GLA UNIVERSITY
ESTD. 2010
ऋते ज्ञानान्न मुक्तिः

GLA
UNIVERSITY
MATHURA
Established vide U.P. Act 21 of 2010.

Institute of Engineering & Technology

MID TERM REPORT

On

TWITTER SENTIMENT ANALYSIS

Submitted by

SAHIL SRIVASTAVA
SHIVANK GARG
SAMARTH AHUJA
KANISHK

Department of Computer Engineering & Applications
Institute of Engineering & Technology

GLA University
Mathura- 281406, INDIA

Department of computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,
Mathura – 281406

Declaration

We hereby declare that the work which is being presented in the Mini Project **“Twitter Sentiment Analysis”,** in partial fulfilment of the requirements for Mini Project viva voce, is an authentic record of my own work carried under the supervision of **“VINAY AGRAWAL SIR” .**

Signature of Candidate:

Name of Candidate: SAHIL SRIVASTAVA, KANISHK, SAMARTH AHUJA, SHIVANK GARG

Roll. No.: 171500281, 171500153, 171500286, 171500325

Course: B. TECH CSE

Year: 3rd Year
Semester: Vth Semester

ACKNOWLEDGEMENT

On successful completion of my project thus I would take a privilege to thank the supporting man power of our faculty members, industrial guide and all my friends & colleagues with gratitude. However, we wish to make special mention of the following.

First of all, I'm are thankful of my internal guide **MR. VINAY AGARWAL** and all the friends and mentor of CEA department under whose guideline I was able to complete my project up to valuable mark. I'm wholeheartedly thankful to them for giving me their value able time and attention, and for providing me a systematic way for completing my project in time.

SAHIL SRIVASTAVA

SAMARTH AHUJA

SHIVANK GARG

KANISHK

ABSTRACT

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools like Brandwatch Analytics make that process quicker and easier than ever before, thanks to real-time monitoring capabilities. Sentiment Analysis also helps organisations measure the ROI of their marketing campaigns and improve their customer service. Since sentiment analysis gives the organisations a sneak peek into their customer's emotions, they can be aware of any crisis that's to come well in time – and manage it accordingly. Sentiment analysis helps you complete your market research by getting to know what your customers' opinions are about your products/services and how you can align your products/services' quality and features with their tastes.

Table of Contents

PROBLEM STATEMENT

The problem in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature.

Whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral.

OBJECTIVE

In this project, we will showcase how to perform Sentiment Analysis on Twitter data using Pig. To begin with, we will be collecting real-time tweets from Twitter using Flume. With the help of AFINN dictionary we can find positive sentiment from the data which we have dumped into our HDFS/FLUME. We can prioritize our own customized words in AFINN dictionary as per the need of Client.

As this project is most advance use case of Hadoop in MNC & has wide impact in sentiment analysis. We need most advance concepts of PIG & MAP-REDUCE for getting into this project.

The data from Twitter is in ‘Json’ format, so a Pig JsonLoader is required to load the data into Pig. You need to download the required jars for the JsonLoader.

PROJECT FUNCTIONALITIES

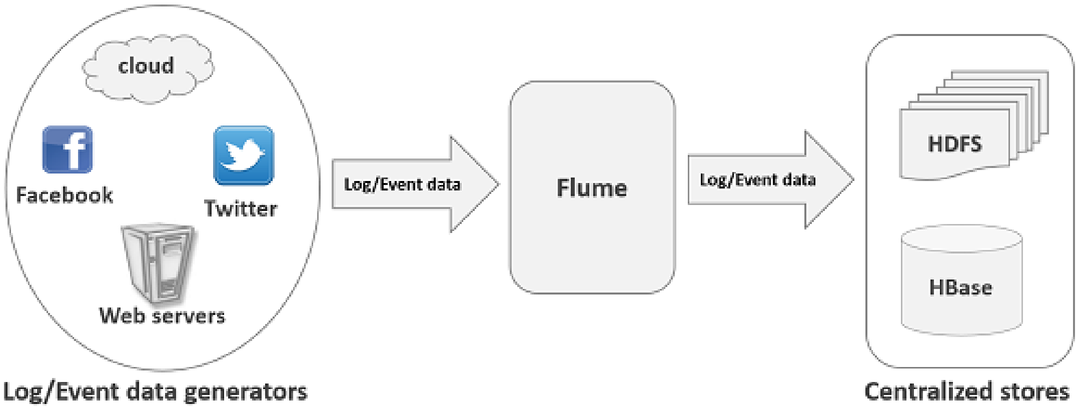
The functionalities of the proposed system are very user friendly and attractive. Some of the functionalities are as follows:

- Add category
- Add keyword
- Select File
- Collect tweets
- Pre-process tweets
- Analysis
- Web Application
- Providing the resultant tweets
- Providing suggestions

INTRODUCTION TO APACHE FLUME

What is flume?

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.
Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.



Applications of Flume

Assume an e-commerce web application wants to analyze the customer behavior from a particular region. To do so, they would need to move the available log data in to Hadoop for analysis. Here, Apache Flume comes to our rescue.
Flume is used to move the log data generated by application servers into HDFS at a higher speed.

Advantages of Flume

1. Using Apache Flume, we can store the data in to any of the centralized stores (HBase, HDFS).

2. Flume is reliable, fault tolerant, scalable, manageable, and customizable.

3. When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data

producers and the centralized stores and provides a steady flow of data between them.

4. Flume provides the feature of contextual routing.

5. The transactions in Flume are channel-based where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.

Features of Flume

1. Flume ingests log data from multiple web servers into a centralized store (HDFS, HBase) efficiently.

2. Using Flume, we can get the data from multiple servers immediately into Hadoop.

3. Along with the log files, Flume is also used to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart.

4. Flume supports a large set of sources and destinations types.

5. Flume supports multi-hop flows, fan-in fan-out flows, contextual routing, etc.

How Flume helps Hadoop to get data from live streaming?

1. Flume allows the user to do the following:
2. Stream data into Hadoop from multiple sources for analysis.
3. Collect high-volume web logs in real-time.
4. It acts as a buffer when the rate of incoming data exceeds the rate at which the data can be written. Thereby preventing data loss.
5. Guarantees data delivery.
6. Scales horizontally (connects commodity system in parallel) to handle additional data volume.

Essential Components Involved in Getting Data from a Live-Streaming Source

There are 3 major components, namely: Source, Channel, and Sink, which are involved in ingesting data, moving data and storing data, respectively. Below is the breakdown of the parts applicable in this scenario:

1. **Event** – A singular unit of data that is transported by Flume (typically a single log entry).

2. **Source** – The entity through which data enters into the Flume. Sources either actively samples the data or passively waits for data to be delivered to them. A variety of sources such as log4j logs and syslog’s, allows data to be collected.

3. **Sink** – The unit that delivers the data to the destination. A variety of sinks allow data to be streamed to a range of destinations. Example: HDFS sink writes events to the HDFS.

4. **Channel** – It is the connection between the Source and the Sink. The Source ingests Event into the Channel and the Sink drains the Channel.

5. **Agent** – Any physical Java virtual machine running Flume. It is a collection of Sources, Sinks and Channels.

6. **Client** – It produces and transmits the Event to the Source operating within the Agent

INTRODUCTION TO BIG DATA

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

What is Big Data?

Big Data is also data but with a huge size. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short, such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Examples of Big Data

The New York Stock Exchange generates about one terabyte of new trade data per day.

Social media

The statistics shows that 500+ terabytes of new data ingested into the databases of social media site Facebook every day

A single jet engine can generate 10+terabytes of data in 30 minutes of flight time. It turns up to petabytes.

Types of Big Data

1. Structured
2. Unstructured
3. Semi-structured

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes

Ex- An Employee table in the database

Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format
Ex-the output returned by google search

Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.
Ex-personal data stored in xml file

Characteristics of Big Data

(i) ***Volume*** – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **volume** is one characteristic which needs to be considered while dealing with Big Data.

(ii) ***Variety*** **–** The next aspect of Big Data is its **variety**. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(iii) ***Velocity*** **–** The term **velocity** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

(iv) Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensor devices, etc. The flow of data is massive and continuous.

*(v)****Variability*** – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Benefits of Big Data Processing

Ability to process Big Data brings in multiple benefits, such as-
Businesses can utilize outside intelligence while taking decisions
Improved customer services

INTRODUCTION TO HADOOP

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop is a framework which helps in storing and processing huge datasets and Sqoop component is used to transfer files from traditional databases like RDBMS to HDFS and vice versa when the data is of the structured type.

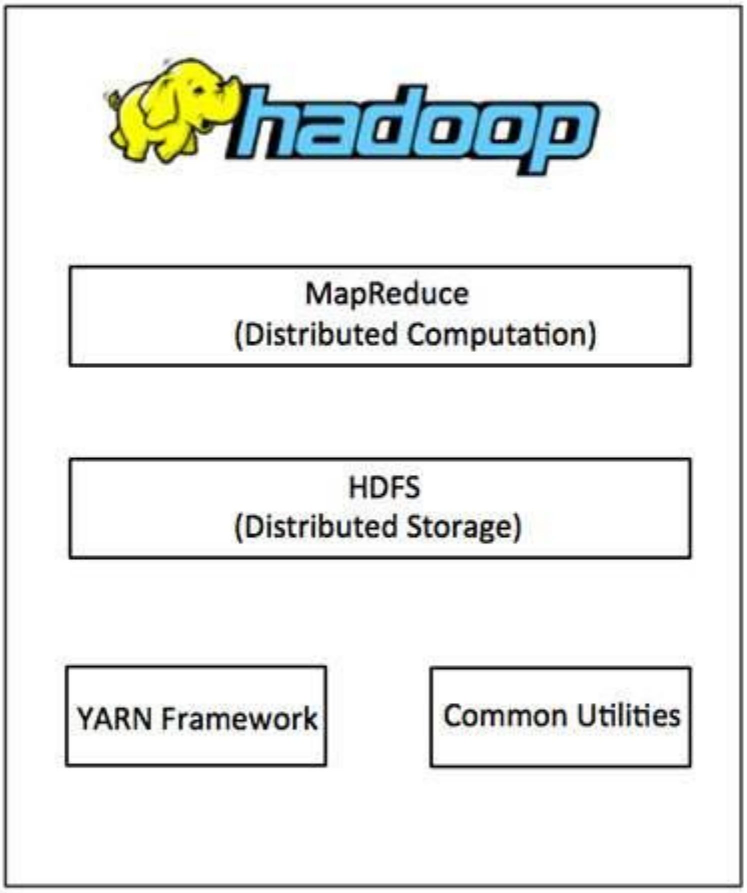
What if we want to load the data which is of type semi-structured and unstructured into the HDFS cluster, or else capture the live streaming data which is generated, from different sources like twitter, weblogs and more into the HDFS cluster, which component of Hadoop ecosystem will be useful to do this kind of job. The solution is FLUME.

Micro blogging today has become a very popular communication tool among internet users. Twitter, one of the largest social media sites receives millions of tweets every day on variety of important issues. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. These posts analysis can be used for decision making in different areas like government, Electronics, Business, Product review etc. Also, sentiment analysis is one of the important areas of analysis of twitter posts that can be very helpful in decision making.
Performing Sentiment Analysis on Twitter is trickier than doing it for large reviews. This is because the tweets are very short (only about 140 characters) and usually contain slangs, emoticons, hash tags and other twitter specific jargon. For the development purpose twitter provides streaming API which allows the developer an access to 1% of tweets tweeted at that time bases on the particular keyword. The object about which we want to perform sentiment analysis is submitted to the twitter API's which does further mining and provides the tweets related to only that object. Twitter data is generally unstructured i.e. use of abbreviations is very high. Also, it allows the use of emoticons which are direct indicators of the author's view on the subject. Tweet messages also consist of a timestamp and the user name.

Hadoop Architecture

At its core, Hadoop has two major layers namely –

1. Processing/Computation layer (MapReduce)
2. Storage layer (Hadoop Distributed File System)



MapReduce

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules –

Hadoop Common − These are Java libraries and utilities required by other Hadoop modules.

Hadoop YARN − This is a framework for job scheduling and cluster resource management.

How Does Hadoop Work?

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So, this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines. Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs –

- Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).

- These files are then distributed across various cluster nodes for further processing.

- HDFS, being on top of the local file system, supervises the processing.
- Blocks are replicated for handling hardware failure.

- Checking that the code was executed successfully.

- Performing the sort that takes place between the map and reduce stages.
- Sending the sorted data to a certain computer.
- Writing the debugging logs for each job.

Advantages of Hadoop

1. Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
2. Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
3. Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
4. Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

INTRODUCTION TO APACHE PIG

Pig is a high-level platform or tool which is used to process the large datasets. It provides a high-level of abstraction for processing over the MapReduce. It provides a high-level scripting language, known as *Pig Latin* which is used to develop the data analysis codes. First, to process the data which is stored in the HDFS, the programmers will write the scripts using the Pig Latin Language. Internally pig engine (a component of Apache Pig) converted all these scripts into a specific map and reduce task. But these are not visible to the programmers in order to provide a high-level of abstraction. Pig Latin and Pig Engine are the two main components of the Apache Pig tool. The result of Pig always stored in the HDFS.

Need of Pig:

One limitation of MapReduce is that the development cycle is very long. Writing the reducer and mapper, compiling packaging the code, submitting the job and retrieving the output is a time-consuming task. Apache Pig reduces the time of development using the multi-query approach. Also, Pig is beneficial for the programmers who are not from Java background. 200 lines of Java code can be written in only 10 lines using the Pig Latin language. Programmers who have SQL knowledge needed less effort to learn Pig Latin.

Evolution of Pig:

Earlier in 2006, Apache Pig was developed by Yahoo’s researchers. At that time, the main idea to develop Pig was to execute the MapReduce jobs on extremely large datasets. In the year 2007, it moved to Apache Software Foundation (ASF) which makes it an open source project. The first version (0.1) of Pig came in the year 2008. The latest version of Apache Pig is 0.18 which came in the year 2017.

Features of Apache Pig:

1. For performing several operations Apache Pig provides rich sets of operators like the filters, join, sort, etc.

2. Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.

3. Apache Pig is extensible so that you can make your own user-defined functions and process.

4. Join operation is easy in Apache Pig.
5. Fewer lines of code.

6. Apache Pig allows splits in the pipeline.

7. The data structure is multivalued, nested and richer.

8. Pig can handle the analysis of both structured and unstructured data.

Applications of Apache Pig:

1. For exploring large datasets Pig Scripting is used.

2. Provides the supports across large data-sets for Ad-hoc queries.

3. In the prototyping of large data-sets processing algorithms.

4. Required to process the time sensitive data loads.

5. For collecting large amounts of datasets in form of search logs and web crawls.

6. Used where the analytical insights are needed using the sampling.

Types of Data Models in Apache Pig:

It consists of the 4 types of data models as follows:

1. **Atom**: It is an atomic data value which is used to store as a string. The main use of this model is that it can be used as a number and as well as a string.

2. **Tuple**: It is an ordered set of the fields.

3. **Bag**: It is a collection of the tuples.
4. **Map**: It is a set of key/value pairs.

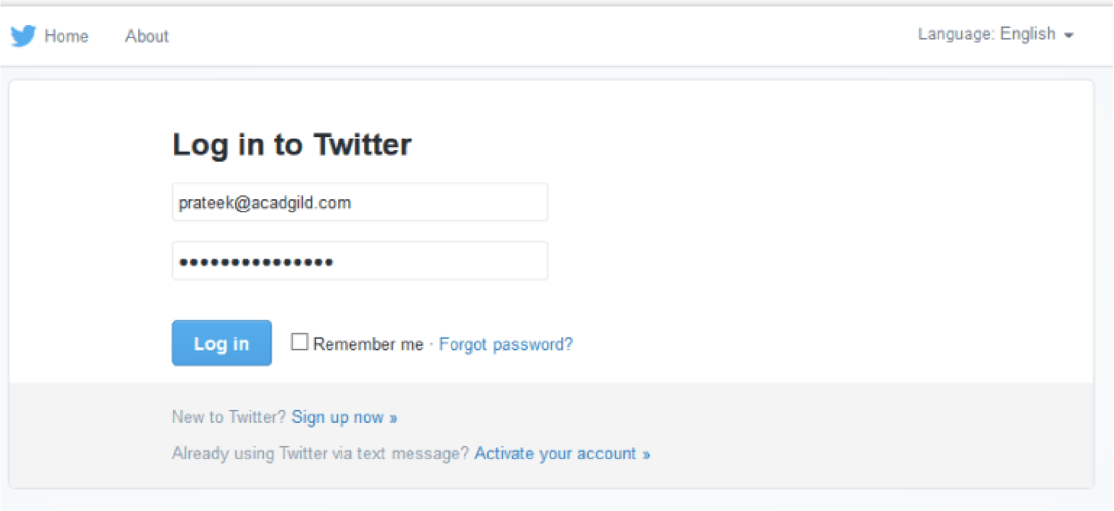
IMPLEMENTATION

Let's look at the necessary pre-requisites:

- Twitter account
- Install Hadoop/Start Hadoop

Data Streaming from Twitter to HDFS

Step 1: Open a Twitter developer account



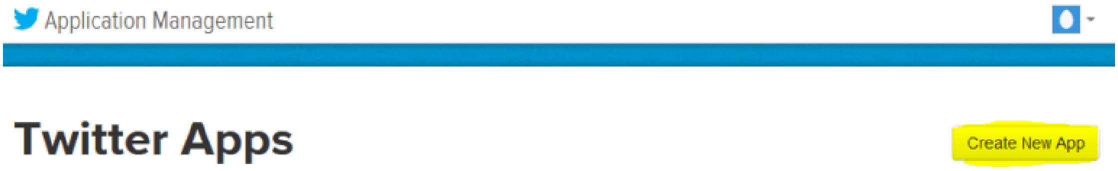
Log in to Twitter

prateek@acadgild.com

••••••••••••••••

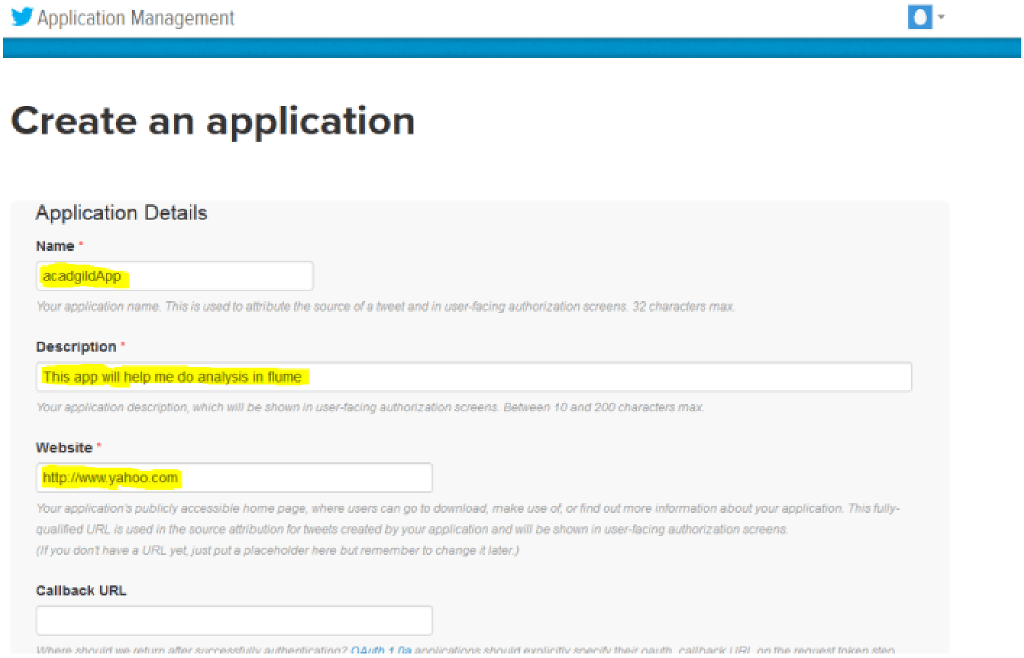
Home
About
Language: English
Log in
☐ Remember me · Forgot password?
New to Twitter? Sign up now »
Already using Twitter via text message? Activate your account »

Step 2: Go to the following link and click on 'create app'. https://apps.twitter.com/app



Twitter Apps

Step 3: Fill in the necessary details.



Create an application

Application Details

Name *

acadgildApp

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

This app will help me do analysis in flume

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

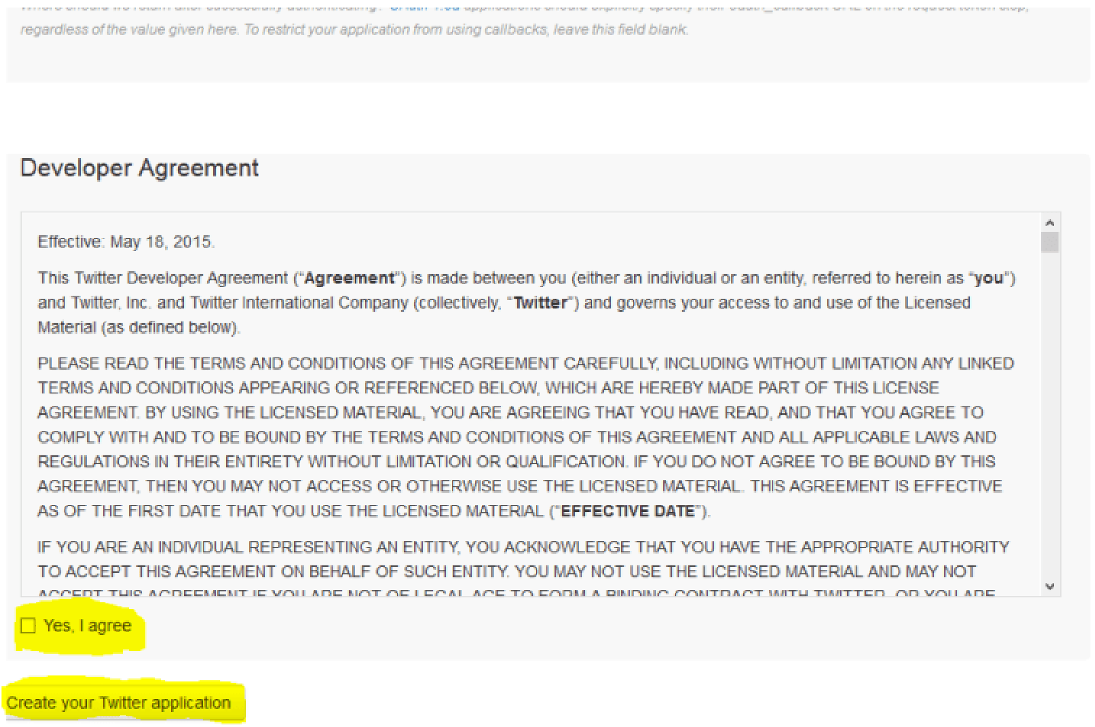
http://www.yahoo.com

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step,

Step 4: Accept the agreement and click on 'create your Twitter application'.



regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Developer Agreement

Effective: May 18, 2015.

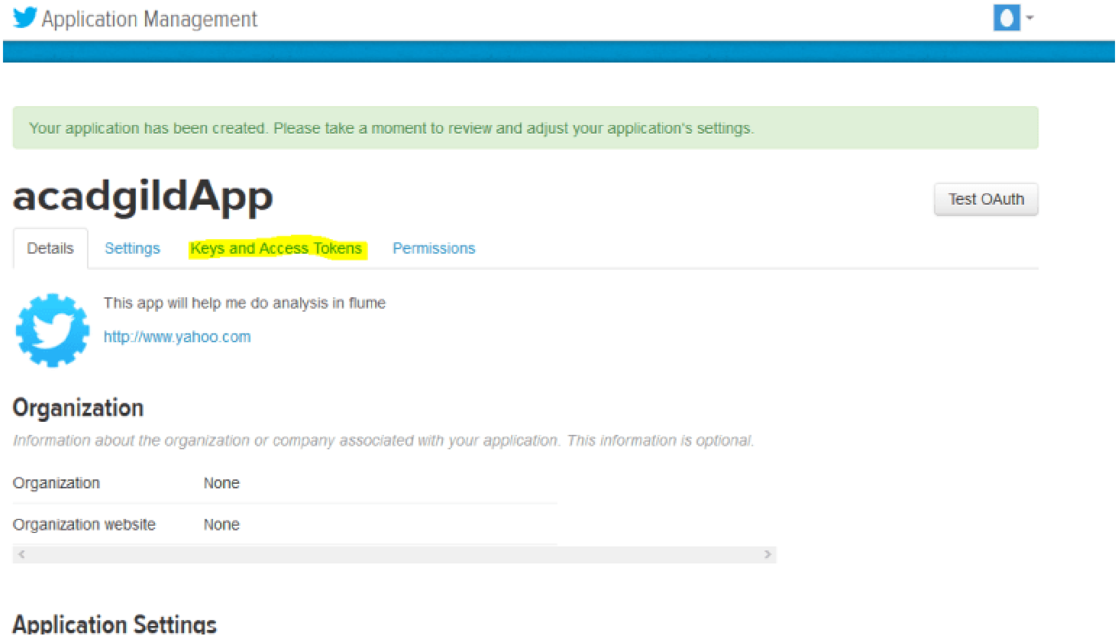
This Twitter Developer Agreement ("**Agreement**") is made between you (either an individual or an entity, referred to herein as "**you**") and Twitter, Inc. and Twitter International Company (collectively, "**Twitter**") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("**EFFECTIVE DATE**").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCEPT THIS AGREEMENT IF YOU ARE NOT OF LEGAL AGE TO FORM A BINDING CONTRACT WITH TWITTER, OR YOU ARE

☐ Yes, I agree

Step 5: Go to 'Keys and Access Token' tab.



Application Management

Your application has been created. Please take a moment to review and adjust your application's settings.

acadgildApp

This app will help me do analysis in flume

http://www.yahoo.com

Organization

Information about the organization or company associated with your application. This information is optional.

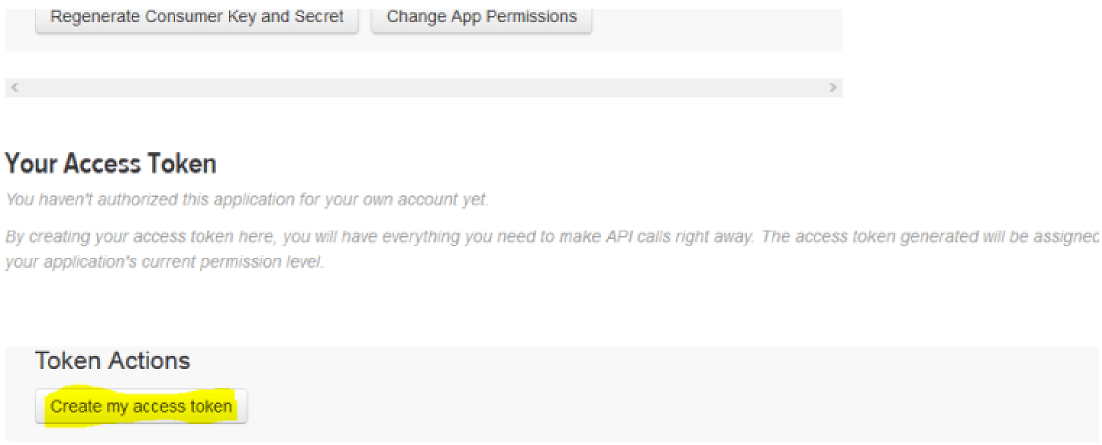
Organization	None
Organization website	None

Application Settings

Step 6: Copy the consumer key and the consumer secret

Step 7: Scroll down further and click on 'create my access token'

.



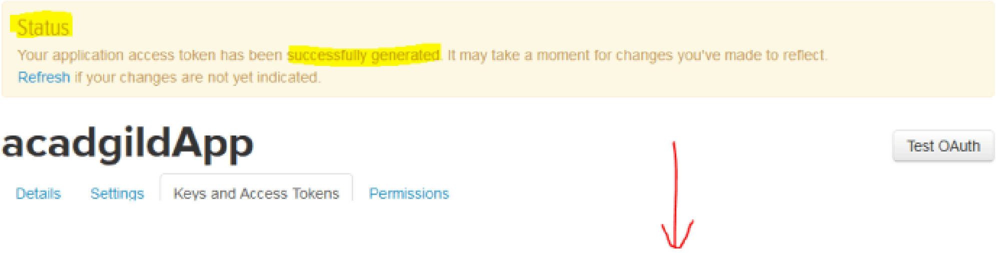
Your Access Token

You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assignec your application's current permission level.

Token Actions

You will now receive a message that says that you have successfully generated your application access token.



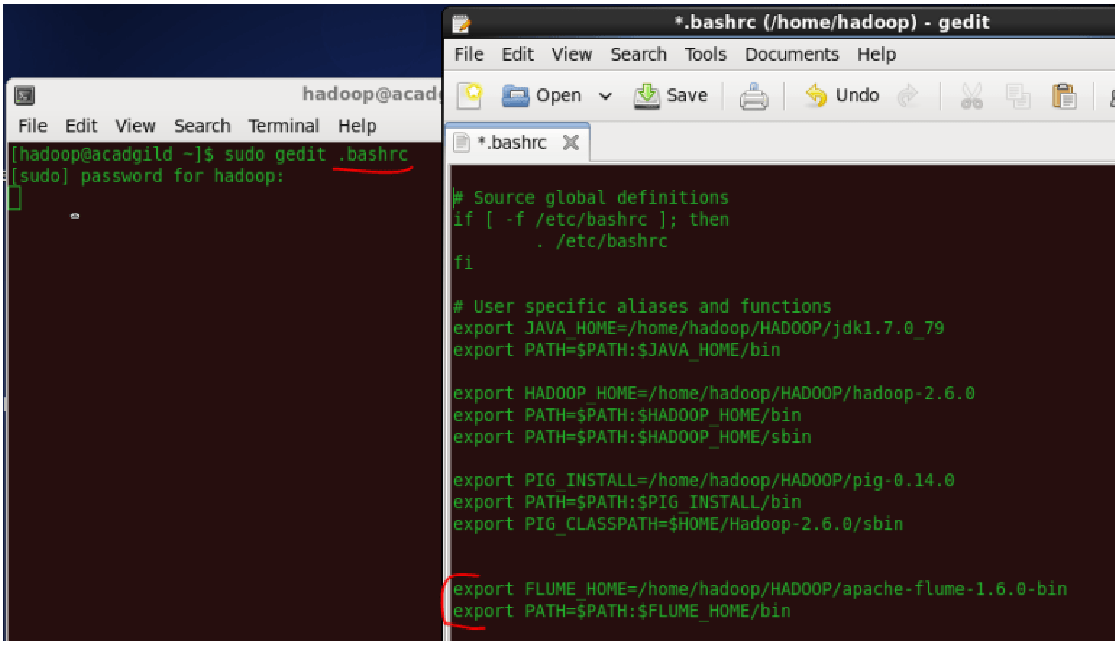
Status

Your application access token has been successfully generated. It may take a moment for changes you've made to reflect.
Refresh if your changes are not yet indicated.

acadgildApp

Step 8: Copy the Access Token and Access token Secret.

Step 9: Download flume tar file from below link and extract it https://drive.google.com/open?id=0B2nmxAJLHEE8ZGlLeE05TEUtdEE
Extract the flume.tar file and update the path of extracted file in .bashrc
NOTE: keep the path same as where the extracted file exists.

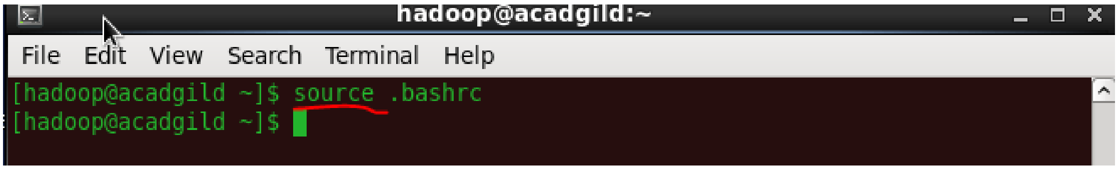


hadoop@acad
File Edit View Search Terminal Help
*.bashrc (/home/hadoop) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*.bashrc

```
[hadoop@acadgild ~]$ sudo gedit .bashrc
[sudo] password for hadoop:
```

```
# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# User specific aliases and functions
export JAVA_HOME=/home/hadoop/HADOOP/jdk1.7.0_79
export PATH=$PATH:$JAVA_HOME/bin

export HADOOP_HOME=/home/hadoop/HADOOP/hadoop-2.6.0
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin

export PIG_INSTALL=/home/hadoop/HADOOP/pig-0.14.0
export PATH=$PATH:$PIG_INSTALL/bin
export PIG_CLASSPATH=$HOME/Hadoop-2.6.0/sbin


export FLUME_HOME=/home/hadoop/HADOOP/apache-flume-1.6.0-bin
export PATH=$PATH:$FLUME_HOME/bin
```

Update the bashrc file with source command.

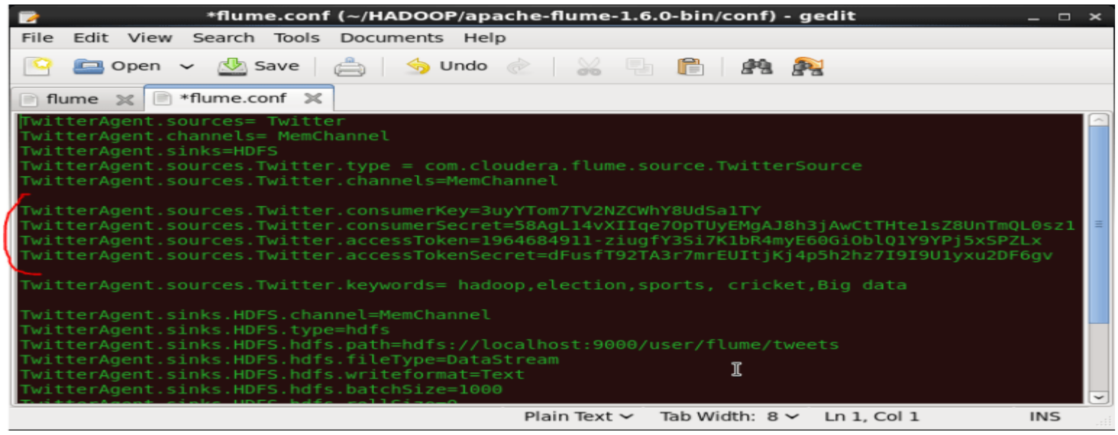


```
[hadoop@acadgild ~]$ source .bashrc
[hadoop@acadgild ~]$
```

Step 10: Create a new file inside the ‘conf’ directory inside the Flume-extracted directory.

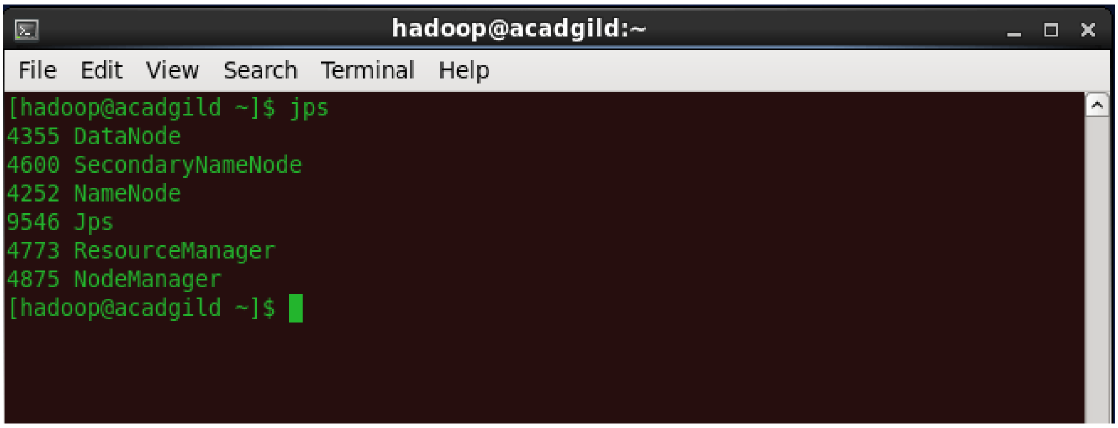
Step 11: Copy the information from the below link and paste it inside the newly created file.
https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk

Step 12: Change the twitter api keys with the keys generated as shown in the step no 6 and step number


*flume.conf (~/HADOOP/apache-flume-1.6.0-bin/conf) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
flume
*flume.conf
Plain Text
Tab Width: 8
Ln 1, Col 1
INS

```
TwitterAgent.sources= Twitter
TwitterAgent.channels= MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels=MemChannel

TwitterAgent.sources.Twitter.consumerKey=3uyYTom7TV2NZCWhY8UdSa1TY
TwitterAgent.sources.Twitter.consumerSecret=58AgL14vXIIqe7OpTUyEMgAJ8h3jAwCtTHte1sZ8UnTmQL0sz1
TwitterAgent.sources.Twitter.accessToken=1964684911-ziugfY3Si7K1bR4myE60GiOblQ1Y9YPj5xSPZLx
TwitterAgent.sources.Twitter.accessTokenSecret=dFusfT92TA3r7mrEUItjKj4p5h2hz7I9I9U1yxu2DF6gv

TwitterAgent.sources.Twitter.keywords= hadoop,election,sports, cricket,Big data

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/user/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
```

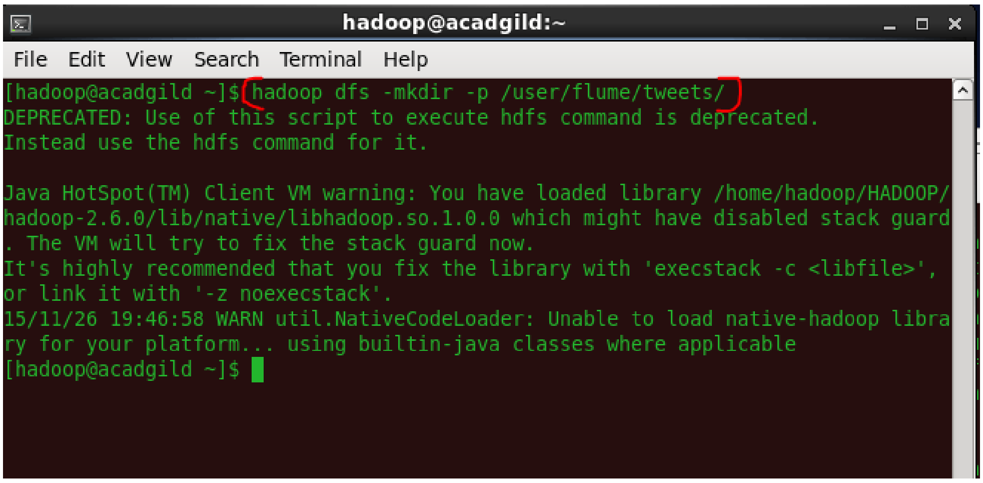
Step 13: Open the terminal to check for all Hadoop daemons running, by using the ‘jps’ command.



```
[hadoop@acadgild ~]$ jps
4355 DataNode
4600 SecondaryNameNode
4252 NameNode
9546 Jps
4773 ResourceManager
4875 NodeManager
[hadoop@acadgild ~]$
```

Step 14: Using the below command, create a directory inside HDFS where Twitter data will be stored.

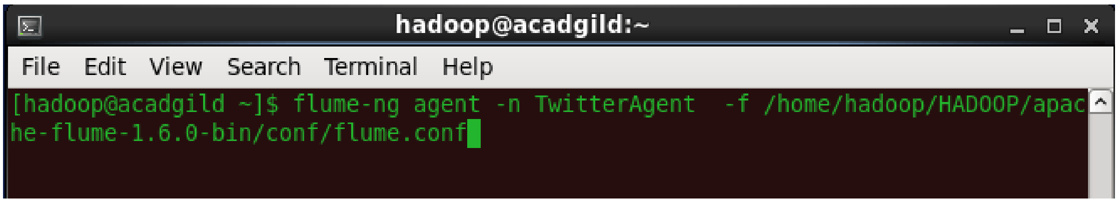
Hadoop dfs –mkdir –p /user/flume/tweets



```
[hadoop@acadgild ~]$ hadoop dfs -mkdir -p /user/flume/tweets/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Java HotSpot(TM) Client VM warning: You have loaded library /home/hadoop/HADOOP/
hadoop-2.6.0/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard
. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>',
or link it with '-z noexecstack'.
15/11/26 19:46:58 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[hadoop@acadgild ~]$ 
```

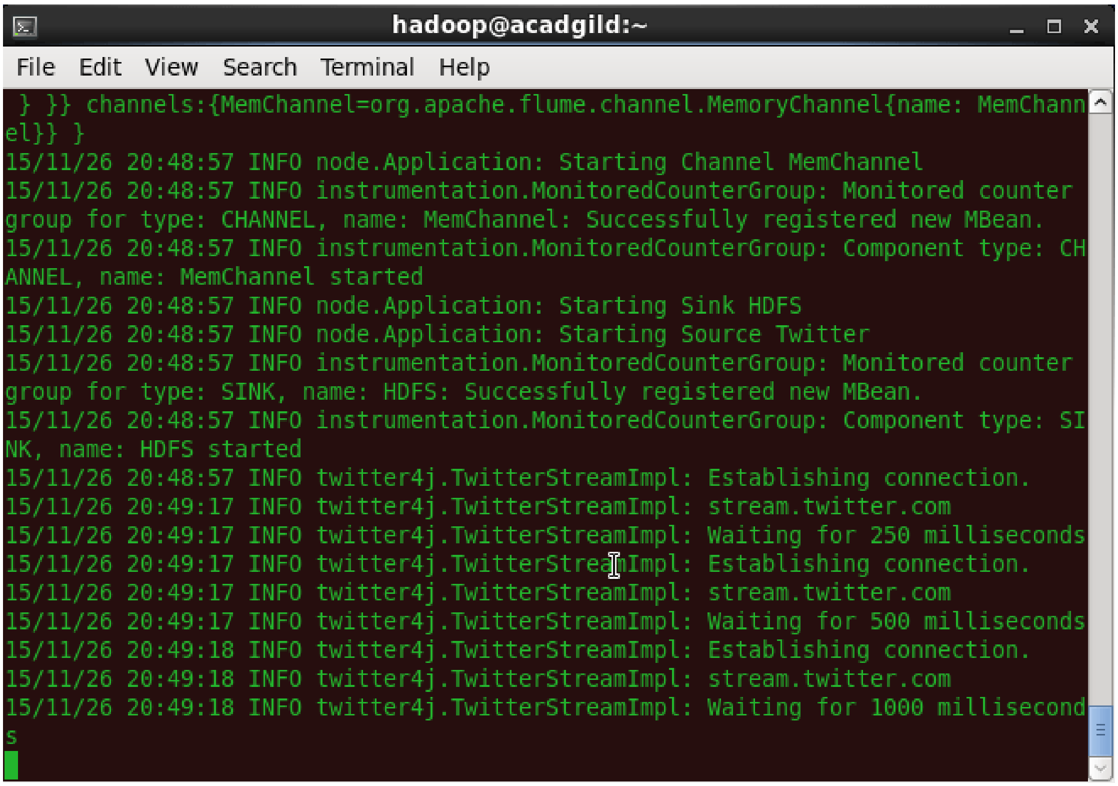
Step 15: For fetching data from Twitter, give the below command in the terminal.

flume-ng agent -n TwitterAgent -f <location of created/edited conf file>



```
[hadoop@acadgild ~]$ flume-ng agent -n TwitterAgent  -f /home/hadoop/HADOOP/apac
he-flume-1.6.0-bin/conf/flume.conf
```

This will start fetching data from Twitter and send it to the HDFS.

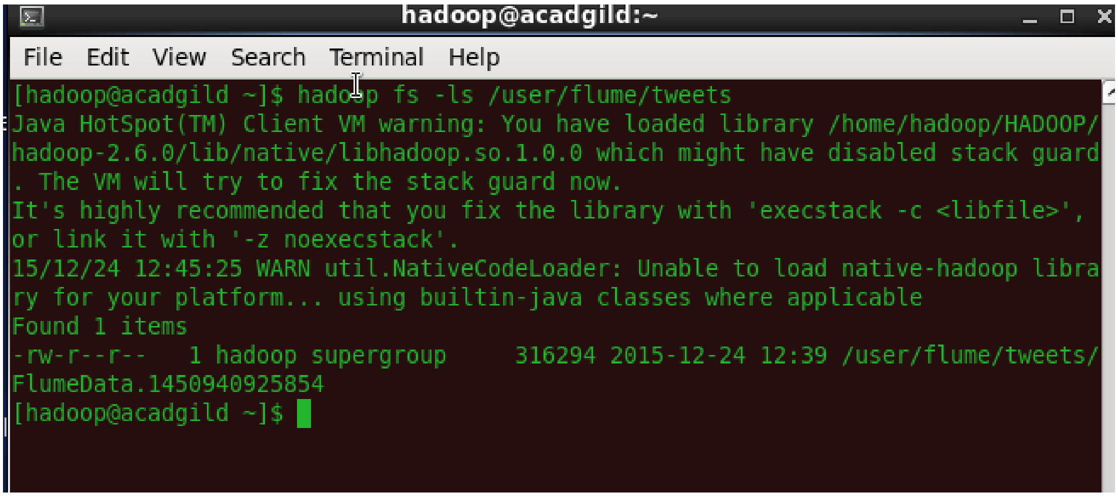


```
 } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name: MemChann
el}} }
15/11/26 20:48:57 INFO node.Application: Starting Channel MemChannel
15/11/26 20:48:57 INFO instrumentation.MonitoredCounterGroup: Monitored counter
group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.
15/11/26 20:48:57 INFO instrumentation.MonitoredCounterGroup: Component type: CH
ANNEL, name: MemChannel started
15/11/26 20:48:57 INFO node.Application: Starting Sink HDFS
15/11/26 20:48:57 INFO node.Application: Starting Source Twitter
15/11/26 20:48:57 INFO instrumentation.MonitoredCounterGroup: Monitored counter
group for type: SINK, name: HDFS: Successfully registered new MBean.
15/11/26 20:48:57 INFO instrumentation.MonitoredCounterGroup: Component type: SI
NK, name: HDFS started
15/11/26 20:48:57 INFO twitter4j.TwitterStreamImpl: Establishing connection.
15/11/26 20:49:17 INFO twitter4j.TwitterStreamImpl: stream.twitter.com
15/11/26 20:49:17 INFO twitter4j.TwitterStreamImpl: Waiting for 250 milliseconds
15/11/26 20:49:17 INFO twitter4j.TwitterStreamImpl: Establishing connection.
15/11/26 20:49:17 INFO twitter4j.TwitterStreamImpl: stream.twitter.com
15/11/26 20:49:17 INFO twitter4j.TwitterStreamImpl: Waiting for 500 milliseconds
15/11/26 20:49:18 INFO twitter4j.TwitterStreamImpl: Establishing connection.
15/11/26 20:49:18 INFO twitter4j.TwitterStreamImpl: stream.twitter.com
15/11/26 20:49:18 INFO twitter4j.TwitterStreamImpl: Waiting for 1000 millisecond
s
```

To stop fetching data, press *'Ctrl+c'*. This will end the process of fetching the data.

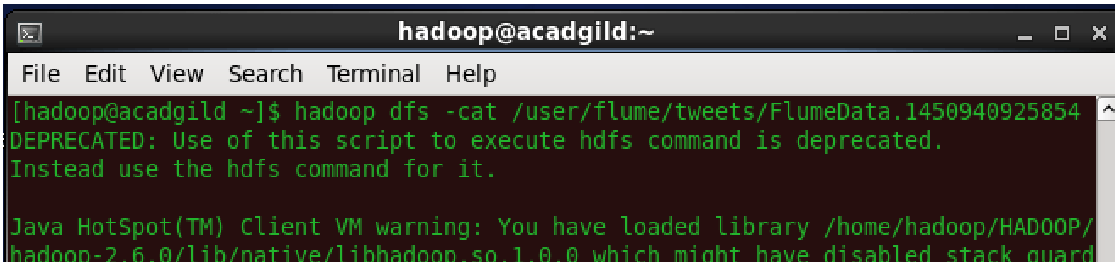
Step 16: To check the contents of the Tweets folder, use the following command:

hadoop dfs –ls /user/flume/tweets

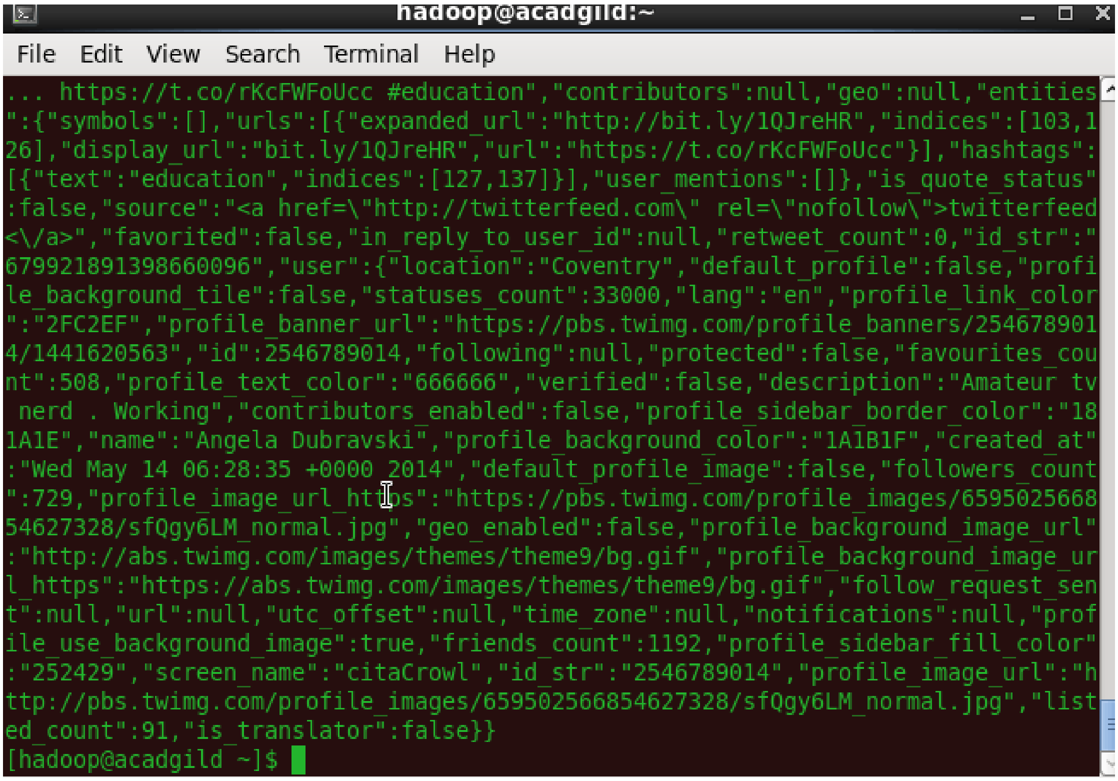


```
[hadoop@acadgild ~]$ hadoop fs -ls /user/flume/tweets
Java HotSpot(TM) Client VM warning: You have loaded library /home/hadoop/HADOOP/
hadoop-2.6.0/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard
. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>',
or link it with '-z noexecstack'.
15/12/24 12:45:25 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 hadoop supergroup     316294 2015-12-24 12:39 /user/flume/tweets/
FlumeData.1450940925854
[hadoop@acadgild ~]$ 
```

Step 17: To see the data inside this file, type the following command: **hadoop dfs –cat /user/flume/tweets/<flumeData file name>**



```
[hadoop@acadgild ~]$ hadoop dfs -cat /user/flume/tweets/FlumeData.1450940925854
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Java HotSpot(TM) Client VM warning: You have loaded library /home/hadoop/HADOOP/
hadoop-2.6.0/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard
```



```
... https://t.co/rKcFWFoUcc #education","contributors":null,"geo":null,"entities
":{"symbols":[],"urls":[{"expanded_url":"http://bit.ly/1QJreHR","indices":[103,1
26],"display_url":"bit.ly/1QJreHR","url":"https://t.co/rKcFWFoUcc"}],"hashtags":
[{"text":"education","indices":[127,137]}],"user_mentions":[]},"is_quote_status"
:false,"source":"<a href=\"http://twitterfeed.com\" rel=\"nofollow\">twitterfeed
<\/a>","favorited":false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"
679921891398660096","user":{"location":"Coventry","default_profile":false,"profi
le_background_tile":false,"statuses_count":33000,"lang":"en","profile_link_color
":"2FC2EF","profile_banner_url":"https://pbs.twimg.com/profile_banners/254678901
4/1441620563","id":2546789014,"following":null,"protected":false,"favourites_cou
nt":508,"profile_text_color":"666666","verified":false,"description":"Amateur tv
 nerd . Working","contributors_enabled":false,"profile_sidebar_border_color":"18
1A1E","name":"Angela Dubravski","profile_background_color":"1A1B1F","created_at"
:"Wed May 14 06:28:35 +0000 2014","default_profile_image":false,"followers_count
":729,"profile_image_url_https":"https://pbs.twimg.com/profile_images/6595025668
54627328/sfQgy6LM_normal.jpg","geo_enabled":false,"profile_background_image_url"
:"http://abs.twimg.com/images/themes/theme9/bg.gif","profile_background_image_ur
l_https":"https://abs.twimg.com/images/themes/theme9/bg.gif","follow_request_sen
t":null,"url":null,"utc_offset":null,"time_zone":null,"notifications":null,"prof
ile_use_background_image":true,"friends_count":1192,"profile_sidebar_fill_color"
:"252429","screen_name":"citaCrowl","id_str":"2546789014","profile_image_url":"h
ttp://pbs.twimg.com/profile_images/659502566854627328/sfQgy6LM_normal.jpg","list
ed_count":91,"is_translator":false}}
[hadoop@acadgild ~]$ 
```

We have completed the action of fetching live-streaming data from Twitter and loaded it to the HDFS, using Flume after streaming data from twitter, we need to analysis of this data and give polarity accordingly.

ANALYSIS OF DATA STREAMED FROM TWITTER IN PREVIOUS STEPS

Register pig json loader:

1- elephant-bird-hadoop-compat-4.1.jar
2- elephant-bird-pig-4.1.jar
3- json-simple-1.1.1.jar

Execution model:

1- Invoke PIG grunt shell in map-reduce mode.
2- Register all the PIG JSON LOADER with PIG.
3- Load AFINN dictionary from HDFS to PIG bag.

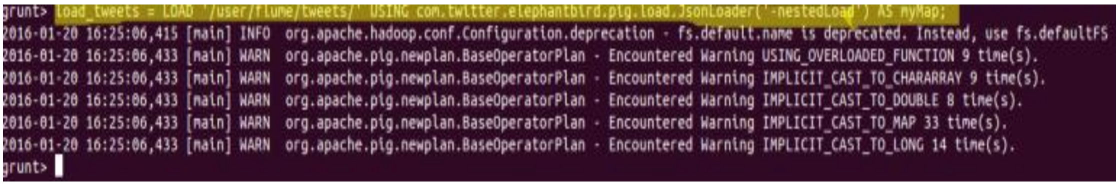
Step 1-Register the downloaded jars in pig by using the below commands:

```
Grunt> REGISTER '/home/edureka/Desktop/elephant-bird-hadoop-compat-4.1.jar;
Grunt> REGISTER '/home/edureka/Desktop/elephant-bird-pig-4.1.jar';
Grunt> REGISTER '/home/edureka/Desktop/json-simple-1.1.1.jar';
```

Note: You need to provide the path of the jar file accordingly.
After registering the required jars, we can now write a Pig script to perform Sentiment Analysis.

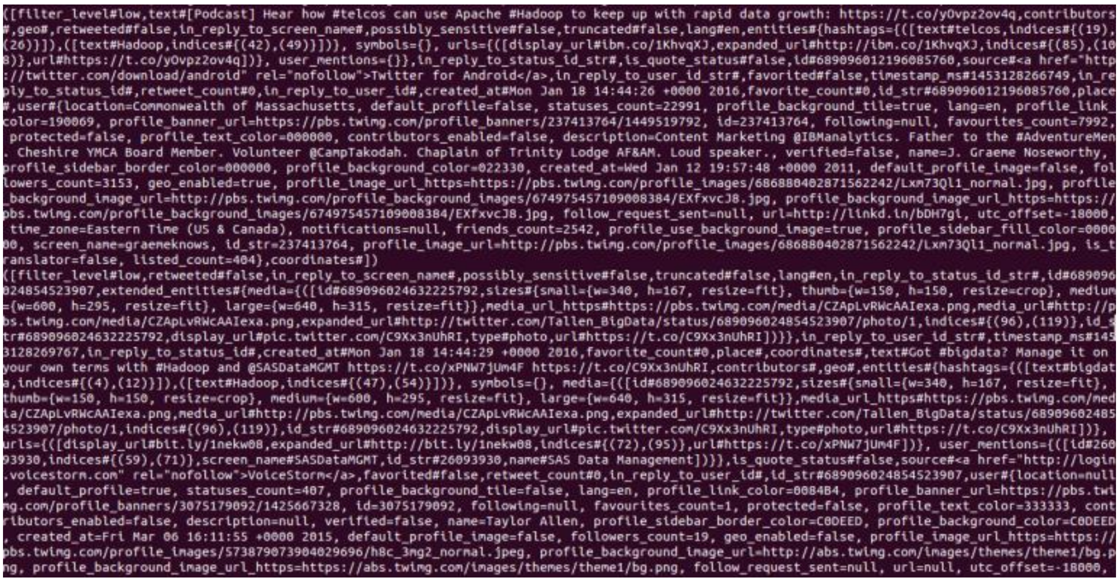
STEP 2- The tweets are in nested Json format and consists of map data types. We need to load the tweets using JsonLoader which supports maps, so we are using **elephant bird JsonLoader** to load the tweets.
Below is the first Pig statement required to load the tweets into Pig: -

```
load_tweets = LOAD '/user/flume/tweets/' USING
com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap;
```



```
grunt> load_tweets = LOAD '/user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap;
2016-01-20 16:25:06,415 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-01-20 16:25:06,433 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 9 time(s).
2016-01-20 16:25:06,433 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 9 time(s).
2016-01-20 16:25:06,433 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:25:06,433 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 33 time(s).
2016-01-20 16:25:06,433 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

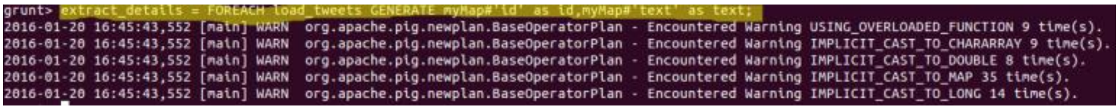
Step 3-When we dump the above relation, we can see that all the tweets got loaded successfully.



```
([filter_level#low,text#[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,contributors
#,geo#,retweeted#false,in_reply_to_screen_name#,possibly_sensitive#false,truncated#false,lang#en,entities#{hashtags={([text#telcos,indices#{(19),
(26)}]),([text#Hadoop,indices#{(42),(49)}])}, symbols={}, urls={([display_url#ibm.co/1KhvqXJ,expanded_url#http://ibm.co/1KhvqXJ,indices#{(85),(10
8)},url#https://t.co/yOvpz2ov4q])}, user_mentions={}},in_reply_to_status_id_str#,is_quote_status#false,id#689096012196085760,source#<a href="http
://twitter.com/download/android" rel="nofollow">Twitter for Android</a>,in_reply_to_user_id_str#,favorited#false,timestamp_ms#1453128266749,in_re
ply_to_status_id#,retweet_count#0,in_reply_to_user_id#,created_at#Mon Jan 18 14:44:26 +0000 2016,favorite_count#0,id_str#689096012196085760,place
#,user#{location=Commonwealth of Massachusetts, default_profile=false, statuses_count=22991, profile_background_tile=true, lang=en, profile_link_
color=190069, profile_banner_url=https://pbs.twimg.com/profile_banners/237413764/1449519792, id=237413764, following=null, favourites_count=7992,
 protected=false, profile_text_color=000000, contributors_enabled=false, description=Content Marketing @IBManalytics. Father to the #AdventureMen
. Cheshire YMCA Board Member. Volunteer @CampTakodah. Chaplain of Trinity Lodge AF&AM. Loud speaker., verified=false, name=J. Graeme Noseworthy,
profile_sidebar_border_color=000000, profile_background_color=022330, created_at=Wed Jan 12 19:57:48 +0000 2011, default_profile_image=false, fol
lowers_count=3153, geo_enabled=true, profile_image_url_https=https://pbs.twimg.com/profile_images/686880402871562242/Lxm73Ql1_normal.jpg, profile
_background_image_url=http://pbs.twimg.com/profile_background_images/674975457109008384/EXfxvcJ8.jpg, profile_background_image_url_https=https://
pbs.twimg.com/profile_background_images/674975457109008384/EXfxvcJ8.jpg, follow_request_sent=null, url=http://linkd.in/bDH7gl, utc_offset=-18000,
 time_zone=Eastern Time (US & Canada), notifications=null, friends_count=2542, profile_use_background_image=true, profile_sidebar_fill_color=0000
00, screen_name=graemeknows, id_str=237413764, profile_image_url=http://pbs.twimg.com/profile_images/686880402871562242/Lxm73Ql1_normal.jpg, is_t
ranslator=false, listed_count=404},coordinates#])
([filter_level#low,retweeted#false,in_reply_to_screen_name#,possibly_sensitive#false,truncated#false,lang#en,in_reply_to_status_id_str#,id#689096
024854523907,extended_entities#{media={([id#689096024632225792,sizes#{small={w=340, h=167, resize=fit}, thumb={w=150, h=150, resize=crop}, medium
={w=600, h=295, resize=fit}, large={w=640, h=315, resize=fit}},media_url_https#https://pbs.twimg.com/media/CZApLvRWcAAIexa.png,media_url#http://p
bs.twimg.com/media/CZApLvRWcAAIexa.png,expanded_url#http://twitter.com/Tallen_BigData/status/689096024854523907/photo/1,indices#{(96),(119)},id_s
tr#689096024632225792,display_url#pic.twitter.com/C9Xx3nUhRI,type#photo,url#https://t.co/C9Xx3nUhRI])}},in_reply_to_user_id_str#,timestamp_ms#145
3128269767,in_reply_to_status_id#,created_at#Mon Jan 18 14:44:29 +0000 2016,favorite_count#0,place#,coordinates#,text#Got #bigdata? Manage it on
your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUm4F https://t.co/C9Xx3nUhRI,contributors#,geo#,entities#{hashtags={([text#bigdat
a,indices#{(4),(12)}]),([text#Hadoop,indices#{(47),(54)}])}, symbols={}, media={([id#689096024632225792,sizes#{small={w=340, h=167, resize=fit},
thumb={w=150, h=150, resize=crop}, medium={w=600, h=295, resize=fit}, large={w=640, h=315, resize=fit}},media_url_https#https://pbs.twimg.com/med
ia/CZApLvRWcAAIexa.png,media_url#http://pbs.twimg.com/media/CZApLvRWcAAIexa.png,expanded_url#http://twitter.com/Tallen_BigData/status/68909602485
4523907/photo/1,indices#{(96),(119)},id_str#689096024632225792,display_url#pic.twitter.com/C9Xx3nUhRI,type#photo,url#https://t.co/C9Xx3nUhRI])},
urls={([display_url#bit.ly/1nekw08,expanded_url#http://bit.ly/1nekw08,indices#{(72),(95)},url#https://t.co/xPNW7jUm4F])}, user_mentions={([id#260
93930,indices#{(59),(71)},screen_name#SASDataMGMT,id_str#26093930,name#SAS Data Management])}},is_quote_status#false,source#<a href="http://login
.voicestorm.com" rel="nofollow">VoiceStorm</a>,favorited#false,retweet_count#0,in_reply_to_user_id#,id_str#689096024854523907,user#{location=null
, default_profile=true, statuses_count=407, profile_background_tile=false, lang=en, profile_link_color=0084B4, profile_banner_url=https://pbs.twi
mg.com/profile_banners/3075179092/1425667328, id=3075179092, following=null, favourites_count=1, protected=false, profile_text_color=333333, cont
ributors_enabled=false, description=null, verified=false, name=Taylor Allen, profile_sidebar_border_color=C0DEED, profile_background_color=C0DEED
, created_at=Fri Mar 06 16:11:55 +0000 2015, default_profile_image=false, followers_count=19, geo_enabled=false, profile_image_url_https=https://
pbs.twimg.com/profile_images/573879073904029696/h8c_3mg2_normal.jpeg, profile_background_image_url=http://abs.twimg.com/images/themes/theme1/bg.p
ng, profile_background_image_url_https=https://abs.twimg.com/images/themes/theme1/bg.png, follow_request_sent=null, url=null, utc_offset=-18000,
```

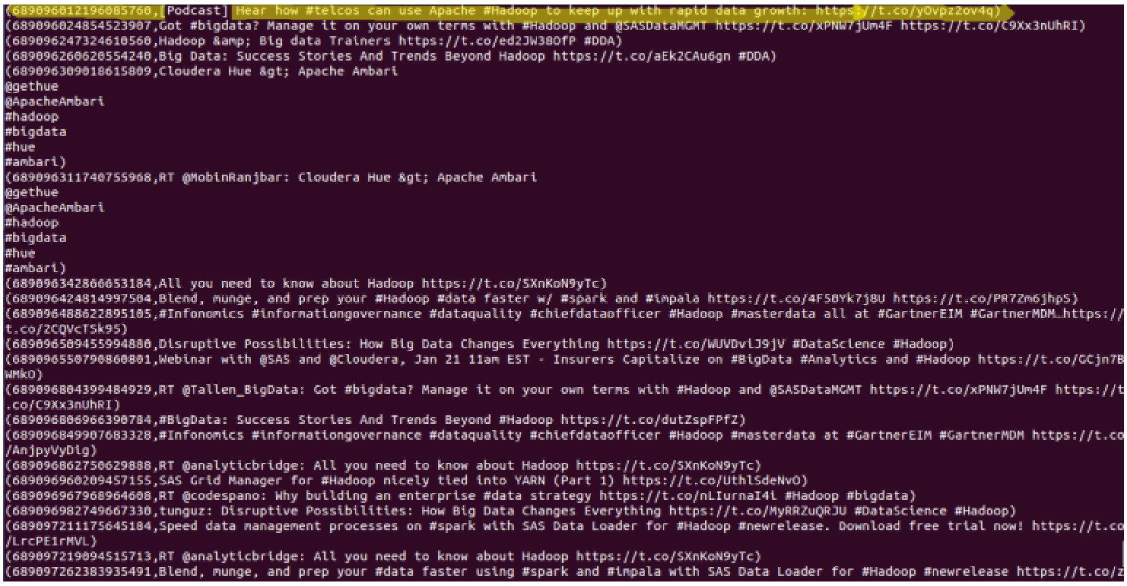
Step 4-Now, we shall extract the **id** and the **tweet text** from the above tweets. The Pig statement necessary to perform this is as shown below:

extract_details = FOREACH load_tweets GENERATE myMap#'id' as id,myMap#'text' as text;



```
grunt> extract_details = FOREACH load_tweets GENERATE myMap#'id' as id,myMap#'text' as text;
2016-01-20 16:45:43,552 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 9 time(s).
2016-01-20 16:45:43,552 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 9 time(s).
2016-01-20 16:45:43,552 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:45:43,552 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 16:45:43,552 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
```

We can see the extracted **id** and **tweet text** from the tweets in the below screen shot.

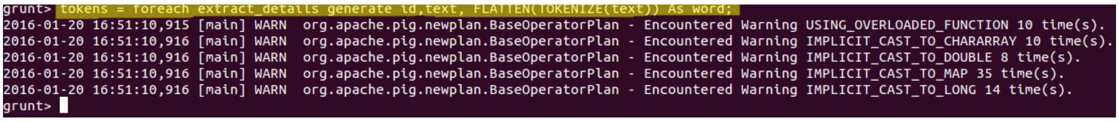


```
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q)
(689096024854523907,Got #bigdata? Manage it on your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUm4F https://t.co/C9Xx3nUhRI)
(689096247324610560,Hadoop &amp; Big data Trainers https://t.co/ed2JW38OfP #DDA)
(689096260620554240,Big Data: Success Stories And Trends Beyond Hadoop https://t.co/aEk2CAu6gn #DDA)
(689096309018615809,Cloudera Hue &gt; Apache Ambari
@gethue
@ApacheAmbari
#hadoop
#bigdata
#hue
#ambari)
(689096311740755968,RT @MobinRanjbar: Cloudera Hue &gt; Apache Ambari
@gethue
@ApacheAmbari
#hadoop
#bigdata
#hue
#ambari)
(689096342866653184,All you need to know about Hadoop https://t.co/SXnKoN9yTc)
(689096424814997504,Blend, munge, and prep your #Hadoop #data faster w/ #spark and #impala https://t.co/4F50Yk7j8U https://t.co/PR7Zm6jhpS)
(689096488622895105,#Infonomics #informationgovernance #dataquality #chiefdataofficer #Hadoop #masterdata all at #GartnerEIM #GartnerMDM…https://
t.co/2CQVcTSk95)
(689096509455994880,Disruptive Possibilities: How Big Data Changes Everything https://t.co/WUVDviJ9jV #DataScience #Hadoop)
(689096550790860801,Webinar with @SAS and @Cloudera, Jan 21 11am EST - Insurers Capitalize on #BigData #Analytics and #Hadoop https://t.co/GCjn7B
WMkO)
(689096804399484929,RT @Tallen_BigData: Got #bigdata? Manage it on your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUm4F https://t
.co/C9Xx3nUhRI)
(689096806966390784,#BigData: Success Stories And Trends Beyond #Hadoop https://t.co/dutZspFPfZ)
(689096849907683328,#Infonomics #informationgovernance #dataquality #chiefdataofficer #Hadoop #masterdata at #GartnerEIM #GartnerMDM https://t.co
/AnjpyVyDig)
(689096862750629888,RT @analyticbridge: All you need to know about Hadoop https://t.co/SXnKoN9yTc)
(689096960209457155,SAS Grid Manager for #Hadoop nicely tied into YARN (Part 1) https://t.co/UthlSdeNvO)
(689096967968964608,RT @codespano: Why building an enterprise #data strategy https://t.co/nLIurnaI4l #Hadoop #bigdata)
(689096982749667330,tunguz: Disruptive Possibilities: How Big Data Changes Everything https://t.co/MyRRZuQRJU #DataScience #Hadoop)
(689097211175645184,Speed data management processes on #spark with SAS Data Loader for #Hadoop #newrelease. Download free trial now! https://t.co
/LrcPE1rMVL)
(689097219094515713,RT @analyticbridge: All you need to know about Hadoop https://t.co/SXnKoN9yTc)
(689097262383935491,Blend, munge, and prep your #data faster using #spark and #impala with SAS Data Loader for #Hadoop #newrelease https://t.co/z
```

Step 5-We have the tweet id and the tweet text in the relation named as **extract_details**.

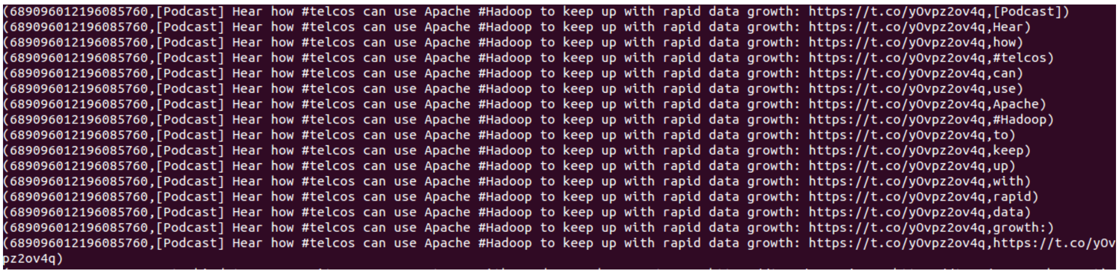
Now, we shall extract the words from the text using the TOKENIZE key word in Pig.

```
tokens = foreach extract_details generate id,text,
FLATTEN(TOKENIZE(text)) As word
```



```
grunt> tokens = foreach extract_details generate id,text, FLATTEN(TOKENIZE(text)) As word;
2016-01-20 16:51:10,915 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 16:51:10,916 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 16:51:10,916 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:51:10,916 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 16:51:10,916 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt> 
```

From the below screen shot, we can see that the text got divided into words.

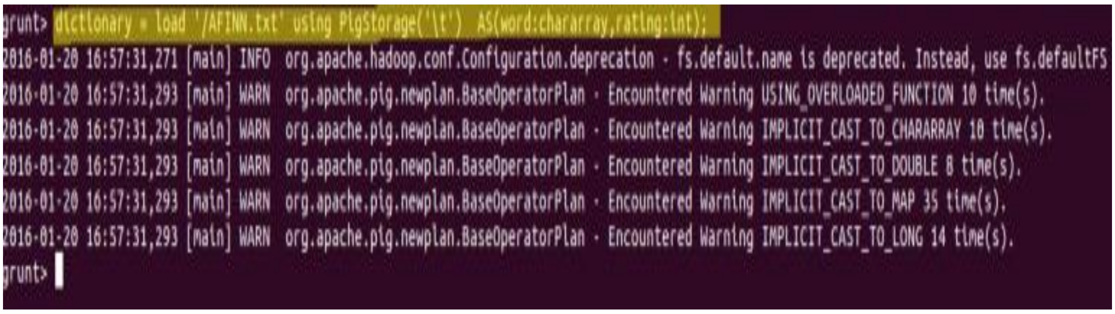


```
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,[Podcast])
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,Hear)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,how)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,#telcos)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,can)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,use)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,Apache)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,#Hadoop)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,to)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,keep)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,up)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,with)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,rapid)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,data)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,growth:)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,https://t.co/yOv
pz2ov4q)
```

STEP 6-Now, we have to analyze the Sentiment for the tweet by using the words in the text. We will rate the word as per its meaning from +5 to -5 using the dictionary AFINN. The AFINN is a dictionary which consists of 2500 words which are rated from +5 to -5 depending on their meaning. You can download the dictionary from the above given link.

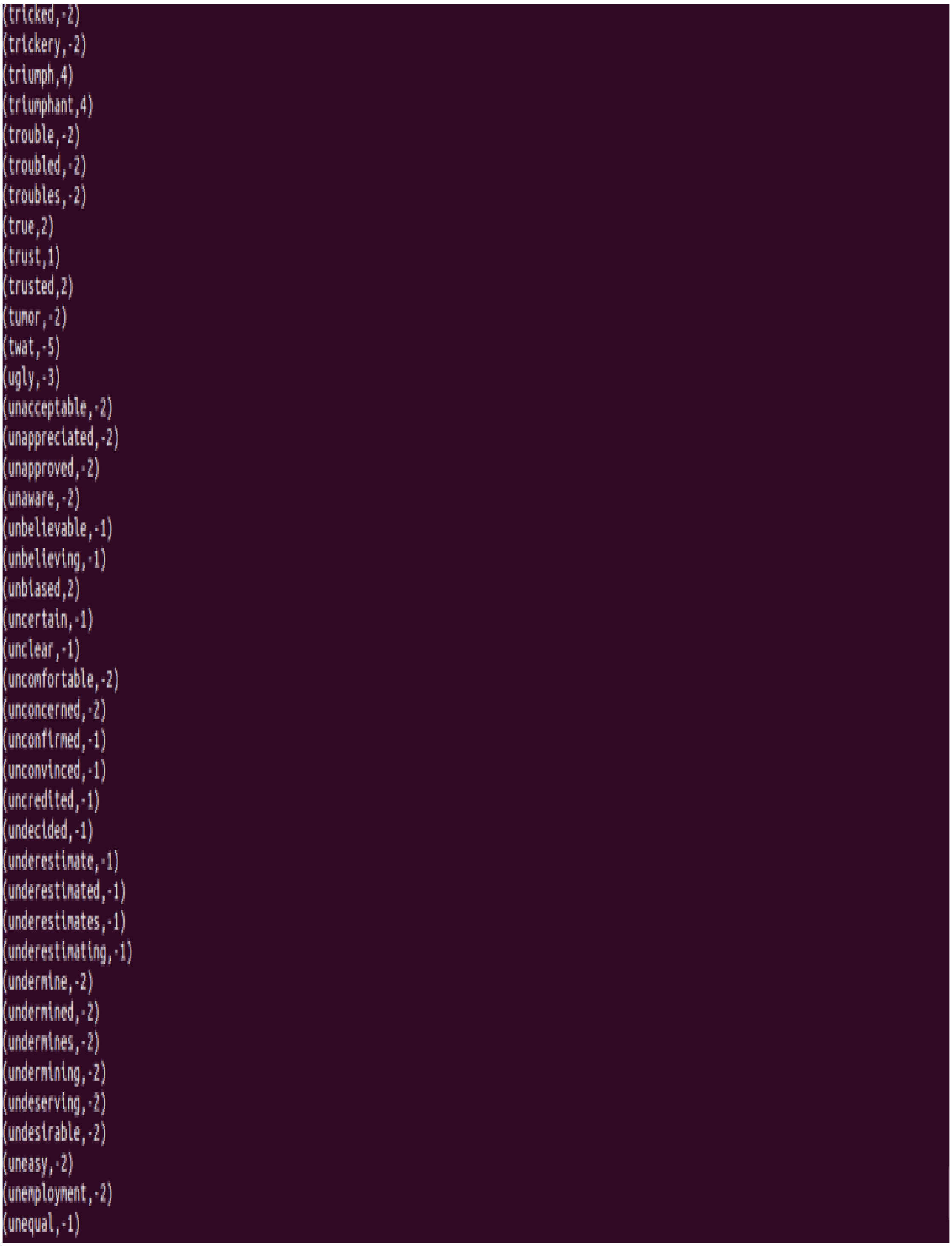
We will load the dictionary into pig by using the below statement:

```
dictionary = load '/AFINN.txt' using PigStorage('\t') AS
(word:chararray,rating:int);
```



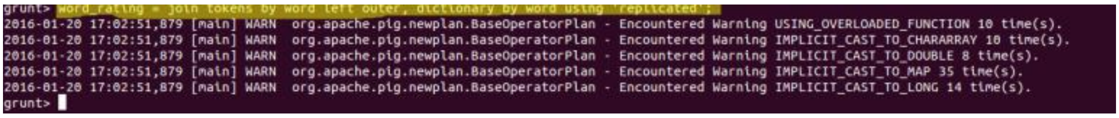
```
grunt> dictionary = load '/AFINN.txt' using PigStorage('\t')  AS(word:chararray,rating:int);
2016-01-20 16:57:31,271 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-01-20 16:57:31,293 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 16:57:31,293 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 16:57:31,293 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:57:31,293 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 16:57:31,293 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt> 
```

The output of this will be look like following image:-



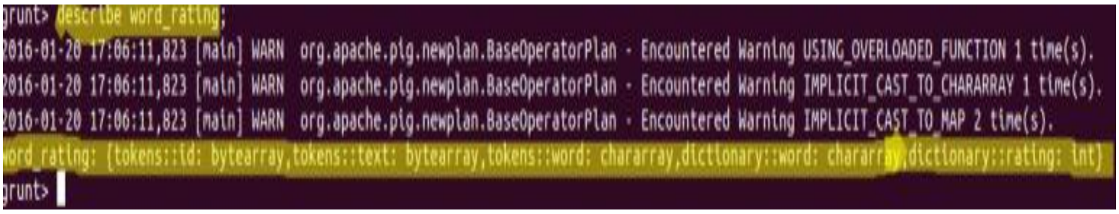
```
(tricked,-2)
(trickery,-2)
(triumph,4)
(triumphant,4)
(trouble,-2)
(troubled,-2)
(troubles,-2)
(true,2)
(trust,1)
(trusted,2)
(tumor,-2)
(twat,-5)
(ugly,-3)
(unacceptable,-2)
(unappreciated,-2)
(unapproved,-2)
(unaware,-2)
(unbelievable,-1)
(unbelieving,-1)
(unbiased,2)
(uncertain,-1)
(unclear,-1)
(uncomfortable,-2)
(unconcerned,-2)
(unconfirmed,-1)
(unconvinced,-1)
(uncredited,-1)
(undecided,-1)
(underestimate,-1)
(underestimated,-1)
(underestimates,-1)
(underestimating,-1)
(undermine,-2)
(undermined,-2)
(undermines,-2)
(undermining,-2)
(undeserving,-2)
(undesirable,-2)
(uneasy,-2)
(unemployment,-2)
(unequal,-1)
```

Step 7-Now, let’s perform a map side join by joining the **tokens** statement and the dictionary contents using this command:
word_rating = join tokens by word left outer, dictionary by word using 'replicated';



```
grunt> word_rating = join tokens by word left outer, dictionary by word using 'replicated';
2016-01-20 17:02:51,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:02:51,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:02:51,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:02:51,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:02:51,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

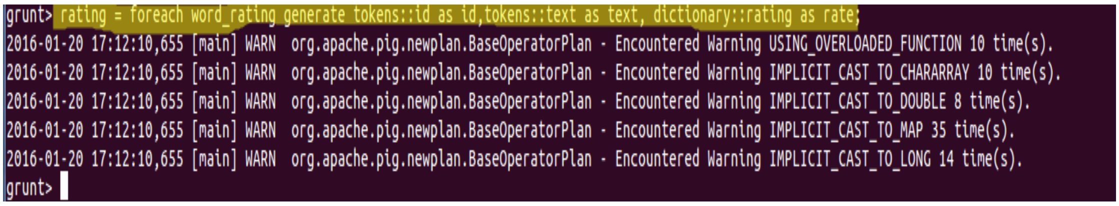
We can see the schema of the statement after performing join operation by using the below command:
describe word_rating;



```
runt> describe word_rating;
016-01-20 17:06:11,823 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
016-01-20 17:06:11,823 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
016-01-20 17:06:11,823 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 2 time(s).
ord_rating: {tokens::id: bytearray,tokens::text: bytearray,tokens::word: chararray,dictionary::word: chararray,dictionary::rating: int}
runt>
```

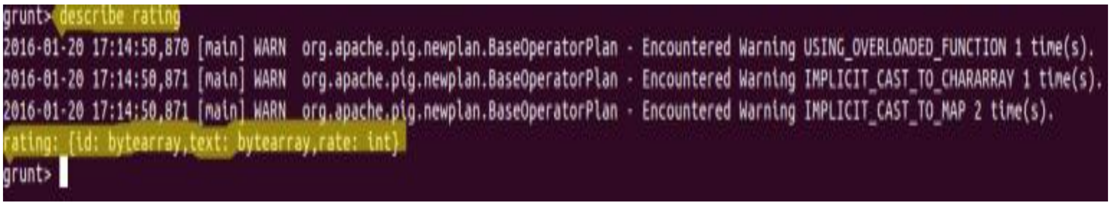
In the above screenshot, we can see that the word_rating has joined the **tokens(**consists of id, tweet text, word**)** statement and the **dictionary**(consists of word, rating).

STEP 8-Now we will extract the **id,tweet text** and **word rating(**from the dictionary**)** by using the below relation:
rating = foreach word_rating generate tokens::id as id,tokens::text as text, dictionary::rating as rate;



```
grunt> rating = foreach word_rating generate tokens::id as id,tokens::text as text, dictionary::rating as rate;
2016-01-20 17:12:10,655 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:12:10,655 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:12:10,655 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:12:10,655 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:12:10,655 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

We can now see the schema of the relation **rating** by using the command describe rating

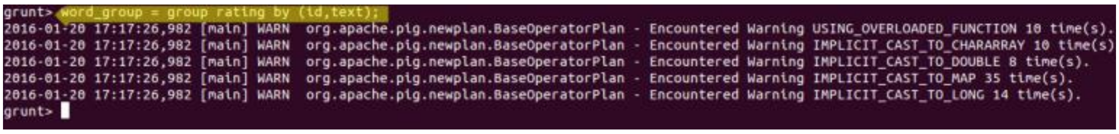


```
grunt> describe rating
2016-01-20 17:14:50,870 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
2016-01-20 17:14:50,871 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2016-01-20 17:14:50,871 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 2 time(s).
rating: {id: bytearray,text: bytearray,rate: int}
grunt>
```

In the above screen shot we can see that our relation now consists of **id,tweet text** and **rate(**for each word**).**

Step9-Now, we will group the **rating of all the words in a tweet** by using the below relation:

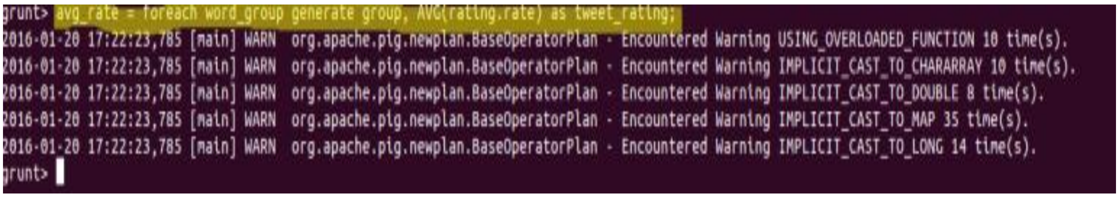
```
word_group = group rating by (id,text);
```



```
grunt> word_group = group rating by (id,text);
2016-01-20 17:17:26,982 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:17:26,982 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s)
2016-01-20 17:17:26,982 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:17:26,982 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:17:26,982 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt> 
```

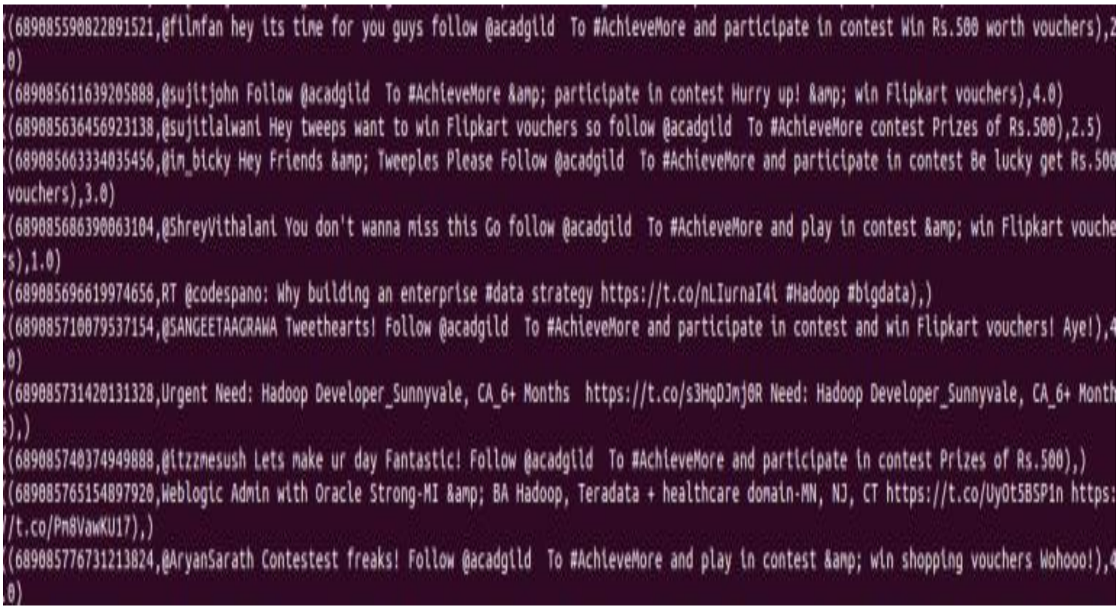
STEP 10-Now, let's perform the **Average** operation on the **rating of the words per each tweet**.

```
avg_rate = foreach word_group generate group, AVG (rating.rate) as tweet_rating;
```



```
grunt> avg_rate = foreach word_group generate group, AVG(rating.rate) as tweet_rating;
2016-01-20 17:22:23,785 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:22:23,785 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:22:23,785 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:22:23,785 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:22:23,785 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt> 
```

Now we have calculated the Average rating of the tweet using the rating of each word**.** You can refer to the below image for the same.



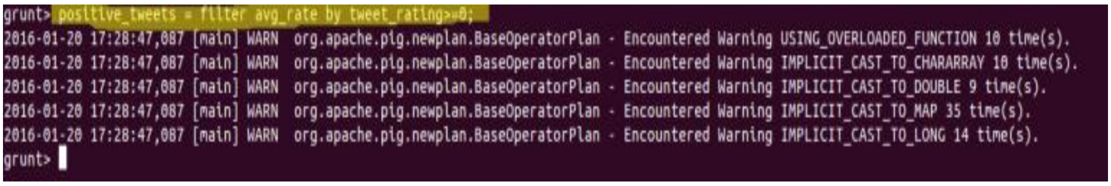
```
((689085590822891521,@filmfan hey its time for you guys follow @acadgild  To #AchieveMore and participate in contest Win Rs.500 worth vouchers),2
0)
((689085611639205888,@sujitjohn Follow @acadgild  To #AchieveMore &amp; participate in contest Hurry up! &amp; win Flipkart vouchers),4.0)
((689085636456923138,@sujitlalwani Hey tweeps want to win Flipkart vouchers so follow @acadgild  To #AchieveMore contest Prizes of Rs.500),2.5)
((689085663334035456,@im_bicky Hey Friends &amp; Tweeples Please Follow @acadgild  To #AchieveMore and participate in contest Be lucky get Rs.500
vouchers),3.0)
((689085686390063104,@ShreyVithalani You don't wanna miss this Go follow @acadgild  To #AchieveMore and play in contest &amp; win Flipkart vouche
rs),1.0)
((689085696619974656,RT @codespano: Why building an enterprise #data strategy https://t.co/nLIurnaI4l #Hadoop #bigdata),)
((689085710079537154,@SANGEETAAGRAWA Tweethearts! Follow @acadgild  To #AchieveMore and participate in contest and win Flipkart vouchers! Aye!),4
0)
((689085731420131328,Urgent Need: Hadoop Developer_Sunnyvale, CA_6+ Months  https://t.co/s3HqDJmj0R Need: Hadoop Developer_Sunnyvale, CA_6+ Month
s),)
((689085740374949888,@itzzmesush Lets make ur day Fantastic! Follow @acadgild  To #AchieveMore and participate in contest Prizes of Rs.500),)
((689085765154897920,Weblogic Admin with Oracle Strong-MI &amp; BA Hadoop, Teradata + healthcare domain-MN, NJ, CT https://t.co/UyOt5BSP1n https:
//t.co/Pm8VawKU17),)
((689085776731213824,@AryanSarath Contestest freaks! Follow @acadgild  To #AchieveMore and play in contest &amp; win shopping vouchers Wohooo!),4
0)
```

STEP 11-From the above relation, we will get all the tweets i.e., both positive and negative. Here, we can classify the positive tweets by taking the rating of the tweet which can be from **0-5.** We can classify the negative tweets by taking the rating of the tweet from **-5 to -1.**

We have now successfully performed the Sentiment Analysis on Twitter data using Pig. We now have the tweets and its rating, so let's perform an operation to filter out the positive tweets.

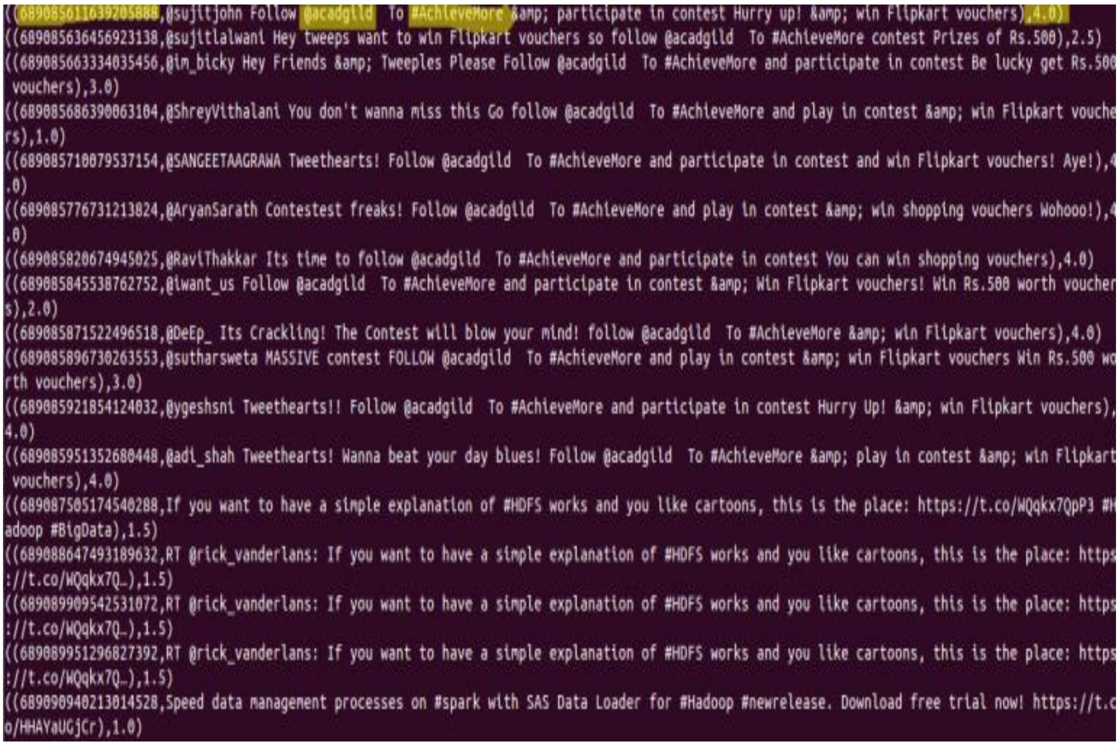
Now we will filter the positive tweets using the below statement:

positive_tweets = filter avg_rate by tweet_rating>=0;



```
grunt> positive_tweets = filter avg_rate by tweet_rating>=0;
2016-01-20 17:28:47,087 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:28:47,087 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:28:47,087 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 9 time(s).
2016-01-20 17:28:47,087 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:28:47,087 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt> 
```

We can see the positive tweets and its rating in the below screen shot.



```
((689085611639205888,@sujitjohn Follow @acadgild  To #AchieveMore &amp; participate in contest Hurry up! &amp; win Flipkart vouchers),4.0)
((689085636456923138,@sujitlalwani Hey tweeps want to win Flipkart vouchers so follow @acadgild  To #AchieveMore contest Prizes of Rs.500),2.5)
((689085663334035456,@im_bicky Hey Friends &amp; Tweeples Please Follow @acadgild  To #AchieveMore and participate in contest Be lucky get Rs.500
 vouchers),3.0)
((689085686390063104,@ShreyVithalani You don't wanna miss this Go follow @acadgild  To #AchieveMore and play in contest &amp; win Flipkart vouche
rs),1.0)
((689085710079537154,@SANGEETAAGRAWA Tweethearts! Follow @acadgild  To #AchieveMore and participate in contest and win Flipkart vouchers! Aye!),4
.0)
((689085776731213824,@AryanSarath Contestest freaks! Follow @acadgild  To #AchieveMore and play in contest &amp; win shopping vouchers Wohooo!),4
.0)
((689085820674945025,@RaviThakkar Its time to follow @acadgild  To #AchieveMore and participate in contest You can win shopping vouchers),4.0)
((689085845538762752,@iwant_us Follow @acadgild  To #AchieveMore and participate in contest &amp; Win Flipkart vouchers! Win Rs.500 worth voucher
s),2.0)
((689085871522496518,@DeEp_ Its Crackling! The Contest will blow your mind! follow @acadgild  To #AchieveMore &amp; win Flipkart vouchers),4.0)
((689085896730263553,@sutharsweta MASSIVE contest FOLLOW @acadgild  To #AchieveMore and play in contest &amp; win Flipkart vouchers Win Rs.500 wo
rth vouchers),3.0)
((689085921854124032,@ygeshsni Tweethearts!! Follow @acadgild  To #AchieveMore and participate in contest Hurry Up! &amp; win Flipkart vouchers),
4.0)
((689085951352680448,@adi_shah Tweethearts! Wanna beat your day blues! Follow @acadgild  To #AchieveMore &amp; play in contest &amp; win Flipkart
 vouchers),4.0)
((689087505174540288,If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https://t.co/WQqkx7QpP3 #H
adoop #BigData),1.5)
((689088647493189632,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https
://t.co/WQqkx7Q…),1.5)
((689089909542531072,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https
://t.co/WQqkx7Q…),1.5)
((689089951296827392,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https
://t.co/WQqkx7Q…),1.5)
((689090940213014528,Speed data management processes on #spark with SAS Data Loader for #Hadoop #newrelease. Download free trial now! https://t.c
o/HHAYaUGjCr),1.0)
```

In the above screen shot we can see the tweet_id,tweet_text and its rating.

POSSIBLE FUTURE WORK

1. Opinion mining can majorly help in discovering hot search keywords. This feature can help the brand in their SEO (Search Engine Optimization). This means that opinion mining will help them make strategies about, how their brand will come up among the top results, when a trending or hot keyword is searched in a search engine.

2. It can be used to give your business valuable insights into how people feel about your product brand or service.

3. We can predict pre elections polls and exit polls. This will help in checking out the popularity of the leaders among the peoples of a country before the elections.

4. This project will impose a barrier on putting a vulgar or abusive comment on any social media platform. It will help the developers to keep an eye on it.

5. The benefits of sentiment analysis extend into your bottom line. With sentiment analysis, you can identify a dissatisfied customer as and when they're chatting with your team. This enables your agents to offer a smooth service and quick resolution to appease, and ultimately retain, the customer.

6. Sentiment analysis also means you'll be able to detect changes in the overall opinion towards your brand. Because it provides insight into the way your customers are feeling when they approach you, you can monitor trends and see if overall opinion towards your company drops or rises.

7. When a company releases a new product or service, it is released as a pilot or beta version. The monitoring of public feedback at this stage is very crucial. So, text mining from social media platforms and review sections greatly helps accelerate this process.

CONCLUSION

We presented results for sentiment analysis on Twitter. We report an overall accuracy for 3-way classification tasks: positive versus negative versus neutral. We presented a comprehensive set of experiments for two level of classification: message level and phrase level on manually annotated data that is a random sample of stream of tweets. We investigated two kinds of models: Baseline and Feature Based Models and demonstrate that combination of both these models perform the best. For our feature-based approach, we do feature analysis which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags. In future work, we will explore even richer linguistic analysis, for example, parsing, semantic analysis and topic modelling.

BIBLIOGRAPHY

Big is next – Anand
https://bigishere.wordpress.com/

Geeks for geeks
https://www.geeksforgeeks.org/

Apache Pig
https://pig.apache.org/docs/r0.15.0/

Apache Hadoop
https://hadoop.apache.org/docs/stable/

Apache Flumes
https://flume.apache.org/documentation.html