CS590RA- Real time bigdata analytics

Project Phase-I

By

L V Sandeep Bathina  
Rakesh Vistarakula  
Balaji Gannavarapu  
Lakshmana Dheeraj Polisetty

**Framework Specification:**

1. **System Architecture Diagram:**

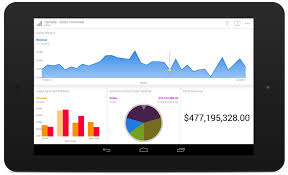


**Collect the movie viewer’s comments on twitter**

**Push the continuous tweets data to Storm**

**Analyze the tweets sent through kafka**

**Store the analyzed data in the form of rows and columns. Display on user Interface**



Graphical representation of the analyzed data

1. **Domain Model:**

**Data sources:**

Here we are using twitter dataset by using twitter4j jar file where we get the tweets in a real time basis. We capture this data and store them in text file format.

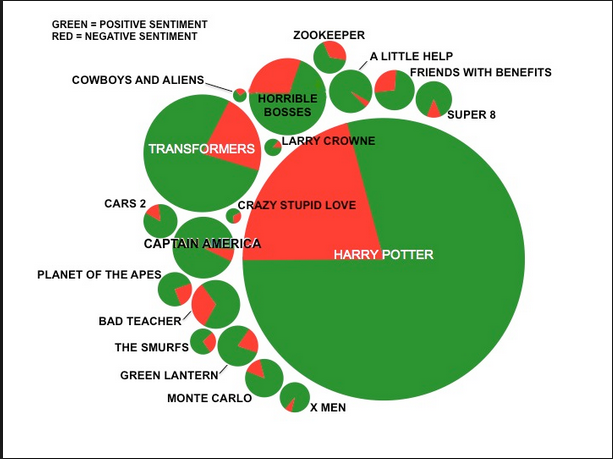
**Sentiment Analysis model and algorithm:**

The algorithm we used here is naïve Bayes classification algorithm. The approach followed here is by using weka tool.

**Analytic tools:**

The tool we are going to use here is WEKA Tool for machine learning algorithms.

**Expected Inputs/outputs:**



1. **Application specification:**

**System specification:**

The system needs the following

**Cloudera**: cloudera CDH 5.0

**Storm**: Storm 0.9.2-incubating

**Kafka**: Apache Kafka 0.8.1

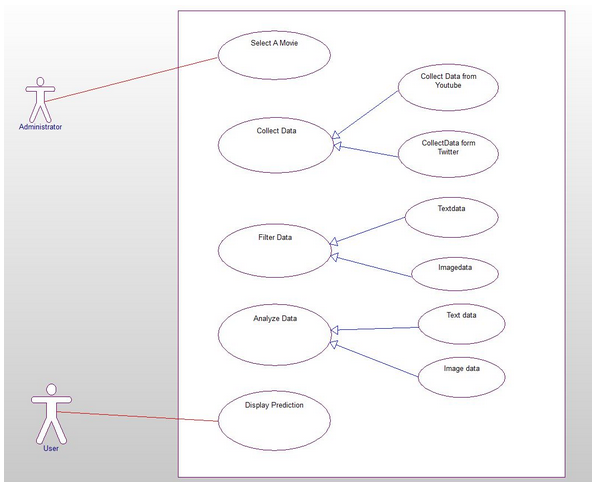
**Tools**: Weka

**Data store**: Hbase

**Existing Applications used**:

Habakkuk project which gets the tweets of bible scripture.

**Use Case Diagram:**



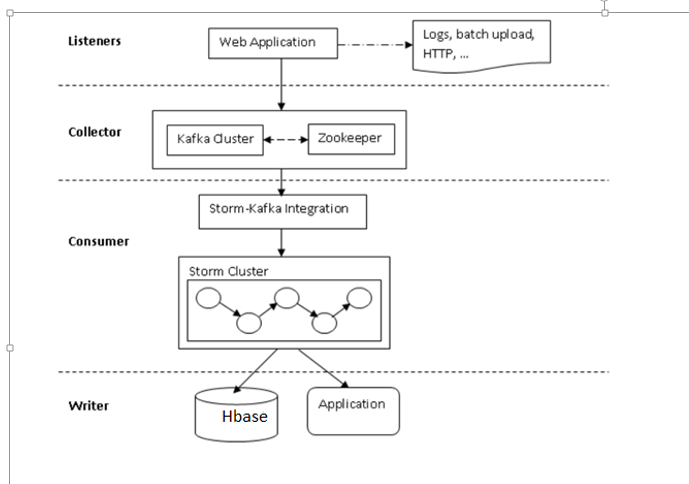
**Design of mobile client:**

**Technologies**: Android sdk, REST Web services

**GUI**: ADT Eclipse

1. **Implementation:**

**Data Model:**



1. Listener:

The raw activity data has to be processed as a stream which is consumed by the storm.

1. Collector:

Kafka here is a messaging system which acts as pipeline between storm and data processing.

1. Consumer:

All data processed by collector consists of stream of data and some associated metadata.

1. Writer:

The standard data writer writes to applications and databases like hbase and file systems.

**Sentiment analysis and algorithms:**

Naive Bayes analysis algorithm: The system is mainly based on naïve-Bayes classifier for detecting polarity of the tweets. The best performance here is calculated based on the polarity categories positive and negative. Evaluation results shows good performance when it is used to detect four sentiment categories.

**Predictive algorithms:**

Not yet used.

**Mobile user interface:**

The interface we developed here is by using Android application.

1. Initially a button is created and when the button gets clicked, all the tweets get displayed over here with their tweet ids.
2. The analysis button is created and when the user clicks on the button the sentiment analysis part of each tweet is calculated and displayed over there.
3. The next button created over here is the visualization button which displays the pie chart of analysis part and number of tweets of those are displayed here.

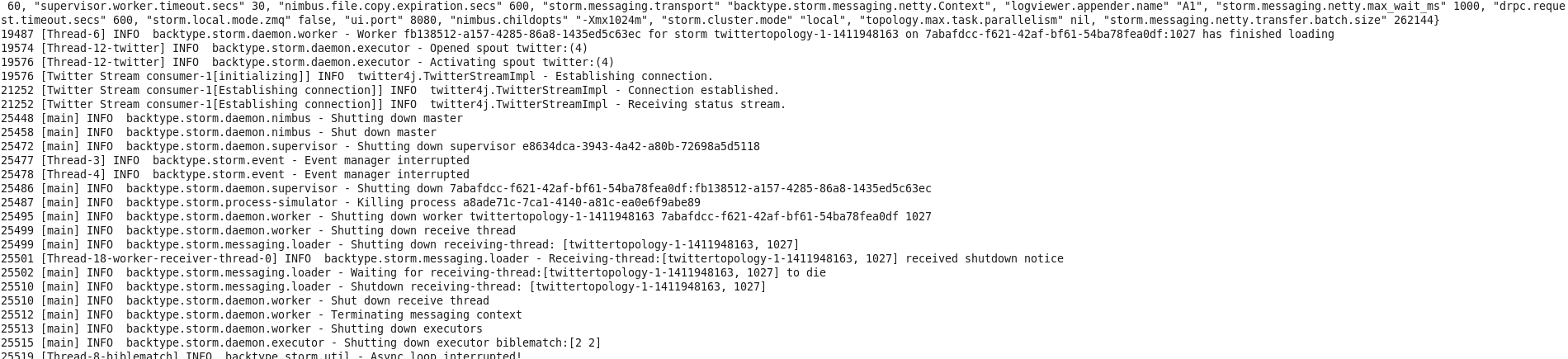
1. **Documentation:**

Outputs:

Strom Local:

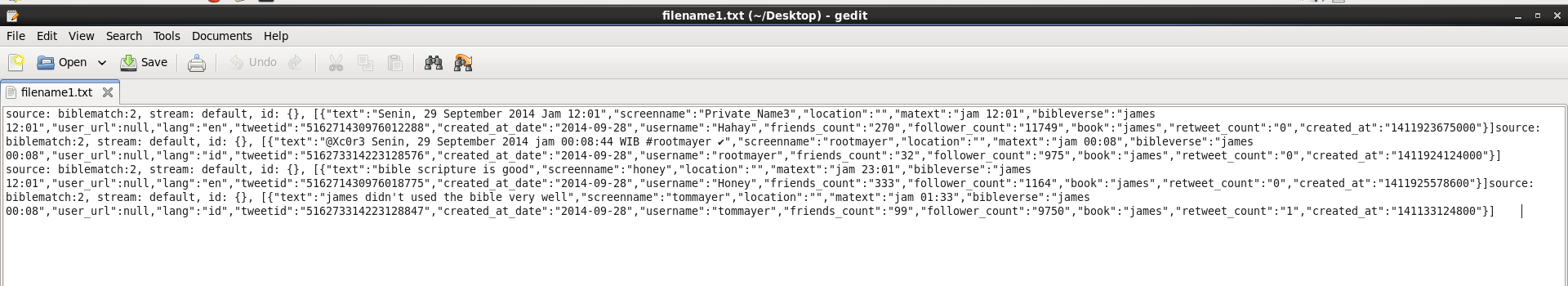
1. Initially install storm , Kafka, zookeeper and hbase
2. Run the below command in storm
3. Before this, modify the Habakkuk. Properties file and provide your twitter id and keys. This is required to generate the live streaming twitter tweets.
4. The tweets gets displayed on the console.





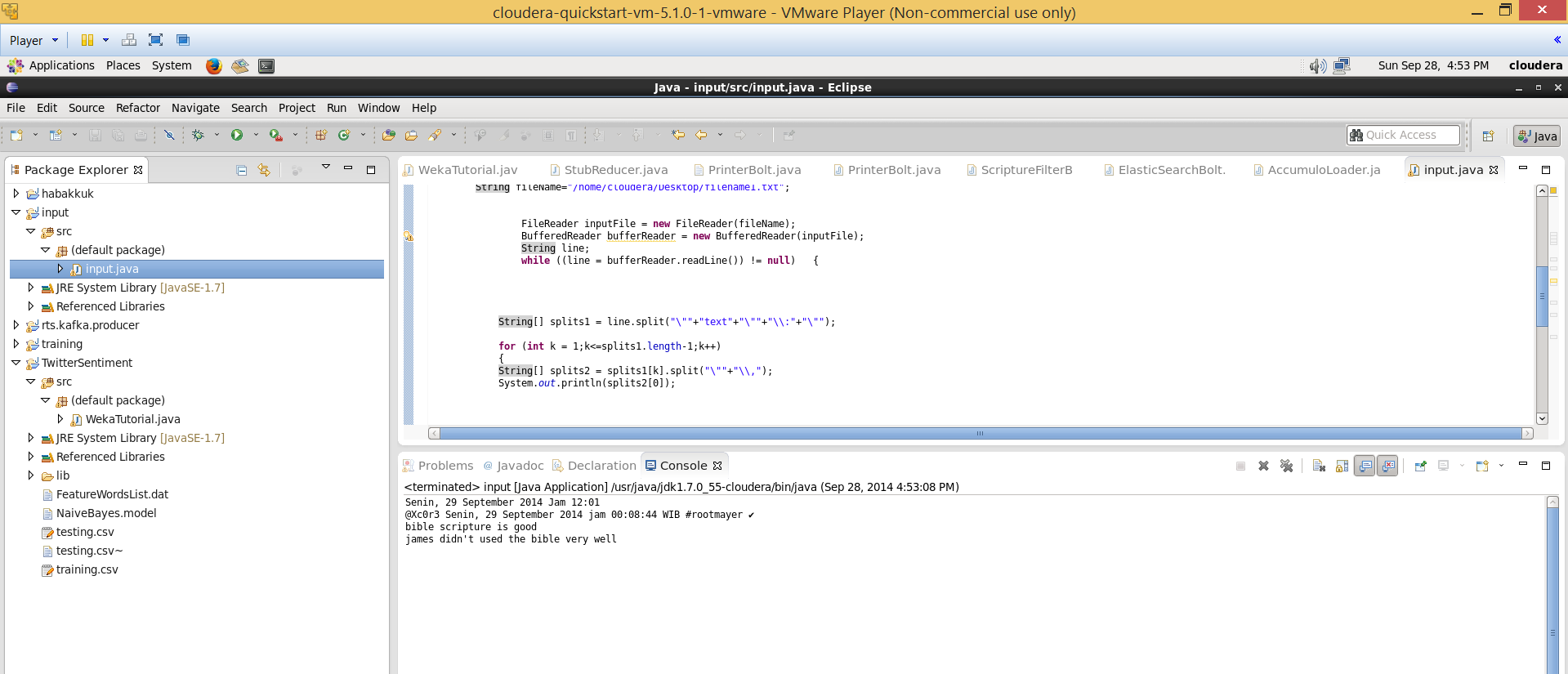
Generated real time streaming output is sent to a file

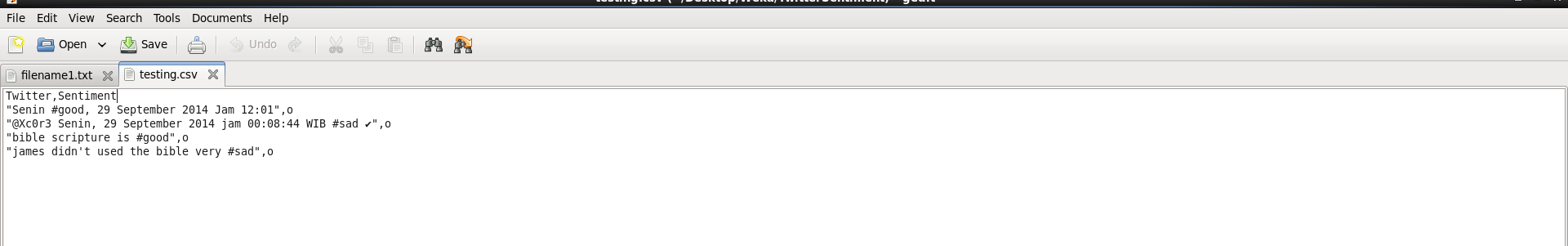
1. Modify the file printerbolt.java
2. Write the file writer code to extract the tuples of tweets to a file.
3. Provide the exact path in the code so that this can be used to analyze the tweets



Now the tweet text messages are collected and sent to the csv file

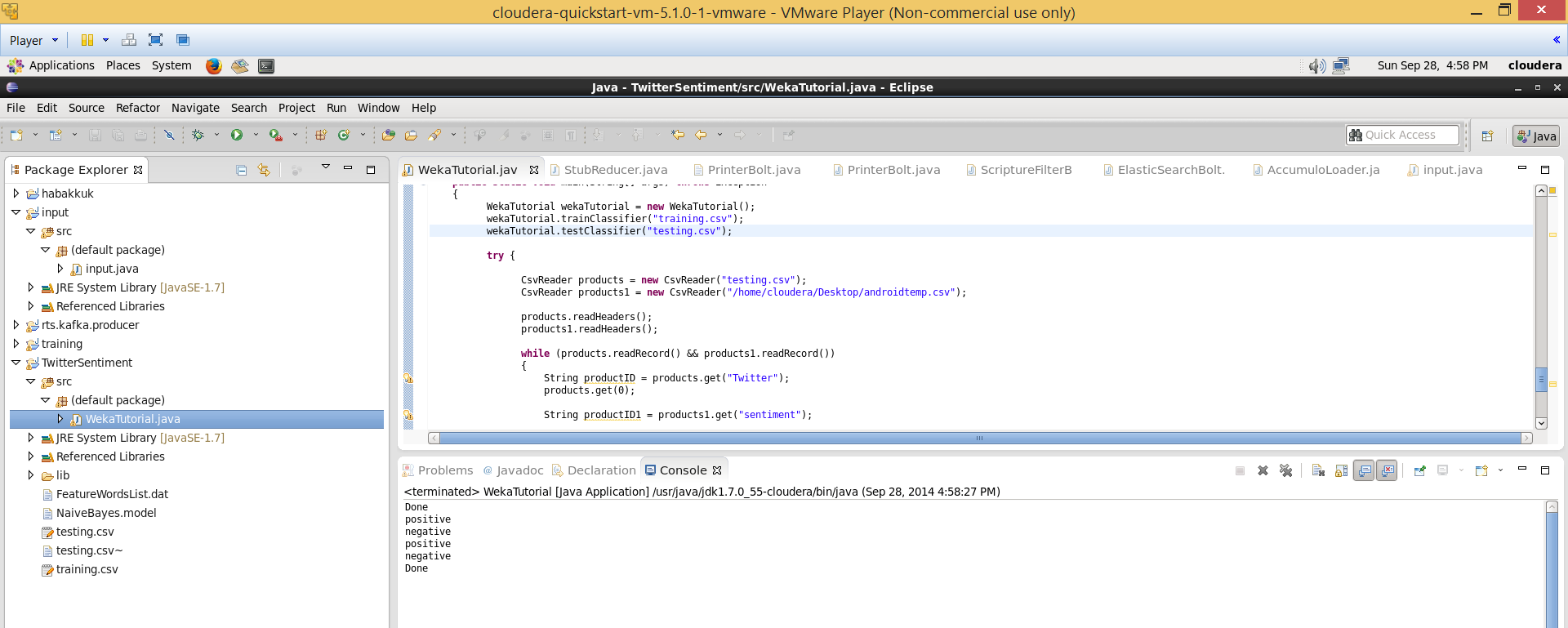
1. In order to analyze the tweets, we need to convert text file of tweets to csv format.
2. We need to append any special character at the end of each tweet.
3. This format must be followed in order to analyze data with training data.





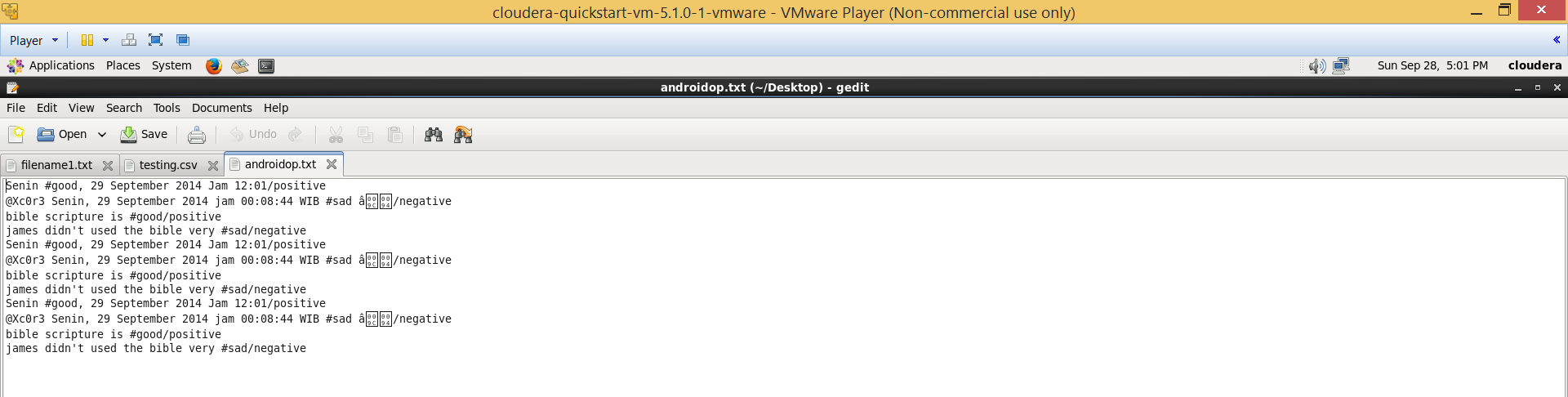
Now run the Analysis algorithm over this generated data

1. The generated csv file is fed to the twitter sentiment code
2. Here we uses weka jar to use the naïve bayes algorithm so that positive negative tweets can be identified over this.
3. Each tweet is analyzed and displayed on the console.



The tweets analysis is displayed here on the console and saved on to a file

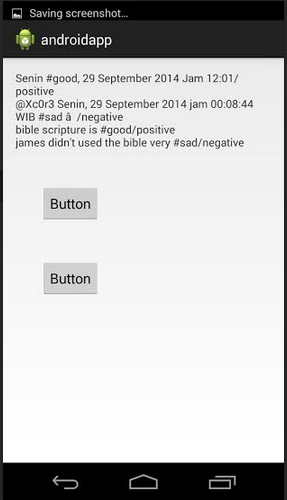
1. Apart from his, he analyzed part is stored in to a file along with their corresponding tweets.
2. This file can be used for showing analysis and visualization to the user by using android application .

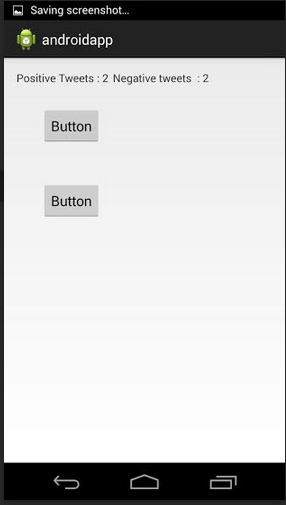


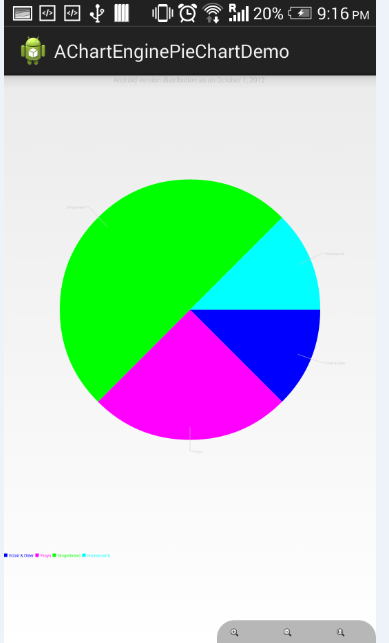
**Android:**

Now run the android app on the generated tweets

1. Initially create two buttons
2. First button displays tweet messages
3. Second button displays no.of positive and negative tweet analysis report.
4. Along with this the visualization part of pie chart is shown displaying percentage of positive and negative tweets.







1. **Project Management:**

**Design by algorithm:**

The algorithm which we going to use here is Naïve Bayes classification algorithm.

Naïve Bayes Classification algorithm:

Naïve Bayes classifiers follows simple probabilistic classifiers by applying Bayes theorem.

Advantage of naïve Bayes classifier is that it requires small amount of training data to estimate the parameters which are necessary for classification.

**Build by features:**

Implementation:

1. The scenario followed here is, initially we retrieve the tweets from the twitter by searching with a specific movie name. This step is followed by running storm command
2. We save the tweets which we got and pass this to the Kafka producer.
3. The Kafka producer sends these tweets stream to the twitter analysis part.
4. The twitter analysis steps we follow here is by using weka tool.
5. The analysis part is now send to the hbase and gets stored over there.
6. The tweets are retrieved from the hbase and displayed to the user by using visualizations and other.

**Status Report:**

**Work Completed:**

The Tweets are getting displayed for the specific movie name and those are getting stored in hbase and the analysis part on the tweets has also been done.

**Responsibility:**

Sandeep/Rakesh: Twitter tweets retrieval part and analysis part

Balaji/Dheeraj: Installations of storm and Kafka, Hbase storage and also user interface part.

**Work to be completed:**

In the next phase, the image analysis part has to be done and also the user interface has to be developed more efficiently.

**Issues:**

The twitter tweets are getting displayed only on local storm but not in storm remote.

**Bibliography**:

<http://blog.csdn.net/hljlzc2007/article/details/13275441>

<http://gramatica.usc.es/~gamallo/artigos-web/TASS2013.pdf>

<https://www.google.com/search?hl=en&site=imghp&tbm=isch&source=hp&biw=1536&bih=755&q=twitter+logo&oq=twitter+l&gs_l=img.3.0.0l10.1103.3765.0.4721.9.8.0.1.1.0.253.934.5j2j1.8.0....0...1ac.1.54.img..0.9.933.IpKE-td4Uzc#facrc=_&imgdii=_&imgrc=Z60PgD9iMKtm5M%253A%3BuY8NrGIitp5hnM%3Bhttp%253A%252F%252Floopele.com%252Fwp-content%252Fuploads%252F2014%252F07%252Ftwitter-logo.jpg%3Bhttp%253A%252F%252Floopele.com%252Ftwitter-logo%252F%3B1024%3B831>

<http://www.csvreader.com/java_csv_samples.php>

<http://stackoverflow.com/questions/2346713/csv-row-reader-question>

<https://blackboard.umkc.edu/webapps/blackboard/content/listContent.jsp?course_id=_114034_1&content_id=_1371866_1&mode=reset>

<http://www.java2s.com/Code/Jar/j/Downloadjavacsvjar.htm>