Sentiment Analysis using Apache Pig

1 Problem description

In this project we tried to do sentiment analysis of the tweets of different people on Twitter. We mainly used Apache Pig to analyze the data.

2 Tools

Hadoop | Apache Pig | Apache Zeppelin

3 Data

We have collected a dataset where twitter tweets are gathered in CSV format. We have used a static dataset for a specific time period. As we have the complete dataset, we have loaded the dataset in Pig using PigStorage.

4 Methodology and algorithm

First, we loaded the data into PigStorage. Then we dumped and divided the tweet texts to words. We took help of the dictionaries like AFINN to rate those words depending on their meaning. Then we compared each word with the dictionary and rated them from -5 to +5 depending on their meanings. Then map side join is performed to join tokens statement (words) and dictionary contents. Schema of the statement could be seen at this stage. Average rating of the tweet is determined by analyzing each word. It is either negative or positive. At last, we filtered out the positive and negative tweets.

5 How to Run the Code

For this project, we are trying to analyze the views of different people from the Twitter tweets. We have installed Hadoop, Zeppelin and Apache Pig (as an interpreter) in a virtual machine. We have collected a **Dataset** where twitter tweets are gathered in CSV format. We have used this dataset for a specific time period. Considering this as a complete dataset, we have loaded it into Pig using pigstorage.

```
load_tweets = LOAD '/demonetization-tweets.csv' USING PigStorage(',');
dump;
```

READY

```
("","text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID","st
atusSource","screenName","retweetCount","isRetweet","retweeted")
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;#8230;",FALSE,0,NA,"2016-11-23 18:40:30",FALSE,NA,"80149565697631846
4",NA,"<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter for Android</a>","HASHTAGFA
RZIWAL",331,TRUE,FALSE)
("2","RT @Hemant_80: Did you vote on #Demonetization on Modi survey app?",FALSE,0,NA,"2016-11-23 18:40:29",
FALSE,NA,"801495654778413057",NA,"<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter
for Android</a>","PRAMODKAUSHIK9",66,TRUE,FALSE)
("3","RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair + Harvard Professor lambaste #Demonetizatio
n.)
```

After loading the data, we have used pig command ‘dump’ to verify that the dataset is successfully loaded. Metadata of the tweets are id, Text, favorite, favoriteCount, replyToSN, created, truncated, replyToSID, replyToUID, statusSource, screenName, retweetCount, isRetweet, retweeted. From columns, we have extracted id and the twitter texts.

```
extract_details = FOREACH load_tweets GENERATE $0 as id,$1 as text;
dump;
```

READY

```
("","text")
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;◆")
("2","RT @Hemant_80: Did you vote on #Demonetization on Modi survey app?")
("3","RT @roshankar: Former FinSec)
(,)
(If not for Aam Aadmi, listen to th◆")
("4","RT @ANI_news: Gurugram (Haryana): Post office employees provide cash exchange to patients in hospital
s #demonetization https://t.co/uGMxUP9◆")
("5","RT @satishacharya: Reddy Wedding! @mail_today cartoon #demonetization #ReddyWedding https://t.co/u7gL
```

After that we have divided the twitter texts into words and then we calculated the sentiment of the whole tweet -

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

READY

```
("","text",text)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;◆",RT)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;◆",@rssurjewala:)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;◆",Critical)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;amp;◆",question:)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
```

Now, we started analyzing the Sentiment for the tweet by using the words in the text. We have rated 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 following link: **AFINN dictionary**

We have loaded this dictionary into Apache Pig -

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

Then we performed a map side join by joining the tokens statement and the dictionary contents using a relation. We can see the schema of the statement after performing join operations too.

```
word_rating = join tokens by word left outer, dictionary by word using 'replicated';
dump

("","text",text,,)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",RT,,)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",@rssurjewala:,,)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",Critical,,)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",question:,,)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",Was,,)
```

We can see the schema of the statement after performing join operation -

```
describe word_rating;
dump;

word_rating: {tokens::id: bytearray,tokens::text: bytearray,tokens::word: chararray,dictionary::word: chararr
ay,dictionary::rating: int}
```

We extracted 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, dictiona
dump

("","text",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearl
y fishy and requires full disclosure &amp;◆",)
```

Now, we grouped the rating of all the words in a tweet by using the below relation:

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

READY

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

READY

```
((And the Oscar goes to ""Mr.<U+092D><U+093E><U+0935><U+0941><U+0915>"" <ed><U+00A0><U+00BD><ed><U+00B8><U+
00A9><ed><U+00A0><U+00BD><ed><U+00B8><U+00A5><ed><U+00A0><U+00BD><ed><U+00B8><U+00A2><ed><U+00A0><U+00BD><e
d><U+00B8><U+00AD>#demonetization... https://t.co/mdywoTgK3t",FALSE),)
((<ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00A0><U+00BD><e
d><U+00B7><U+00B3><ed><U+00A0><U+00BD><ed><U+00B1><U+0087><ed><U+00A0><U+00BC><ed><U+00BF><U+00BC>Vote<e
d><U+00A0><U+00BD><ed><U+00B1><U+0087><ed><U+00A0><U+00BC><ed><U+00BF><U+00BC>Here<ed><U+00A0><U+00BD><e
d><U+00B9><U+008F><ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00
A0><U+00BD><ed><U+00B7><U+00B3> https://t.co/pPYE◆",FALSE),)
((<ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00A0><U+00BD><ed><U+00B7><U+00B3><ed><U+00A0><U+00BD><e
d><U+00B7><U+00B3><ed><U+00A0><U+00BD><ed><U+00B1><U+0087><ed><U+00A0><U+00BC><ed><U+00BF><U+00BC>Vote<e
```

Here we have grouped by two constraints, id and tweet text. Then we performed the Average operation on the rating of the words per each tweet.

Now we have calculated the Average rating of the tweet using the rating of each word. From the above relation, we got 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 we performed an operation to filter out the positive tweets.

```
positive_tweets = filter avg_rate by tweet_rating>=0;
dump;
```

READY

```
(("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clear
ly fishy and requires full disclosure &amp;amp;◆"),1.0)
(("12","RT @Joydas: Question in Narendra Modi App where PM is taking feedback if people support his #DeMone
tization strategy https://t.co/pYgK8Rmg7r"),2.0)
(("19","RT @pGurus1: #Demonetization The co-operative banking sector in Kerala is as good as a tax haven. I
s Kerala a Black Money HQ? https://t.co/◆"),3.0)
(("44","@dineshgrao you have 12.5 k followers yet you are not getting enough likes for your tweets against
#demonetization. You need to introspect."),2.0)
(("50","RT @Punitspeaks: Survey result so far by @PMOIndia on #Demonetization. 5 lakh response in 24 hrs. 9
0% supports note ban. #MeraDeshBadalRaha◆"),2.0)
(("52","RT @Punitspeaks: Survey result so far by @PMOIndia on #Demonetization. 5 lakh response in 24 hrs. 9
0% supports note ban. #MeraDeshBadalRaha◆"),2.0)
```

Alike positive tweets, we also filtered the negative tweets -

```
negative_tweets = filter avg_rate by tweet_rating<0;
dump;
```

READY

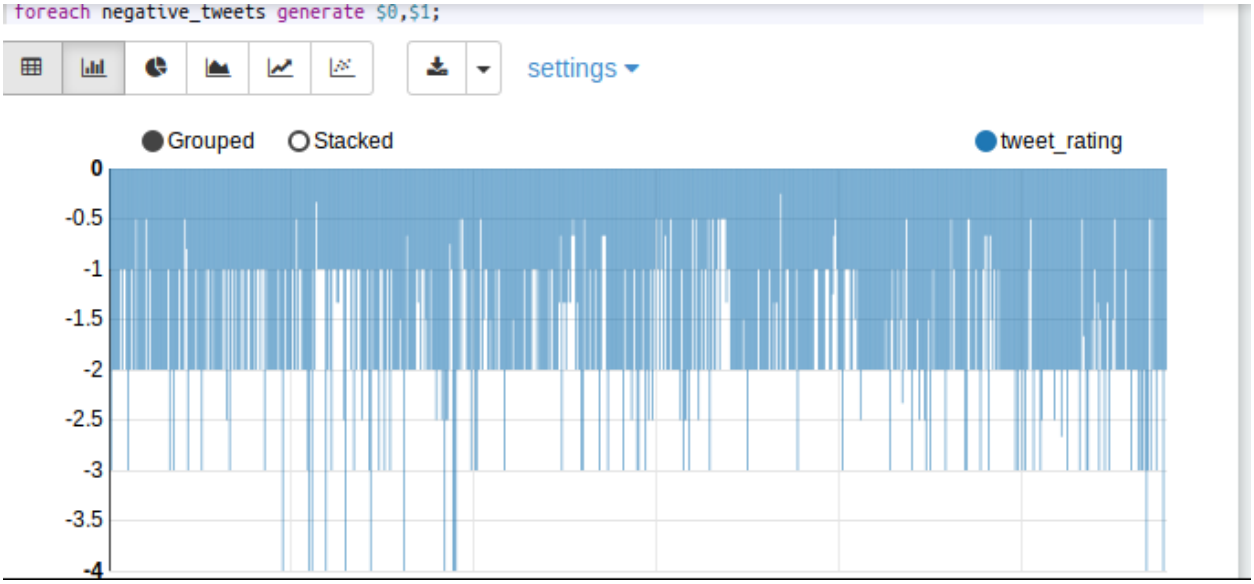
```
((So, if you really think that this #Demonetization move has struck at fake◆"),-1.0)
(("10","National reform now destroyed even the essence of sagan. Such instances urge giving #demonetization
a second though◆ https://t.co/eyySIREiUq"),-3.0)
(("13","@Jaggesh2 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00B8><U+0082>Those who  are protesting #dem
onetization  are all different party leaders."),-2.0)
(("16","RT @Dipankar_cpiml: The Modi app on #DeMonetization proves once again that the govt is totally indi
fferent to the mounting misery and hards◆"),-2.0)
(("27","RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the same has
rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> ""acquired funds"" No wo◆"),-2.0)
(("29","RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the same has
```

calhost:8080/#/ <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> ""acquired funds"" No wo◆"),-2.0)

Pig Query has been used to visualize the data in Apache Zeppelin.

Negative tweets has been visualized in this bar chart depending on their negativity (determined by dictionary)

```
foreach negative_tweets generate $0,$1;
```



Positive tweets has been visualized in this **line chart** depending on their positivity (determined by dictionary).

Following image shows the that we could export these information for further analyses -