Diseases.

1. Introduction

Big data, a concept which is most popular in this tech era is created from a variety of sources such as Internet clicks, mobile transactions, user-generated data, especially in the social media and also the content which is generated through the bank transactions and market demands. Apart from these sectors such as health care, engineering, and many more add to big data environment. The data obtaineed from various sources requires the use of powerful computational tools or applications which can view trends within or in between these large datasets. Data anlysis and new insights created from this data information can be very beneficial in stastical analysis, surveys or the sources where we can find an enormous amount of data , which can add a value to the data obtained in real time and also narrowing the gap between the useful information and time.

In current scenario the technological improvements have introduced a very huge amount of data from distictive domains such as social media, health care, user generated content over internet, etc.The term "Big Data" was coined when scientists saw a massive data over social media such as Twitter. The definition of this huge volume of data over Twitter or in general is much different than the traditional form of data. It has unique characteristics over traditional form of data. Big data can be structured, semi structured, or unstructured thus requires perceptive techniques for analysis. The advances in new technology requires creation of new systems and architectures for the collection, storage and processing of data and algorithm mechanisms to be carried out during the whole process, this is where Hadoop software framework comes under use. Twitter is an online social media network where anyone can register and can do share or post a message which is 140 characters or less called tweets. It is a platform where people from all over the world commute and stay connected to each other by tweeting or exchanging the latest and frequent messages. People follow others from all over the world and get themselves updated known as followers. Furthermore, the Tweets have been created on day to day basis of conversations, news, comments on some events, movies, politics, life, catastrophic events, wars, diseases. Recently there is massive increase in volume of data that the existing technologies will be inefficient to handle this growth. The daily tweets accumulate in huge volume of data, if this kind of data is analyzed in sophisticated way, it can be of great use for us and provides various useful information for decision making. This paper aims to present a Java-Hadoop application developed for big data analytics. The definition of the big data is first introduced followed by a discussion of some attributes regarding big data analytics. A systematic framework is presented to extract important information from big dataset using Hadoop Java Eclipse application on an Oracle Linux OS. The rest of the paper is organized as the follows: Section (II) presents the literature review related to the big data analysis past researchers. Section (III) contains the proposed framework, while the results, and the discussion are in the subject of Section (IV). Finally, section (V) concludes the paper and directions for further work.

2. Literature Review

Manoj Kumar Danthala (February, 2015) published a paper which discusses the term ‘Big Data’ and how it can be used for social media data. In the paper he has discussed that ‘BIG DATA’ has been getting much importance in different industries over the last year or two, on a scale that has generated lots of data every day. Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a reasonable amount of time. It has tremendous potential to transform business and power in several ways. Here the challenge is not only storing the data, but also accessing and analyzing the required data in specified amount of time. He has argued about the massive increase in amount of data that is being generated since 2005 and how the term big data evolved after that. The term big data refers to the data that is generating around us everyday life. It is generally exceeds the capacity of normal conventional traditional databases. For example by combining a large number of signals from the user’s actions and those of their friends, Facebook developed the large network area to the users to share their views, ideas and lot many things. The major characteristics and challenges of big data are Volume, Velocity and Variety. These are called as 3V’s of big data which are used to characterize different aspects of big data. The “Velocity” is defined as the speed at which the data is created, modified, retrieved and processed. This is one of the major challenges of big data because the amount of time for processing and performing different operations are considered a lot when dealing with massive amounts of data. Traditional databases are incapabable of processing this data and hence another major challenge is the volume. The second challenge is the “Volume” i.e amount of data which is processed. There are many business areas where we are uploading and processing the data in very high rate in terms of terabytes and zeta bytes. If we take present statistics, every day we are uploading around 25 terabytes of data into facebook, 12 terabytes of data in twitter and around 10 terabytes of data from various devices like RFID, telecommunications and networking. The storing process requires large clusters and large servers with high bandwidth. Here the problem is not of storing but also the processing which is a major issue nowadays in organisations. Third callenge is the “Variety” i.e. the data can be of any form like structured, semi-structured or unstructured. The data generated from Facebook or Twitter gives ypu the information about the sentiment analysis while the sensory data will give you information about how a product is used and what are the mistakes. So this is the major issue to process information from different sets of data. Big data “Veracity” refers to the biases, noise and abnormally in data. It is the data that is being stored, and mined meaningful to the problem being analysed. In other words, Veracity can be treated as the uncertainty of data due to data inconsistency and incompleteness, ambiguities, latency, model approximations in the process of analysing data. After that he discussed about Hadoop and Map Reduce an open source framework for solving the big data challenges.One of the popular implementation to solve the above challenges of big data is using Hadoop. Hadoop is well-known open-source implementation of the MapReduce programming model for processing big data in parallel of data-intensive jobs on clusters of commodity servers. It is highly scalable compute platform. Hadoop enables users to store and process bulk amount which is not possible while using less scalable techniques. Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. Also in the paper he provides a way of analyzing big data such as twitter data using Apache Hadoop which will process and analyze the tweets on Hadoop clusters. This also includes visualizing the results into a pictorial representations of twitter users and their tweets.The paper proposes three modules for finding and performing operation on social media data sets.The main scope of the paper is to analyzing and fetching the Twitter IDs of those users whose statuses have been retweeted the most by the user whose tweets are being analyzed. First the system involves collecting the tweets from the social network using the twitter API's. Then second, this consists of standard platform as Hadoop to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling. And finally includes analysing the collected tweets and fetching the Twitter IDs of those users whose statuses have been retweeted the most by the user whose tweets are being analysed[1].Sultan Alenezi and Saleh Mesbah(October, 2015) published a paper which developed a Java-Hadoop application. The paper aims to analyze social media big data to identify the widespread of certain keywords. A Java-Hadoop application is developed to analyze data obtained from Twitter social network. The application is used to identify number of people (Tweets) who mentioned specific medical keywords (e.g. Cancer) classified by location. The application is developed using Java Eclipse on CentOS Linux operating System and runs on Oracle Virtual Machine. The analysis aims to help in decision making according to the number of people tweeting about cancer or any related word (like tumor) and analyze them according to their cities all over the world. The results are used to create a GIS layer to spatially enhance the visualization of the obtained results [2]. Manoj Kumar Danthala andDr. Siddhartha Ghosh(May, 2015) published a paper for Twitter analysis of data using Apache Hadoop and Visualizing using BigInsights.Here in the paper twitter data, which is the largest social networking area where data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphere BigInsights tool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets [3]. IlkyuHa, BonghyunBack and ByoungchulAhn (2015) published a research article on “Map Reduce functions to analyze Sentiment Information from Social Big Data”. In this paper they have discussed about opinion mining and how Hadoop Map Reduce model can be used for this system.Opinion mining which extracts large amount of opinion information from social media data is gaining so much popularity in research area. In particular, opinion mining has been used to understand the true meaning and intent of social networking siteusers. It requires special techniques to collect, store and process of data and extract meaningful information from them. Therefore, in this paper, they propose a method to extract sentiment information from various types of unstructured social media text data from social networks by using a parallel Hadoop Distributed File System(HDFS) to save social multimedia data and using Map Reduce functions for sentiment analysis.The proposed method has stably performed data gathering and data loading and maintained stable load balancing of memory and CPU resources during data processing by the HDFS system [4]. Prachi Gokhale Pradnya Kulkarni University at Buffalo made a paper on“Twitter Data Analysis: Hadoop2 MapReduce Framework”. The paper was aimed to find the IPL trends from Twitter. They used the MapReduce programming model and Hadoop Framework for storing and processing of data. And finally used R studio for visulization using graphs [5].

3. Framework for the Proposed System

3.1 Architecture

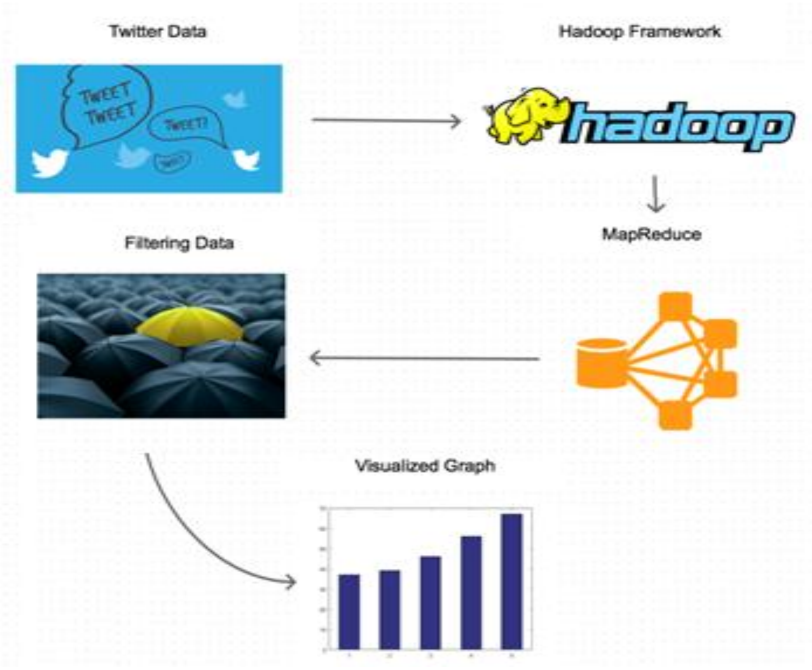


Fig. 1: Proposed System Architecture.

1. Data Sets: First we gather data sets from social network website especially Twitter. We can also get raw data set from snap repository for analysis purpose. But we are fetching dataset from Twitter only. We are using mahout for fetching the Twitter dataset through Eclipse IDE console.
2. UBUNTU: Linux OS i.e. UBUNTU 12.04. is used for implementing the whole paper work as it is open source and its more easier to work in Linux than any other OS like windows for Big Data papers.
3. Hadoop Framework: The Hadoop architecture will play a most important role in as it is the framework which is going to store the whole dataset. Hadoop Framework 2.7.2 supported by Java 1.7.2 version is used. The data sets are stored on HDFS.
4. Map Reduce: The whole processing of data takes place in map reduce programming model of Hadoop. Code is written in Java using Eclipse Luna where all Hadoop libraries are imported and Maven and POM (Paper Object Model) is used for all this java programming purpose.
5. Visualized Graph : A java program with all the libraries required for generating a bar graph are imported and the output is visualized and can be filtered in various ways.

The filtering is done on the basis of disease dataset and the bar graph will be generated for all the common disease which we have provided through raw sample dataset.

3.2 Framework

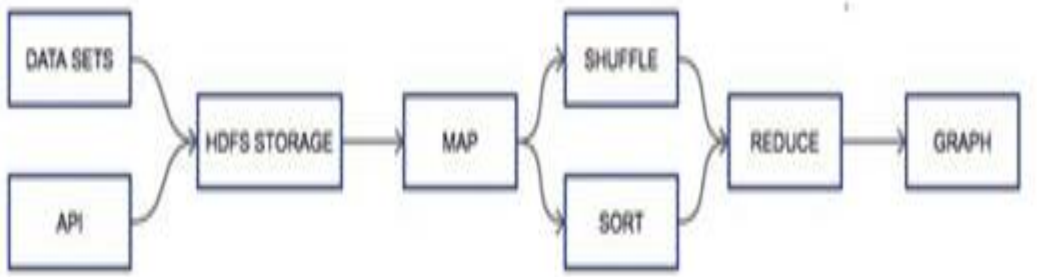


Fig.2: Proposed System Framework.

1. Collecting raw data sets related from Social Networks API or SNAP Stanford University repository.
2. Store the data sets in HDFS (Hadoop Distributed File System).
3. Split the data up and store at the collection of machines known as cluster.
4. Next step is to process the data rather than retrieving data from Node server.
5. Process with Map Reduce programming, where files are broken into chunks and parallel data processing is performed.

6. As the data already owns the cluster, so we can process in that place.
7. Data is mapped and later data is reduced (Map Reduce).
8. Obtained results will be in the form of (Key, Value) pairs.
9. Filter the data on the basis of disease dataset and visualization can be done for different location on the basis of disease and tweets in form of Bar Graph.
10. Compare the different graphs obtained as a final result and the trending disease or the most common disease tweeted by the people can be seen as a core part of the paper.

3.3 Module

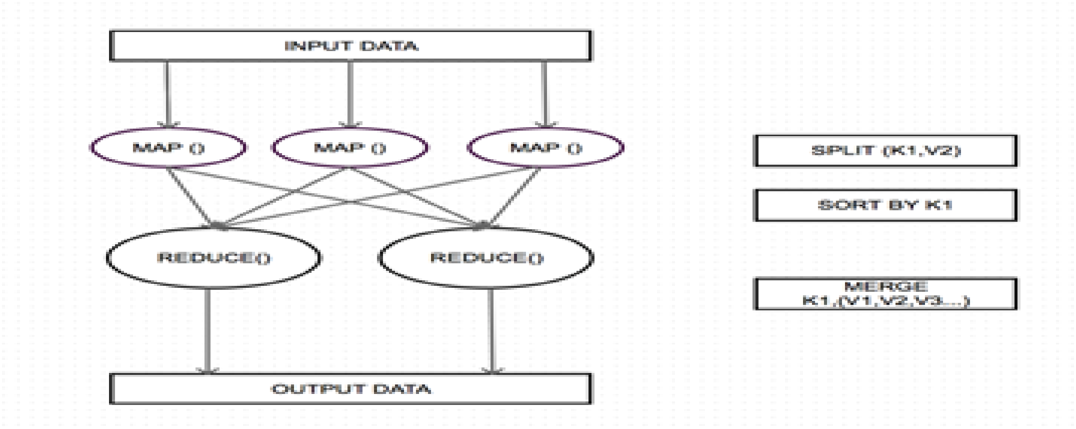


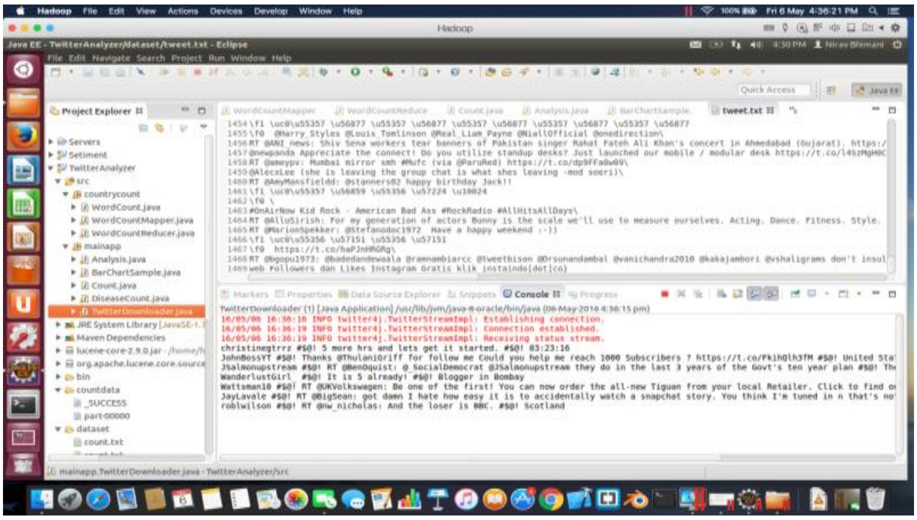
Fig. 3: Map Reduce module.

1. The initial step of map reduce job is to get the input data or raw data where stored data file is broken down in to chunks. Map Reduce processes the data in parallel as it is the core functionality of Map Reduce programming model. In serial processing of data it will take a lot of time so map reduce is preferred.
2. After the input data file, the we will feed data to mapper class which will split it according to key value pair. In our case each word acts as a key and the occurrence of word will act as a value. The mapper program will generate the intermediate keys in the form of key value pairs.
3. The reducer class starts with shuffling and sorting of key values. The reducer program will sort the data according to key. The sorting of data is done in alphabetical order and in ascending order.
4. Now the job of Reducer comes in where the sorted keys with the number of occurrences will be merged together. Here the reducer will merge the key and the number of values it is having of a particular key and group them together.
5. Finally after mapping and reducing, the output file will be generated with each key and its value count in alphabetical order.

6. Now we can use this output file for filtering the key and values which are required by us for the analysis purpose. We have filtered the map reduce output by comparing it with the raw disease dataset.

4. Results & Discussions

1. The initial output we get from fetching data from Twitter is saved in "tweets.txt". The process of fetching data from Indian server took time like 200 tweets in 10 mins, comparing to general server (random fetching from any countries) took 1000 tweets in less than 5 minutes.

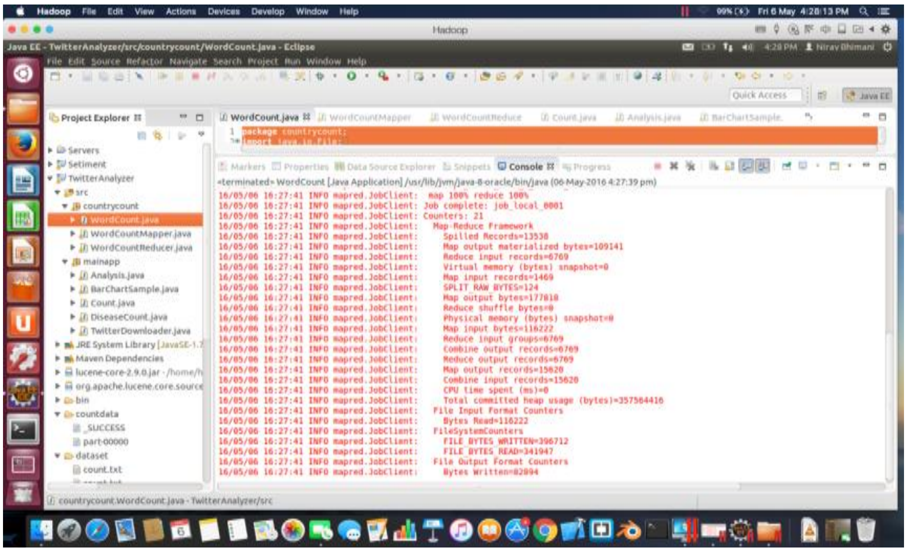


```
\f1 \uc0\u55357 \u56877 \u55357 \u56877 \u55357 \u56877 \u55357 \u56877 \u55357 \u56877
\f0  @Harry_Styles @Louis_Tomlinson @Real_Liam_Payne @NiallOfficial @onedirection\
RT @ANI_news: Shiv Sena workers tear banners of Pakistan singer Rahat Fateh Ali Khan's concert in Ahmedabad (Gujarat). https:/
@newpanda Appreciate the connect! Do you utilize standup desks? Just launched our mobile / modular desk https://t.co/l4SzMgH0C
RT @ameypv: Mumbai mirror smh #Mufc (via @ParuRed) https://t.co/dp9FFa0w09\
@AlecxLee (she is leaving the group chat is what shes leaving -mod soeri)\
RT @AmyMansfieldd: @stanners02 happy birthday Jack!!
\f1 \uc0\u55357 \u56859 \u55356 \u57224 \u10024
\f0 \
#OnAirNow Kid Rock - American Bad Ass #RockRadio #AllHitsAllDays\
RT @AlluSirish: For my generation of actors Bunny is the scale we'll use to measure ourselves. Acting. Dance. Fitness. Style.
RT @MarionSpekker: @Stefanodoc1972  Have a happy weekend :-))
\f1 \uc0\u55356 \u57151 \u55356 \u57151
\f0  https://t.co/haPJnHRORg\
RT @bgopu1973: @badedandewaala @ramnambiarcc @tweetbison @Drsunandambal @vanichandra2010 @kakajambori @vshaligrams don't insul
web Followers dan Likes Instagram Gratis klik instaindo[dot]co)
```

```
TwitterDownloader (1) [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (06-May-2016 4:36:15 pm)
16/05/06 16:36:18 INFO twitter4j.TwitterStreamImpl: Establishing connection.
16/05/06 16:36:19 INFO twitter4j.TwitterStreamImpl: Connection established.
16/05/06 16:36:19 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
christinegtrrz #$@! 5 more hrs and lets get it started. #$@! 03:23:16
JohnBossYT #$@! Thanks @ThulaniGriff for follow me Could you help me reach 1000 Subscribers ? https://t.co/FkihQlh3fM #$@! United Sta
JSalmonupstream #$@! RT @BenOquist: @_SocialDemocrat @JSalmonupstream they do in the last 3 years of the Govt's ten year plan #$@! The
WanderlustGirl  #$@! It is 5 already! #$@! Blogger in Bombay
Wattsman10 #$@! RT @UKVolkswagen: Be one of the first! You can now order the all-new Tiguan from your local Retailer. Click to find o
JayLavale #$@! RT @BigSean: got damn I hate how easy it is to accidentally watch a snapchat story. You think I'm tuned in n that's no
roblwilson #$@! RT @nw_nicholas: And the loser is BBC. #$@! Scotland
```

Fig. 4 : Twitter Data Fetch.

2. Then, comes the MapReduce job where sorting and merging of tweets occurs and data is saved, but the output file will contains all the words occurring in input file (tweets.txt) and count the occurrence.



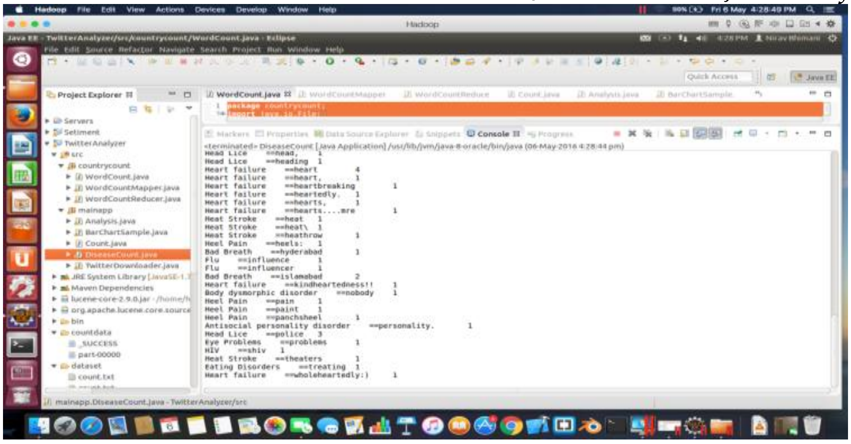
```
package countrycount;
import java.io.File;
```

```
<terminated> WordCount [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (06-May-2016 4:27:39 pm)
16/05/06 16:27:41 INFO mapred.JobClient:  map 100% reduce 100%
16/05/06 16:27:41 INFO mapred.JobClient: Job complete: job_local_0001
16/05/06 16:27:41 INFO mapred.JobClient: Counters: 21
16/05/06 16:27:41 INFO mapred.JobClient:   Map-Reduce Framework
16/05/06 16:27:41 INFO mapred.JobClient:     Spilled Records=13538
16/05/06 16:27:41 INFO mapred.JobClient:     Map output materialized bytes=109141
16/05/06 16:27:41 INFO mapred.JobClient:     Reduce input records=6769
16/05/06 16:27:41 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
16/05/06 16:27:41 INFO mapred.JobClient:     Map input records=1469
16/05/06 16:27:41 INFO mapred.JobClient:     SPLIT_RAW_BYTES=124
16/05/06 16:27:41 INFO mapred.JobClient:     Map output bytes=177818
16/05/06 16:27:41 INFO mapred.JobClient:     Reduce shuffle bytes=0
16/05/06 16:27:41 INFO mapred.JobClient:     Physical memory (bytes) snapshot=0
16/05/06 16:27:41 INFO mapred.JobClient:     Map input bytes=116222
16/05/06 16:27:41 INFO mapred.JobClient:     Reduce input groups=6769
16/05/06 16:27:41 INFO mapred.JobClient:     Combine output records=6769
16/05/06 16:27:41 INFO mapred.JobClient:     Reduce output records=6769
16/05/06 16:27:41 INFO mapred.JobClient:     Map output records=15620
16/05/06 16:27:41 INFO mapred.JobClient:     Combine input records=15620
16/05/06 16:27:41 INFO mapred.JobClient:     CPU time spent (ms)=0
16/05/06 16:27:41 INFO mapred.JobClient:     Total committed heap usage (bytes)=357564416
16/05/06 16:27:41 INFO mapred.JobClient:   File Input Format Counters
16/05/06 16:27:41 INFO mapred.JobClient:     Bytes Read=116222
16/05/06 16:27:41 INFO mapred.JobClient:   FileSystemCounters
16/05/06 16:27:41 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=396712
16/05/06 16:27:41 INFO mapred.JobClient:     FILE_BYTES_READ=341947
16/05/06 16:27:41 INFO mapred.JobClient:   File Output Format Counters
16/05/06 16:27:41 INFO mapred.JobClient:     Bytes Written=82894
```

Hadoop
Java EE - TwitterAnalyzer/src/countrycount/WordCount.java - Eclipse
Project Explorer
WordCount.java
Console

Fig. 5: Map Reduce Job Running.

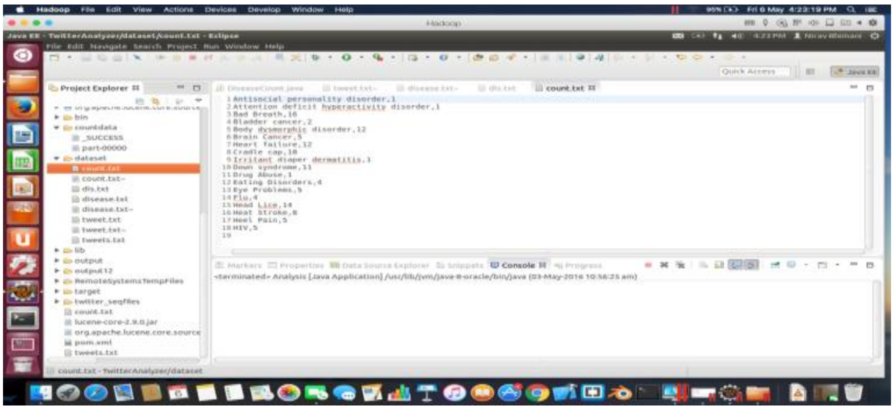
3. Later come the filtering of keywords from the MapReduce output. At this point, we will have input file containing list of Medical disease names and that file will be filtered wrt to MapReduce code which will again generate output of Diseases and number of occurrences.



```
<terminated> DiseaseCount [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (06-May-2016 4:28:44 pm)
Head Lice    ==head,    1
Head Lice    ==heading  1
Heart failure    ==heart        4
Heart failure    ==heart,       1
Heart failure    ==heartbreaking        1
Heart failure    ==heartedly.   1
Heart failure    ==hearts,      1
Heart failure    ==hearts....mre        1
Heat Stroke    ==heat   1
Heat Stroke    ==heat\  1
Heat Stroke    ==heathrow       1
Heel Pain    ==heels:   1
Bad Breath    ==hyderabad       1
Flu    ==influence      1
Flu    ==influencer     1
Bad Breath    ==islamabad       2
Heart failure    ==kindheartedness!!    1
Body dysmorphic disorder    ==nobody    1
Heel Pain    ==pain     1
Heel Pain    ==paint    1
Heel Pain    ==panchsheel       1
Antisocial personality disorder    ==personality.       1
Head Lice    ==police   3
Eye Problems    ==problems      1
HIV    ==shiv   1
Heat Stroke    ==theaters       1
Eating Disorders    ==treating  1
Heart failure    ==wholeheartedly:)     1
```

Fig. 6: Disease Output File.

4. All the Filtered output will be count the occurrences of diseases and group everything.



```
Antisocial personality disorder,1
Attention deficit hyperactivity disorder,1
Bad Breath,16
Bladder cancer,2
Body dysmorphic disorder,12
Brain Cancer,5
Heart failure,12
Cradle cap,10
Irritant diaper dermatitis,1
Down syndrome,11
Drug Abuse,1
Eating Disorders,4
Eye Problems,5
Flu,4
Head Lice,14
Heat Stroke,8
Heel Pain,5
HIV,5
```

Fig. 7: Disease Count Filtered.

5. At the end, the output file is generated to Bar graph to visualise the output in more sensible way. This will provide more insight of data being tweeted on Twitter. The output graph and previous step output can be used in further Sentimental Data Analysis.

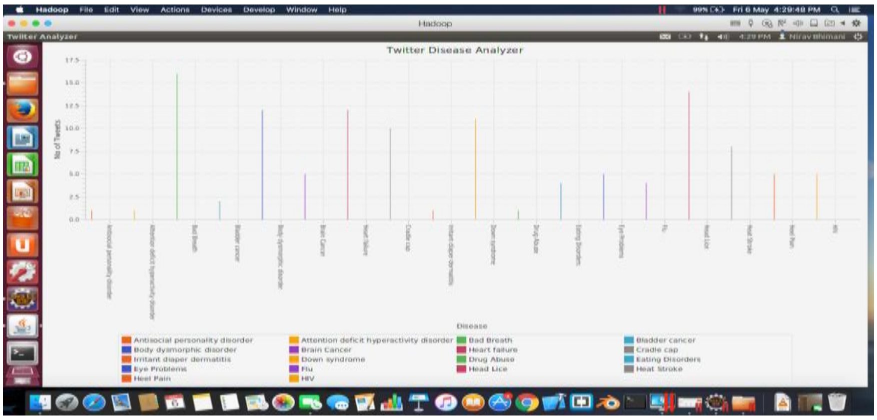


Fig. 8: Bar Graph India.

5. Conclusion

The paper has been very challenging at each stages. Each stages have different problems which need to be overcome. We specify the country and data being being fetched is processed later to generate output. The output received in Bar graph can be used in industry depending upon the filtering done, in our paper we focussed on Medical Disease and Country. This output can help to generate stats for each country or state which can be used in various ways. This may help to point the people getting affected by disease or related to. Also the Hospitals, Government, Medical Industry can gain lot from this data, since people uses SNS 1-2 hrs daily which appears to be good source to collect information about people. Hence, this paper is good approach to Sentimental Analysis in Big Data.

6. References

1. Manoj Kumar Danthala. “Tweet Analysis: Twitter Data processing Using Apache Hadoop.” International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015.
2. Sultan Alenezi and Saleh Mesbah. “Big Data Spatial Analytics in Social Networks using Hadoop.” International Journal of ComputerApplications 128(14):21-26, October 2015. Published by Foundation of Computer Science (FCS), NY, USA.
3. Manoj Kumar Danthala and Dr. Siddhartha Ghosh. “Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights”. International Journal of Engineering and Technical ResearchV4(05) · May 2015.
4. ResearchArticle “MapReduce Functions to Analyze Sentiment Information from Social Big Data” IkyuHa,1 BonghyunBack,2 and ByoungchulAhn2. Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 417502, 11 pages http://dx.doi.org/10.1155/2015/417502.
5. http://docplayer.net/900732-Twitter-data-analysis-hadoop2-map-reduce-framework.html.
6. Mahesh G Huddar; Manjula M Ramannavar, “A Survey on Big Data Analytical Tools”, International Journal of Latest Trends in Engineering and Technology (IJLTET), Special Issue – IDEAS 2013.
7. Mukkamala, R.R.; Hussain, A.; Vatrapu, R., Towards a Set Theoretical Approach to Big Data Analytics, IT University of Copenhagen, 2014 IEEE International Congress on Big Data.
8. http://dev.twitter.com/
9. https://snap.stanford.edu/data/

Corresponding Author:

Anbarasi. M*,

Email: *manbarasi@vit.ac.in*