5350 Assignment 1

Raghu Kothapalli

11336826

Sentiment analysis

Process of computationally identifying and categorizing opinions from a piece of text, and determine whether the writer's attitude towards a particular topic or the product, is positive, negative, or neutral.

For instance, this sentiment analysis is used to know whether a product is doing well in the market or not.

How it works?

Step 1: Tokenization

Step 2: Cleaning the data

Step 3: Analyzing the data

Step 4: Classification (Classify them as positive, negative, and neutral)

First, we ingest the data into hdfs and once all tweets stored, we will analyze all the tools using hive.

Task: 1

Correctly process and store the files in Hive. All tables created for the solution must have your student_id as a prefix to table name. For example, if I were to store the dictionary table, I would name it dictionary_ks0776 (5 points)

- Processing JSON data in hive using get_json_object
- Create a database called twitter_sentiment_analysis.

```
hive> create database twitter_sentiment_analysis;
OK
Time taken: 0.782 seconds

hive> use twitter_sentiment_analysis;
OK
Time taken: 0.034 seconds
```

- Copied twitter.json file to VM desktop.

- Created single column table in hive using following command.

```
create table dictionary_rk0426 (line string);
```

```
hive> create table dictionary_rk0426(json string);
OK
Time taken: 0.097 seconds
```

- Now, loaded data from desktop into the dictionary_rk0426 table which is already created.

```
load data local inpath 'Desktop/Twitter.json' overwrite into
table dictionary_rk0426;
```

```
hive> load data local inpath 'Desktop/Twitter.json' overwrite into table dictionary_rk0426;
Loading data to table twitter_sentiment_analysis.dictionary_rk0426
Table twitter_sentiment_analysis.dictionary_rk0426 stats: [numFiles=1, numRows=0, totalSize=443237, rawDataSize=0]
OK
Time taken: 0.586 seconds
```

- Now, separated the column using GET_JSON_OBJECT function and loaded the data into a different table named dictionary_rk0426_out.

```
create table dictionary_rk0426_out as
select get_json_object(json,'$.retweet_count') as retweet_count,
get_json_object(json,'$.created_at') as created_at,
get_json_object(json,'$.text') as text,
get_json_object(json,'$.id') as id,
get_json_object(json,'$.source') as source,
get_json_object(json,'$.in_reply_to_screen_name') as in_reply_to_screen_name,
get_json_object(json,'$.user.location') as u_location,
get_json_object(json,'$.user.id') as u_id,
get_json_object(json,'$.user.id_str') as u_id_str,
get_json_object(json,'$.user.name') as u_name,
get_json_object(json,'$.user.screen_name') as u_screen_name,
get_json_object(json,'$.user.geo_enabled') as u_geo_enabled,
get_json_object(json,'$.user.lang') as u_lang,
get_json_object(json,'$.user.protected') as u_protected,
get_json_object(json,'$.user.verified') as u_verified,
get_json_object(json,'$.user.followers_count') as u_followers_count,
get_json_object(json,'$.user.friends_count') as u_friends_count,
get_json_object(json,'$.user.listed_count') as u_listed_count,
get_json_object(json,'$.user.favourites_count') as u_favourites_count,
get_json_object(json,'$.user.statuses_count') as u_statuses_count,
get_json_object(json,'$.user.profile_background_color') as
u_profile_background_color,
get_json_object(json,'$.contributors') as contributors,
get_json_object(json,'$.is_quote_status') as is_quote_status,
get_json_object(json,'$.entities.user_mentions.screen_name') as e_screen_name,
```

```
get_json_object(json,'$.entities.user_mentions.name') as e_name,
get_json_object(json,'$.entities.user_mentions.id') as e_id
from dictionary_rk0426;
```

- Above they were stored as string, now we create the table first then store them with respective to their datatype.

```
create table dictionary_rk0426_final( retweet_count int,
created_at string, text string, id int, source string,
in_reply_to_screen_name string, u_location string, u_id int,
u_id_str string, u_name string, u_screen_name string,
u_geo_enabled boolean, u_lang string, u_protected boolean,
u_verified boolean, u_followers_count int, u_friends_count int,
u_listed_count int, u_favourites_count int, u_statuses_count int,
u_profile_background_color string, contributors string,
is_quote_status boolean, e_screen_name string, e_name string,
e_id int);
```

```
insert into table dictionary_rk0426_final  select
get_json_object(json,'$.retweet_count') as retweet_count,
get_json_object(json,'$.created_at') as created_at,
get_json_object(json,'$.text') as text,
get_json_object(json,'$.id') as id,
get_json_object(json,'$.source') as source,
get_json_object(json,'$.in_reply_to_screen_name') as
in_reply_to_screen_name, get_json_object(json,'$.user.location')
as u_location, get_json_object(json,'$.user.id') as u_id,
get_json_object(json,'$.user.id_str') as u_id_str,
get_json_object(json,'$.user.name') as u_name,
get_json_object(json,'$.user.screen_name') as u_screen_name,
get_json_object(json,'$.user.geo_enabled') as u_geo_enabled,
get_json_object(json,'$.user.lang') as u_lang,
get_json_object(json,'$.user.protected') as u_protected,
get_json_object(json,'$.user.verified') as u_verified,
get_json_object(json,'$.user.followers_count') as
u_followers_count, get_json_object(json,'$.user.friends_count')
as u_friends_count,get_json_object(json,'$.user.listed_count') as
u_listed_count, get_json_object(json,'$.user.favourites_count')
as u_favourites_count,
get_json_object(json,'$.user.statuses_count') as
u_statuses_count,
get_json_object(json,'$.user.profile_background_color') as
u_profile_background_color,
get_json_object(json,'$.contributors') as contributors,
get_json_object(json,'$.is_quote_status') as is_quote_status,
get_json_object(json,'$.entities.user_mentions.screen_name') as
e_screen_name,
```

```
get_json_object(json,'$.entities.user_mentions.name') as e_name,
get_json_object(json,'$.entities.user_mentions.id') as e_id from
dictionary_rk0426;
```

```
hive> insert into table dictionary_rk0426_final  select get_json_object(line,'$.
retweet_count') as retweet_count, get_json_object(line,'$.created_at') as create
d_at, get_json_object(line,'$.text') as text, get_json_object(line,'$.id') as id
, get_json_object(line,'$.source') as source, get_json_object(line,'$.in_reply_t
o_screen_name') as in_reply_to_screen_name, get_json_object(line,'$.user.locatio
n') as u_location, get_json_object(line,'$.user.id') as u_id, get_json_object(li
ne,'$.user.id_str') as u_id_str, get_json_object(line,'$.user.name') as u_name,
get_json_object(line,'$.user.screen_name') as u_screen_name, get_json_object(lin
e,'$.user.geo_enabled') as u_geo_enabled, get_json_object(line,'$.user.lang') as
 u_lang, get_json_object(line,'$.user.protected') as u_protected, get_json_objec
t(line,'$.user.verified') as u_verified, get_json_object(line,'$.user.followers_
count') as u_followers_count, get_json_object(line,'$.user.friends_count') as u_
friends_count,get_json_object(line,'$.user.listed_count') as u_listed_count, get
_json_object(line,'$.user.favourites_count') as u_favourites_count, get_json_obj
ect(line,'$.user.statuses_count') as u_statuses_count, get_json_object(line,'$.u
ser.profile_background_color') as u_profile_background_color, get_json_object(li
ne,'$.contributors') as contributors, get_json_object(line,'$.is_quote_status')
as is_quote_status, get_json_object(line,'$.entities.user_mentions.screen_name')
 as e_screen_name, get_json_object(line,'$.entities.user_mentions.name') as e_na
me, get_json_object(line,'$.entities.user_mentions.id') as e_id from dictionary_
rk0426;
Query ID = training_20200219101616_420dbfb8-1836-4c15-a2cd-3b30560e884d
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582122509627_0005, Tracking URL = http://localhost:8088/prox
y/application_1582122509627_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-19 10:16:45,681 Stage-1 map = 0%,  reduce = 0%
2020-02-19 10:17:01,930 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.56 se
c
MapReduce Total cumulative CPU time: 5 seconds 560 msec
Ended Job = job_1582122509627_0005
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dictionary_rk0426_fina
l/.hive-staging_hive_2020-02-19_10-16-29_532_593198869096274932-1/-ext-10000
Loading data to table default.dictionary_rk0426_final
Table default.dictionary_rk0426_final stats: [numFiles=1, numRows=550, totalSize
=168623, rawDataSize=168073]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 5.56 sec   HDFS Read: 450517 HDFS Write:
 168715 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 560 msec
OK
Time taken: 35.29 seconds
```

```
hive> desc dictionary_rk0426_final;
OK
retweet_count           int
created_at              timestamp
text                    string
id                      int
source                  string
in_reply_to_screen_name string
u_location              string
u_id                    int
u_id_str                string
u_name                  string
u_screen_name           string
u_geo_enabled           boolean
u_lang                  string
u_protected             boolean
u_verified              boolean
u_followers_count       int
u_friends_count         int
u_listed_count          int
u_favourites_count      int
u_statuses_count        int
u_profile_background_color      string
contributors            string
is_quote_status         boolean
e_screen_name           string
e_name                  string
e_id                    int
Time taken: 0.208 seconds, Fetched: 26 row(s)
```

- Processing the json data using hive

```
hive> select b.* from dictionary_rk0426 a lateral view json_tuple(a.line, 'retwe
et_count', 'created_at', 'text', 'id', 'source', 'reply_to_screen_name', 'u_loca
tion', 'u_id', 'u_id_str', 'u_name', 'u_screen_name', 'u_geo_enabled', 'u_lang',
 'u_protected', 'u_verified', 'u_followers_count', 'u_friends_count', 'u_listed_
count', 'u_favourites_count', 'u_statuses_count', 'u_profile_background_color',
'contributors', 'is_quote_status', 'e_screen_name', 'e_name', 'e_id') b as retwe
et_count, created_at, text, id, source, reply_to_screen_name, u_location, u_id,
u_id_str, u_name, u_screen_name, u_geo_enabled, u_lang, u_protected, u_verified,
 u_followers_count, u_friends_count, u_listed_count, u_favourites_count, u_statu
ses_count, u_profile_background_color, contributors, is_quote_status, e_screen_n
ame, e_name, e_id;
OK
```

Below image illustrates the sample result that is the first five of them.

```
7       Fri Jul 29 12:59:31 +0000 2016  It is being reported by virtually everyo
ne, and is a fact,that the media pile on against me is the worst in American pol
itical history! 641766061380228000      href="http://twitter.com" rel="nofollow"
>Twitter Web Client</a> NULL    NULL    NULL    NULL    NULL    NULL    NULL   N
ULL     NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL           F
ALSE    NULL    NULL    NULL
2       Fri Jul 29 12:59:31 +0000 2016  I am now in Texas doing a big fundraiser
 for the Republican Party and a @FoxNews Special on the BORDER and with victims
of border crime!        872283722010901000      href="http://twitter.com" rel="n
ofollow">Twitter Web Client</a> NULL    NULL    NULL    NULL    NULL    NULL   N
ULL     NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL   F
ALSE    NULL    NULL    NULL
8       Fri Jul 29 12:59:51 +0000 2016  The @WashingtonPost quickly put together
 a hit job book on me comprised of copies of some of their inaccurate stories. D
on't buy, boring!       534262624541602000      href="http://twitter.com" rel="n
ofollow">Twitter Web Client</a> NULL    NULL    NULL    NULL    NULL    NULL   N
ULL     NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL   T
RUE     NULL    NULL    NULL
6       Fri Jul 29 12:59:31 +0000 2016  .@AnnCoulter's new book, 'In Trump We Tr
ust, comes out tomorrow. People are saying it's terrific - knowing Ann I am sure
 it is! 791731982444207000      href="http://twitter.com" rel="nofollow">Twitter
 Web Client</a> NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL   N
ULL     NULL    NULL    NULL    NULL    NULL    NULL    NULL            FALSE  N
ULL     NULL    NULL
0       Sat Sep 10 22:23:38 +0000 2011  Just leaving Akron, Ohio, after a packed
 rally. Amazing people! Going now to Texas.     516680352480685000      href="ht
```

- Creating one more table for date format

```
create table dictionary_rk0426_final1( retweet_count int,
created_at string, text string, id int, source string,
in_reply_to_screen_name string, u_location string, u_id int,
u_id_str string, u_name string, u_screen_name string,
u_geo_enabled boolean, u_lang string, u_protected boolean,
u_verified boolean, u_followers_count int, u_friends_count int,
u_listed_count int, u_favourites_count int, u_statuses_count int,
u_profile_background_color string, contributors string,
is_quote_status boolean, e_screen_name string, e_name string,
e_id int, date_format string);
```

```
hive> create table dictionary_rk0426_final1( retweet_count int, created_at string, text string, id int, source string
, in_reply_to_screen_name string, u_location string, u_id int, u_id_str string, u_name string, u_screen_name string,
u_geo_enabled boolean, u_lang string, u_protected boolean, u_verified boolean, u_followers_count int, u_friends_count
 int, u_listed_count int, u_favourites_count int, u_statuses_count int, u_profile_background_color string, contributo
rs string, is_quote_status boolean, e_screen_name string, e_name string, e_id int, date_format string);
OK
Time taken: 0.127 seconds
```

```
hive> insert into table dictionary_rk0426_final1 select *, from_unixtime(unix_timestamp(created_at,'EEE MMM dd HH:mm:
ss Z yyyy'), 'yyyy-MM-dd') from dictionary_rk0426_out;
Query ID = training_20200220163232_fd229085-f270-4ed4-ba7e-be116aea98f6
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582122509627_0014, Tracking URL = http://localhost:8088/proxy/application_1582122509627_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-20 16:32:38,345 Stage-1 map = 0%,  reduce = 0%
2020-02-20 16:32:54,941 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.33 sec
MapReduce Total cumulative CPU time: 4 seconds 330 msec
Ended Job = job_1582122509627_0014
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_analysis.db/dictionary_rk0426_final1/.hiv
e-staging_hive_2020-02-20_16-32-22_406_7296821449830188505-1/-ext-10000
Loading data to table twitter_sentiment_analysis.dictionary_rk0426_final1
Table twitter_sentiment_analysis.dictionary_rk0426_final1 stats: [numFiles=1, numRows=550, totalSize=190073, rawDataS
ize=189523]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.33 sec   HDFS Read: 213131 HDFS Write: 190185 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 330 msec
OK
Time taken: 35.58 seconds
```

Task 2:

a. **What were the hashtags used in the file, and how many times each hashtag was used? (10 points)**

- We execute the following DML statement.

```
SELECT text, word FROM dictionary_rk0426_final LATERAL VIEW
explode(split(text,' ')) text_ex as word;
```

This explode() is function which is built in User Defined Table Generating Function which helps to break down an array into elements. In this case it breaks the text into words.

The LATERAL VIEW helps to join the output of explode() to input row (tweet) creating a result set containing n rows.

- Now, with the help of this result we can set #hashtags can be easily filtered from these words using the following command.

```
SELECT word, count(1) as wcount FROM dictionary_rk0426_final
LATERAL VIEW explode(split(text,' ')) text_ex as word
WHERE word LIKE '#%'
GROUP BY word
ORDER BY wcount DESC;
```

```
hive> SELECT word, count(1) as wcount FROM dictionary_rk0426_final LATERAL VIEW
explode(split(text,' ')) text_ex as word like '#%' GROUP BY word ORDER BY wcount
 DESC;
NoViableAltException(151@[()* loopback of 132:59: ( ( COMMA )=> COMMA identifier
 )*])
        at org.antlr.runtime.DFA.noViableAlt(DFA.java:158)
        at org.antlr.runtime.DFA.predict(DFA.java:116)
        at org.apache.hadoop.hive.ql.parse.HiveParser_FromClauseParser.lateralVi
ew(HiveParser_FromClauseParser.java:3399)
        at org.apache.hadoop.hive.ql.parse.HiveParser_FromClauseParser.fromSourc
e(HiveParser_FromClauseParser.java:3717)
        at org.apache.hadoop.hive.ql.parse.HiveParser_FromClauseParser.joinSourc
e(HiveParser_FromClauseParser.java:1829)
        at org.apache.hadoop.hive.ql.parse.HiveParser_FromClauseParser.fromClaus
e(HiveParser_FromClauseParser.java:1483)
        at org.apache.hadoop.hive.ql.parse.HiveParser.fromClause(HiveParser.java
:44538)
        at org.apache.hadoop.hive.ql.parse.HiveParser.singleSelectStatement(Hive
Parser.java:41655)
        at org.apache.hadoop.hive.ql.parse.HiveParser.selectStatement(HiveParser
.java:41340)
        at org.apache.hadoop.hive.ql.parse.HiveParser.regularBody(HiveParser.jav
a:41277)
        at org.apache.hadoop.hive.ql.parse.HiveParser.queryStatementExpressionBo
.....
.....
.....
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1582122509627_0007, Tracking URL = http://localhost:8088/prox
y/application_1582122509627_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0007
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-02-20 10:34:57,362 Stage-2 map = 0%,  reduce = 0%
2020-02-20 10:35:09,961 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.82 se
c
2020-02-20 10:35:23,816 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.26
sec
MapReduce Total cumulative CPU time: 4 seconds 260 msec
Ended Job = job_1582122509627_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.73 sec   HDFS Read: 182490
HDFS Write: 2329 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.26 sec   HDFS Read: 6660 HD
FS Write: 1101 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 990 msec
OK
```

The output gives us the hashtags used in the file, and tells us how many times each hashtag was used in a descending order.

```
#ImWithYou      23
#MAGA   18
#TrumpPence16   16
#CrookedHillary 16
#MakeAmericaGreatAgain  14
#AmericaFirst   13
#Trump2016      12
#RNCinCLE       10
#NeverHillary   5
#MakeAmericaGreatAgain! 4
#MakeAmericaSafeAgain   4
#Trump  4
#LESM   3
#Hillary        3
#tcdisrupt      2
#TrumpTrain     2
#ThrowbackThursday      2
#MAGATickets    2
#LawandOrder    2
#ICYMI: 2
#GOPConvention#AmericaFirst     2
#2A     2
#RiggedSystem   1
#TrumpMovement  1
#POTUS  1
#Obamacare,     1
#Obamacare      1
#NotoTrump      1
#NeverTrump     1
#TrumpTrain"    1
#NATO   1
#MakeAmericaWorkAgain-  1
#MakeAmericaWorkAgain#TrumpPence16      1
#MakeAmericaSafeAgain!#GOPConvention    1
#Trumpforlife   1
#UPUPUP 1
#USA    1
#MSM    1
#MAGAhttps://t.co/wnpVlq6oe4    1
#AlwaysTrump    1
#MAGA3pm        1
#VoterFraud     1
#LeadRight2016  1
#60Minutes      1
#WCS16  1
#ImWithYouhttps://t.co/Eph6qy7zyB       1
#ImWithYouVideo:        1
#ImWithYouTranscript:   1
#WakeUpAmerica  1
#2A!    1
#Hillaryforlife 1
#WheresHillary? 1
#Hate   1
#GOPConvention. 1
#1      1
#GOPConvention  1
#Frozen 1
#ElectionDay:   1
#DemsInPhilly   1
#DefendtheSecond        1
#DNC    1
#CrookedHillary!        1
#smallbiz       1
#AmericasMerkel 1
#taxplan,       1
#TeamUSA#OpeningCeremony        1
#TeamTrump      1
#TBT    1
#SCOTUS 1
#Rio2016        1
Time taken: 107.241 seconds, Fetched: 70 row(s)
```

Reference https://bennyaustin.com/2014/07/28/hashtags/

b. Identify the most trending hashtag by the day. How many times the most trending hashtag was tweeted? (10 points)
[Note: day should be in the format 'yyyy-mm-dd']

- Creating a new table to store the hashtags and their count.

```
hive> create table date_hashtag_count(date string, hashtag string, count int);
OK
Time taken: 0.147 seconds
```

```
hive> insert into table date_hashtag_count select date_format, word as hashword,count(1) from dictionary_rk0426_final
1 lateral view explode(split(text,' ')) text as word where word like "#%" group by date_format, text.word;
Query ID = training_20200220164444_d320a018-2cfd-4cac-8e75-62bf699399cc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1582122509627_0015, Tracking URL = http://localhost:8088/proxy/application_1582122509627_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-02-20 16:44:49,503 Stage-1 map = 0%,  reduce = 0%
2020-02-20 16:45:07,305 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.22 sec
2020-02-20 16:45:26,015 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.2 sec
MapReduce Total cumulative CPU time: 9 seconds 200 msec
Ended Job = job_1582122509627_0015
Loading data to table twitter_sentiment_analysis.date_hashtag_count
Table twitter_sentiment_analysis.date_hashtag_count stats: [numFiles=1, numRows=111, totalSize=2930, rawDataSize=2819
]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.2 sec   HDFS Read: 205967 HDFS Write: 3034 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 200 msec
OK
Time taken: 57.122 seconds
```

- Creating a new table to store named date_hashtag_count_rk0426 the hashtags and their count with respective to the dates and then insert the values that is the table created.

```
hive> create table date_hashtag_count_rk0426 (date string, tag string, count int);
OK
Time taken: 0.18 seconds
hive> insert into table date_hashtag_count_rk0426 select date, word as hashtag,count from date_hashtag_count lateral
view explode(split(tag, '#')) text as word;
FAILED: SemanticException [Error 10004]: Line 1:130 Invalid table alias or column reference 'tag': (possible column n
ames are: date, hashtag, count)
hive> insert into table date_hashtag_count_rk0426 select date, word as hashtag,count from date_hashtag_count lateral
view explode(split(hashtag, '#')) text as word;
Query ID = training_20200220165151_3bc17dfd-a218-47ab-8cd1-bd80844ec13c
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582122509627_0016, Tracking URL = http://localhost:8088/proxy/application_1582122509627_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-20 16:51:19,192 Stage-1 map = 0%,  reduce = 0%
2020-02-20 16:51:37,905 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.4 sec
MapReduce Total cumulative CPU time: 3 seconds 400 msec
Ended Job = job_1582122509627_0016
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_analysis.db/date_hashtag_count_rk0426/.hi
ve-staging_hive_2020-02-20_16-51-00_145_4562129815260963600-1/-ext-10000
Loading data to table twitter_sentiment_analysis.date_hashtag_count_rk0426
Table twitter_sentiment_analysis.date_hashtag_count_rk0426 stats: [numFiles=1, numRows=226, totalSize=4430, rawDataSi
ze=4204]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.4 sec   HDFS Read: 7904 HDFS Write: 4541 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 400 msec
OK
```

- The following query helps us to get the required result, that is first displays the result and then the name of the hashtag after that the count of each hashtag.

```
hive> select a.date, a.tag, a.count from date_hashtag_count_rk0426 a INNER JOIN (select date, MAX(count) AS MOS_TRE F
ROM date_hashtag_count_rk0426 where tag is not NULL GROUP BY date) b ON a.date = b.date and a.count = b.MOS_TRE where
 a.tag is not NULL;
Query ID = training_20200220170101_ed601ac3-610b-46f3-9f3f-278dc35f2197
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1582122509627_0017, Tracking URL = http://localhost:8088/proxy/application_1582122509627_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-02-20 17:02:02,399 Stage-1 map = 0%,  reduce = 0%
2020-02-20 17:02:20,399 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.73 sec
2020-02-20 17:02:38,181 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.5 sec
MapReduce Total cumulative CPU time: 7 seconds 500 msec
Ended Job = job_1582122509627_0017
Execution log at: /tmp/training/training_20200220170101_ed601ac3-610b-46f3-9f3f-278dc35f2197.log
2020-02-20 05:02:51     Starting to launch local task to process map join;      maximum memory = 1013645312
2020-02-20 05:02:55     Dump the side-table for tag: 0 with group count: 38 into file: file:/tmp/training/74e9354f-7d
98-4410-a416-7f56ec51980b/hive_2020-02-20_17-01-43_881_1141543690803803427-1/-local-10004/HashTable-Stage-4/MapJoin-m
apfile00--.hashtable
2020-02-20 05:02:55     Uploaded 1 File to: file:/tmp/training/74e9354f-7d98-4410-a416-7f56ec51980b/hive_2020-02-20_1
7-01-43_881_1141543690803803427-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile00--.hashtable (3840 bytes)
2020-02-20 05:02:55     End of local task; Time Taken: 3.5 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582122509627_0018, Tracking URL = http://localhost:8088/proxy/application_1582122509627_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582122509627_0018
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2020-02-20 17:03:14,410 Stage-4 map = 0%,  reduce = 0%
2020-02-20 17:03:29,570 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
MapReduce Total cumulative CPU time: 2 seconds 500 msec
Ended Job = job_1582122509627_0018
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.5 sec   HDFS Read: 11723 HDFS Write: 705 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 2.5 sec   HDFS Read: 6218 HDFS Write: 1457 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 0 msec
OK
```

```
2009-03-18              2
2009-03-18      CrookedHillary  2
2011-09-10              13
2011-09-10      MakeAmericaGreatAgain   13
2015-12-15              1
2015-12-15      RiggedSystem    1
2015-12-18              2
2015-12-18      CrookedHillary  2
2015-12-18              2
2015-12-18      ImWithYou       2
2015-12-22              3
2015-12-22      ImWithYou       3
2015-12-24              1
2015-12-24      MakeAmericaGreatAgain!  1
2015-12-24              1
2015-12-24      Trump2016       1
2015-12-24              1
2015-12-24      TrumpPence16    1
2016-01-19              2
2016-01-19      RNCinCLE        2
2016-01-25              1
2016-01-25      Hillary 1
2016-01-25              1
2016-01-25      MAGA    1
2016-01-29              2
2016-01-29      ImWithYou       2
2016-02-20              1
2016-02-20      MakeAmericaGreatAgain   1
2016-03-20              2
2016-03-20      CrookedHillary  2
2016-04-04              1
2016-04-04      DNC     1
2016-04-04              1
2016-04-04      DemsInPhilly    1
2016-04-04              1
2016-04-04      ImWithYou       1
2016-04-04              1
2016-04-04      NATO    1
2016-04-26              1
2016-04-26      Trumpforlife    1
2016-05-21              2
2016-05-21      MAGA    2
2016-05-24              2
2016-05-24      ImWithYou       2
2016-05-24              2
2016-05-24      Trump2016       2
2016-06-07              1
2016-06-07      MAGA    1
2016-07-08              2
2016-07-08      NeverHillary    2
2016-07-11              1
2016-07-11      CrookedHillary  1
2016-07-11              1
2016-07-11      Hillary 1
2016-07-11              1
2016-07-11      ImWithYou       1
2016-07-11              1
2016-07-11      MAGAhttps://t.co/wnpVlq6oe4      1
2016-07-11              1
2016-07-11      MakeAmericaSafeAgain    1
2016-07-11              1
2016-07-11      ThrowbackThursday       1
2016-07-11              1
2016-07-11      Trump   1
2016-07-11              1
2016-07-11      TrumpPence16    1
2016-07-11              1
2016-07-11      TrumpTrain      1
2016-07-24              2
2016-07-24      CrookedHillary  2
2016-07-29              2
2016-07-29      TrumpPence16    2
2016-08-23              2
2016-08-23      tcdisrupt       2
Time taken: 106.911 seconds, Fetched: 74 row(s)
```

c. Determine the score for each tweet that was posted? Identify whether the tweet had a positive or negative sentiment? Use the dictionary.txt file for determining the score. Note: Include the date ('yyyy-mm-dd'), tweet_id, user_name, and the score in the resulting query. (20 points)

- Now we will separate the columns using GET_JSON_OBJECT function and load the data into a different table in tdata_rk0426 from table dictionary_rk0426.

```
create table tdata__rk0426 as select get_json_object(json, '$.id') as t_id, get_json_object(json, '$.user.name') as user_name, get_json_object(json, '$.text') as text, unix_timestamp(get_json_object(json, '$.created_at'),'EEE MMM dd HH:mm:ss Z yyyy') as created_date from dictionary_rk0426;
```

```
hive> create table tdata__rk0426 as select get_json_object(json, '$.id') as t_id
, get_json_object(json, '$.user.name') as user_name, get_json_object(json, '$.te
xt') as text, unix_timestamp(get_json_object(json, '$.created_at'),'EEE MMM dd H
H:mm:ss Z yyyy') as created_date from dictionary_rk0426;
Query ID = training_20200224091818_9772ca26-8432-472b-8e42-cace9b9060f8
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582556327891_0001, Tracking URL = http://localhost:8088/prox
y/application_1582556327891_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582556327891_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-24 09:19:01,241 Stage-1 map = 0%,  reduce = 0%
2020-02-24 09:19:15,356 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.56 se
c
MapReduce Total cumulative CPU time: 3 seconds 560 msec
Ended Job = job_1582556327891_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/.hive-staging_hive_2020-02-24_09-18-35_599_2476603779911443679-1/-ext-10
001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/tdata__rk0426
Table twitter_sentiment_analysis.tdata__rk0426 stats: [numFiles=1, numRows=550,
totalSize=75933, rawDataSize=75383]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.56 sec   HDFS Read: 446804 HDFS Write:
 76033 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 560 msec
OK
Time taken: 44.22 seconds
```

- With these text columns now, we split them into words so that we can get only the hashtag words and store them in a separate table called splitter.

```
create table splitter as select t_id, user_name, from_unixtime(created_date,'yyyy-MM-dd') as created_date, word from tdata__rk0426 LATERAL VIEW explode(split(text, ' ')) text_ex as word;
```

```
hive> create table splitter as select t_id, user_name, from_unixtime(created_dat
e,'yyyy-MM-dd') as created_date, word from tdata__rk0426 LATERAL VIEW explode(sp
lit(text, ' ')) text_ex as word;
Query ID = training_20200224092626_a45a03d6-4c57-4423-bac4-d9933ba56491
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582556327891_0002, Tracking URL = http://localhost:8088/prox
y/application_1582556327891_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582556327891_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-24 09:26:50,451 Stage-1 map = 0%,  reduce = 0%
2020-02-24 09:27:05,609 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.46 se
c
MapReduce Total cumulative CPU time: 4 seconds 460 msec
Ended Job = job_1582556327891_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/.hive-staging_hive_2020-02-24_09-26-34_725_6234532225794500147-1/-ext-10
001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/splitter
Table twitter_sentiment_analysis.splitter stats: [numFiles=1, numRows=9148, tota
lSize=390509, rawDataSize=381361]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.46 sec   HDFS Read: 80898 HDFS Write:
390606 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 460 msec
OK
Time taken: 32.399 seconds
```

- Now, as we have to analyze the data that is the hashtags whether they are positive or negative we need to have the date.
- Here, we import Dictionary.txt which helps in analyzing the data.
- We first create table named dictionary_txt_rk0426 and then load data into it.

```
create table dictionary_txt_rk0426 (dictionary_txt_tag string,
dictionary_txt_score string) row format delimited fields terminated by

    > ;
```

```
hive> create table dictionary_txt_rk0426 (dictionary_txt_tag string, dictionary_
txt_score string) row format delimited fields terminated by '\t';
OK
Time taken: 0.239 seconds
hive> load  data local inpath 'Desktop/Dictionary.txt' overwrite into table dict
ionary_txt_rk0426;
Loading data to table twitter_sentiment_analysis.dictionary_txt_rk0426
Table twitter_sentiment_analysis.dictionary_txt_rk0426 stats: [numFiles=1, numRo
ws=0, totalSize=28094, rawDataSize=0]
OK
Time taken: 0.481 seconds
```

- Now we create another table named scores which stores the combined scores data of the tweets and the dictionary data that we have loaded.

```
create table scores as select a.t_id, a.user_name, a.created_date,
a.word, t.dictionary_txt_score from splitter a LEFT OUTER JOIN
dictionary_txt_rk0426 t on (a.word = t.dictionary_txt_tag);
```

```
hive> create table scores as select a.t_id, a.user_name, a.created_date, a.word,
 t.dictionary_txt_score from splitter a LEFT OUTER JOIN dictionary_txt_rk0426 t
on (a.word = t.dictionary_txt_tag);
Query ID = training_20200224094040_3f742c5f-8c07-4a0b-9226-cd74f186317b
Total jobs = 1
Execution log at: /tmp/training/training_20200224094040_3f742c5f-8c07-4a0b-9226-
cd74f186317b.log
2020-02-24 09:40:49     Starting to launch local task to process map join;     m
aximum memory = 1013645312
2020-02-24 09:40:51     Dump the side-table for tag: 1 with group count: 2477 in
to file: file:/tmp/training/96cddc85-98f0-41f3-8e73-010b4ae9587e/hive_2020-02-24
_09-40-39_989_9100682595202654735-1/-local-10003/HashTable-Stage-4/MapJoin-mapfi
le01--.hashtable
2020-02-24 09:40:51     Uploaded 1 File to: file:/tmp/training/96cddc85-98f0-41f
3-8e73-010b4ae9587e/hive_2020-02-24_09-40-39_989_9100682595202654735-1/-local-10
003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable (73293 bytes)
2020-02-24 09:40:51     End of local task; Time Taken: 2.152 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582556327891_0003, Tracking URL = http://localhost:8088/prox
y/application_1582556327891_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582556327891_0003
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2020-02-24 09:41:06,916 Stage-4 map = 0%,  reduce = 0%
2020-02-24 09:41:20,722 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 4.17 se
c
MapReduce Total cumulative CPU time: 4 seconds 170 msec
Ended Job = job_1582556327891_0003
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/scores
Table twitter_sentiment_analysis.scores stats: [numFiles=1, numRows=9148, totalS
ize=417704, rawDataSize=408556]
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1   Cumulative CPU: 4.17 sec   HDFS Read: 397174 HDFS Write:
 417799 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 170 msec
OK
Time taken: 43.229 seconds
```

- Now we evaluate the scores of each twitter id and store them by creating a new table named dictionary_score_rk0426.

```
create table dictionary_score_rk0426 as select t_id, user_name,
created_date, sum(dictionary_txt_score) as dictionary_score_rk0426
from scores GROUP BY t_id, user_name, created_date;
```

```
hive> create table dictionary_score_rk0426 as select t_id, user_name, created_da
te, sum(dictionary_txt_score) as dictionary_score_rk0426 from scores GROUP BY t_
id, user_name, created_date;
Query ID = training_20200224094646_f6c0616d-6385-4d73-b7f4-2a4354fab99d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1582556327891_0004, Tracking URL = http://localhost:8088/prox
y/application_1582556327891_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582556327891_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-02-24 09:46:22,457 Stage-1 map = 0%,  reduce = 0%
2020-02-24 09:46:36,240 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.36 se
c
2020-02-24 09:46:49,163 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.81
sec
MapReduce Total cumulative CPU time: 5 seconds 810 msec
Ended Job = job_1582556327891_0004
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_anal
ysis.db/dictionary_score_rk0426
Table twitter_sentiment_analysis.dictionary_score_rk0426 stats: [numFiles=1, num
Rows=550, totalSize=22190, rawDataSize=21640]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.81 sec   HDFS Read: 425545
HDFS Write: 22300 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 810 msec
OK
Time taken: 42.835 seconds
```

- With the help of tables created above now we analyze the sentiment analysis that is whether the tweet data is positive negative or none and store the resulting information in a new table named score_analyze_rk0426.

```
create table score_analyze_rk0426 as select t_id, user_name,
created_date, dictionary_score_rk0426, case WHEN
dictionary_score_rk0426 < 0 THEN 'NEGATIVE Sentiment' WHEN
dictionary_score_rk0426 > 0 THEN 'POSITIVE Sentiment' ELSE 'NONE'END
as score_analyze_rk0426 from dictionary_score_rk0426;
```

```
hive> create table score_analyze_rk0426 as select t_id, user_name, created_date,
 dictionary_score_rk0426, case WHEN dictionary_score_rk0426 < 0 THEN 'NEGATIVE S
entiment' WHEN dictionary_score_rk0426 > 0 THEN 'POSITIVE Sentiment' ELSE 'NONE'
END as score_analyze_rk0426 from dictionary_score_rk0426;
Query ID = training_20200224095757_7b6fc8a9-432b-431b-81c1-72254d9db062
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1582556327891_0005, Tracking URL = http://localhost:8088/prox
y/application_1582556327891_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1582556327891_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-02-24 09:57:47,878 Stage-1 map = 0%,  reduce = 0%
2020-02-24 09:58:01,583 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.1 sec
MapReduce Total cumulative CPU time: 3 seconds 100 msec
Ended Job = job_1582556327891_0005
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_analysis.db/.hive-staging_hive_2020-02-24
_09-57-32_215_5511933385433681345-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/twitter_sentiment_analysis.db/score_analyze_rk0426
Table twitter_sentiment_analysis.score_analyze_rk0426 stats: [numFiles=1, numRows=550, totalSize=28874, rawDataSize=2
8324]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.1 sec   HDFS Read: 26428 HDFS Write: 28981 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 100 msec
OK
Time taken: 30.881 seconds
```

```
select created_date, t_id, user_name, dictionary_score_rk0426,
score_analyze_rk0426 from score_analyze_rk0426;
```

- With the help of this query we can print all the columns that we need that is the date, id, username, and sentiment score.

```
hive> select created_date, t_id, user_name, dictionary_score_rk0426, score_analyze_rk0426 from score_analyze_rk0426;
OK

2016-06-07      100696141719273000      Michelle        -5.0    NEGATIVE Sentiment
2016-04-04      101270445959461000      Frances 2.0     POSITIVE Sentiment
2011-09-10      103386304747066000      Marion  9.0     POSITIVE Sentiment
2011-09-10      105743242835042000      Alex    NULL    NONE
2011-09-10      105859228650774000      Judith  NULL    NONE
2011-09-10      106128175972515000      Nicholas        -2.0    NEGATIVE Sentiment
2016-07-24      106841024367780000      Annie   NULL    NONE
2011-09-10      107921004221536000      Gene    NULL    NONE
2011-09-10      108397478405756000      Rhonda  NULL    NONE
2011-09-10      109813990267358000      Jennifer        -1.0    NEGATIVE Sentiment
2011-09-10      110624316674377000      Valerie -2.0    NEGATIVE Sentiment
2015-12-15      111442643526292000      Daniel  NULL    NONE
2011-09-10      111565792997795000      Amy     NULL    NONE
2015-12-22      113523162983355000      Marc    NULL    NONE
2011-09-10      119416231902478000      Shirley 2.0     POSITIVE Sentiment
2011-09-10      121033230143694000      Dan     -2.0    NEGATIVE Sentiment
2011-09-10      121873790298340000      Joshua  2.0     POSITIVE Sentiment
2009-03-18      124202382045715000      Lester  -4.0    NEGATIVE Sentiment
2011-09-10      127363447354159000      Jackie  NULL    NONE
2016-04-04      128176348280256000      Johnny  NULL    NONE
2016-06-07      128390143224214000      Sue     2.0     POSITIVE Sentiment
2015-12-22      128919251286755000      Dianne  1.0     POSITIVE Sentiment
2016-01-25      132527313979710000      Judith  -2.0    NEGATIVE Sentiment
2011-09-10      132658794295731000      Valerie NULL    NONE
2011-09-10      134237176213536000      Jonathan        NULL    NONE
2011-09-10      136262215022259000      Lynn    4.0     POSITIVE Sentiment
2015-12-15      137400163920091000      Alexandra       0.0     NONE
2011-09-10      138226334124559000      Kristin NULL    NONE
2011-09-10      138496495922264000      Rita    1.0     POSITIVE Sentiment
2015-12-24      144454247826644000      Seth    NULL    NONE
2016-01-19      145475220714925000      Michelle        -1.0    NEGATIVE Sentiment
2016-07-11      151582047634508000      Alexander       0.0     NONE
2016-07-11      153054238040334000      Calvin  NULL    NONE
2011-09-10      154630825816463000      April   NULL    NONE
2011-09-10      156381402331895000      Lori    -3.0    NEGATIVE Sentiment
2011-09-10      157615883181756000      Megan   NULL    NONE
```

2016-07-11	159040022728300000	Matthew	2.0	POSITIVE Sentiment
2015-12-18	159070043681009000	Marian	NULL	NONE
2016-01-25	162170071258695000	Troy	2.0	POSITIVE Sentiment
2015-12-24	162902219062145000	David	2.0	POSITIVE Sentiment
2011-09-10	164894051801320000	Clyde	7.0	POSITIVE Sentiment
2016-07-11	166009066168663000	Marshall	-3.0	NEGATIVE Sentiment
2011-09-10	167934489363447000	Heather	2.0	POSITIVE Sentiment
2016-07-24	169967281681239000	Nathan	2.0	POSITIVE Sentiment
2016-02-20	171071624081952000	Seth	NULL	NONE
2011-09-10	171571588722167000	Craig	1.0	POSITIVE Sentiment
2016-01-19	171928727133382000	Peter	-3.0	NEGATIVE Sentiment
2015-12-18	173587769001895000	Gayle	NULL	NONE
2011-09-10	174896291654366000	Dale	2.0	POSITIVE Sentiment
2016-01-25	178185921882025000	Carlos	NULL	NONE
2016-01-29	178834874759199000	Thelma	2.0	POSITIVE Sentiment
2011-09-10	179123919010692000	Tara	NULL	NONE
2015-12-22	184375879987218000	Renee	NULL	NONE
2011-09-10	184664637864848000	Linda	NULL	NONE
2011-09-10	185057215625865000	Bruce	-2.0	NEGATIVE Sentiment
2016-07-11	186477348751941000	Jimmy	1.0	POSITIVE Sentiment
2011-09-10	188805051315213000	Louis	1.0	POSITIVE Sentiment
2011-09-10	189602173921151000	Ben	NULL	NONE
2015-12-24	190838041367656000	Ken	NULL	NONE
2009-03-18	192176130274611000	Russell	-4.0	NEGATIVE Sentiment
2011-09-10	192958758219646000	Troy	2.0	POSITIVE Sentiment
2011-09-10	193927851702605000	Elsie	NULL	NONE
2016-07-11	194358557551882000	Christian	-1.0	NEGATIVE Sentiment
2016-01-29	195582399958342000	Peter	NULL	NONE
2016-04-26	198193787642777000	Bruce	-1.0	NEGATIVE Sentiment
2016-04-26	202192313966323000	Carol	NULL	NONE
2011-09-10	202312634011898000	Brandon	-5.0	NEGATIVE Sentiment
2016-01-25	202324113168205000	Ruby	NULL	NONE
2011-09-10	203391274818647000	Gail	-2.0	NEGATIVE Sentiment
2011-09-10	205831028169281000	Philip	NULL	NONE
2016-05-24	206001685287459000	Alison	NULL	NONE
2011-09-10	207317556162850000	Debra	-2.0	NEGATIVE Sentiment
2016-07-11	207877252896344000	Terri	NULL	NONE
2016-01-25	211014694431747000	Craig	NULL	NONE
2016-07-24	213614835353592000	Ronnie	4.0	POSITIVE Sentiment
2016-07-24	215124999743772000	Theresa	2.0	POSITIVE Sentiment
2016-04-26	216297210501385000	Herman	NULL	NONE
2016-07-29	217004108290494000	Ellen	NULL	NONE
2016-07-11	218896750493188000	Larry	-3.0	NEGATIVE Sentiment
2016-03-20	222343594314034000	Kelly	NULL	NONE
2011-09-10	222454051649708000	Elizabeth	4.0	POSITIVE Sentiment
2011-09-10	224101912737994000	Katherine	NULL	NONE

2016-07-24	240301354138822000	Anna	-3.0	NEGATIVE Sentiment
2016-07-29	241360499949704000	Clifford	NULL	NONE
2011-09-10	241361331445150000	Terri	3.0	POSITIVE Sentiment
2016-04-04	242787639499629000	Jessica	-2.0	NEGATIVE Sentiment
2011-09-10	243288421754373000	Gretchen	6.0	POSITIVE Sentiment
2016-06-07	243553305572198000	Michele	-3.0	NEGATIVE Sentiment
2011-09-10	246629171140842000	Thomas	NULL	NONE
2016-01-25	250077982102068000	Marshall	-3.0	NEGATIVE Sentiment
2016-03-20	252561219999538000	Stanley	-2.0	NEGATIVE Sentiment
2016-03-20	254381223127613000	Charlie	-1.0	NEGATIVE Sentiment
2011-09-10	254612845851866000	Yvonne	NULL	NONE
2011-09-10	256122741074442000	Carlos	2.0	POSITIVE Sentiment
2011-09-10	256149922478944000	Arnold	NULL	NONE
2011-09-10	256847268203662000	Nina	NULL	NONE
2011-09-10	258076332672408000	Martha	NULL	NONE
2011-09-10	259456041721387000	Shirley	3.0	POSITIVE Sentiment
2011-09-10	260041661336467000	Betsy	NULL	NONE
2011-09-10	262058165569671000	Ethel	NULL	NONE
2011-09-10	262429684879633000	Marguerite	5.0	POSITIVE Sentiment
2011-09-10	265999352650190000	Michael	-3.0	NEGATIVE Sentiment
2016-08-23	268653953230593000	George	NULL	NONE
2016-01-25	269099495447532000	Shirley	-5.0	NEGATIVE Sentiment
2011-09-10	272429712983819000	Tracy	2.0	POSITIVE Sentiment
2011-09-10	273264888595046000	Ruth	-1.0	NEGATIVE Sentiment
2011-09-10	274259281179532000	Christopher	NULL	NONE
2011-09-10	274451911902931000	Martin	-3.0	NEGATIVE Sentiment
2011-09-10	275628796345959000	Douglas	NULL	NONE
2011-09-10	276423954974353000	Rose	1.0	POSITIVE Sentiment
2015-12-22	276613818614046000	Joe	NULL	NONE
2011-09-10	278879505549078000	Jacqueline	NULL	NONE
2011-09-10	282405070431276000	Alex	-2.0	NEGATIVE Sentiment
2011-09-10	282936306789819000	Ben	NULL	NONE
2016-07-11	288084415685059000	Tina	3.0	POSITIVE Sentiment
2016-07-11	290437490003616000	Geraldine	3.0	POSITIVE Sentiment
2016-07-11	292186470720315000	Marion	-2.0	NEGATIVE Sentiment
2011-09-10	292467805382516000	Eddie	1.0	POSITIVE Sentiment
2011-09-10	294030125965307000	Andrea	NULL	NONE
2011-09-10	297748840388795000	Charlene	NULL	NONE
2011-09-10	300969903297498000	Steve	5.0	POSITIVE Sentiment
2016-05-24	302377820720290000	Jeffrey	NULL	NONE
2011-09-10	305449626035206000	Maria	-3.0	NEGATIVE Sentiment
2011-09-10	305634812075752000	Hugh	-1.0	NEGATIVE Sentiment
2011-09-10	308147925851409000	Brian	-1.0	NEGATIVE Sentiment
2015-12-24	310151220166586000	Arnold	-4.0	NEGATIVE Sentiment
2011-09-10	310789874224621000	Vickie	-2.0	NEGATIVE Sentiment
2011-09-10	313233427544621000	Gloria	NULL	NONE

2011-09-10	313233427544621000	Gloria	NULL	NONE
2011-09-10	313814562482422000	Gene	-1.0	NEGATIVE Sentiment
2009-03-18	313967997302001000	Sandra	NULL	NONE
2011-09-10	319417215505289000	Eugene	-2.0	NEGATIVE Sentiment
2016-01-29	319508781393190000	Brad	4.0	POSITIVE Sentiment
2011-09-10	319695465393128000	Cindy	NULL	NONE
2016-07-24	320144721281132000	Patrick	NULL	NONE
2011-09-10	322284287857803000	Randall	-2.0	NEGATIVE Sentiment
2011-09-10	322434901379863000	Stanley	NULL	NONE
2011-09-10	324137567448785000	April	2.0	POSITIVE Sentiment
2011-09-10	324695959464957000	Sandra	2.0	POSITIVE Sentiment
2011-09-10	324787394042523000	Ricky	NULL	NONE
2011-09-10	327955765983088000	Theresa	-2.0	NEGATIVE Sentiment
2011-09-10	328579674622915000	Glenda	NULL	NONE
2016-06-07	329223296619955000	Ronald	3.0	POSITIVE Sentiment
2015-12-24	329308654707767000	Wayne	1.0	POSITIVE Sentiment
2016-07-08	329490439231855000	Leigh	NULL	NONE
2015-12-24	329494044338539000	Esther	NULL	NONE
2011-09-10	330899242233600000	Herbert	4.0	POSITIVE Sentiment
2011-09-10	332480958832166000	Lucille	NULL	NONE
2016-07-11	335587326934988000	Shelley	NULL	NONE
2011-09-10	337016422277141000	Dean	NULL	NONE
2011-09-10	337294969259209000	Cameron	NULL	NONE
2011-09-10	337960517625769000	Wesley	-6.0	NEGATIVE Sentiment
2011-09-10	339186342910956000	Terry	3.0	POSITIVE Sentiment
2011-09-10	339334937755358000	Timothy	4.0	POSITIVE Sentiment
2015-12-18	344000670375452000	Rose	-2.0	NEGATIVE Sentiment
2011-09-10	345065552445521000	John	NULL	NONE
2016-06-07	345854651336432000	Eddie	1.0	POSITIVE Sentiment
2016-04-04	346297534285257000	Lawrence	1.0	POSITIVE Sentiment
2016-07-08	348043877046730000	Jonathan	NULL	NONE
2016-02-20	350763384924572000	Luis	NULL	NONE
2011-09-10	352489261542599000	Milton	NULL	NONE
2016-02-20	354471312387569000	Sandy	2.0	POSITIVE Sentiment
2016-01-19	356931863493133000	Sue	2.0	POSITIVE Sentiment
2011-09-10	363176182392657000	Ryan	NULL	NONE
2011-09-10	364642343822423000	Kay	NULL	NONE
2009-03-18	366912068027820000	Heather	0.0	NONE
2016-02-20	368093268772857000	Edwin	-3.0	NEGATIVE Sentiment
2011-09-10	368316954944439000	Steven	-2.0	NEGATIVE Sentiment
2016-07-29	368779777700339000	Kay	3.0	POSITIVE Sentiment
2011-09-10	369659769951160000	Marlene	NULL	NONE
2016-07-24	371137043010186000	Justin	NULL	NONE
2016-05-21	372822813791814000	Lee	NULL	NONE
2009-03-18	375611165970496000	Diane	NULL	NONE
2011-09-10	376011600443087000	Peggy	-4.0	NEGATIVE Sentiment

2016-01-25	378003117741832000	Sean	NULL	NONE
2016-04-04	380459189437637000	Diane	2.0	POSITIVE Sentiment
2011-09-10	386990559936906000	Arthur	-3.0	NEGATIVE Sentiment
2016-04-04	387684965559844000	Tammy	-2.0	NEGATIVE Sentiment
2011-09-10	389720483502088000	Bryan	-4.0	NEGATIVE Sentiment
2011-09-10	394082755444620000	Lewis	NULL	NONE
2016-05-21	394544985816104000	Jesse	2.0	POSITIVE Sentiment
2016-06-07	398182394732417000	Harvey	NULL	NONE
2016-01-29	400719186005870000	Michael	NULL	NONE
2011-09-10	400987861074127000	Jackie	1.0	POSITIVE Sentiment
2016-03-20	403057895566782000	Amanda	1.0	POSITIVE Sentiment
2011-09-10	406117734630901000	John	2.0	POSITIVE Sentiment
2016-06-07	406316224813848000	Jeanne	1.0	POSITIVE Sentiment
2016-07-08	410679544463800000	Greg	NULL	NONE
2011-09-10	411108839217448000	Charles	NULL	NONE
2011-09-10	412335317705604000	Harold	NULL	NONE
2016-07-24	413165938863990000	Ashley	-5.0	NEGATIVE Sentiment
2011-09-10	413199933923568000	Lloyd	-4.0	NEGATIVE Sentiment
2011-09-10	415428676583483000	Alison	4.0	POSITIVE Sentiment
2016-04-04	418627988413317000	Lauren	-4.0	NEGATIVE Sentiment
2011-09-10	419085311636146000	Clarence	NULL	NONE
2011-09-10	420378316299325000	Ricky	3.0	POSITIVE Sentiment
2016-07-11	420598885266397000	Andrea	NULL	NONE
2016-01-29	422284497512967000	Robin	NULL	NONE
2015-12-18	423153622532133000	Kathleen	NULL	NONE
2015-12-18	423470739242248000	Justin	NULL	NONE
2011-09-10	423522795404109000	Max	-1.0	NEGATIVE Sentiment
2011-09-10	423789358539063000	Alison	2.0	POSITIVE Sentiment
2011-09-10	425237326343322000	Louis	NULL	NONE
2015-12-22	426458092589721000	Molly	-1.0	NEGATIVE Sentiment
2011-09-10	427982246458003000	Suzanne	NULL	NONE
2011-09-10	428691317581544000	Marlene	1.0	POSITIVE Sentiment
2011-09-10	429600450217876000	Jerome	1.0	POSITIVE Sentiment
2011-09-10	431074808378375000	Wanda	12.0	POSITIVE Sentiment
2016-01-19	437885755599374000	Alfred	-3.0	NEGATIVE Sentiment
2011-09-10	438175005926512000	Cathy	NULL	NONE
2011-09-10	441400698424755000	Leon	NULL	NONE
2011-09-10	442086002508395000	Charles	-1.0	NEGATIVE Sentiment
2016-07-11	445255906427603000	Cameron	NULL	NONE
2015-12-22	446479624424538000	Renee	NULL	NONE
2016-07-08	446610161997524000	Kevin	NULL	NONE
2011-09-10	449403860041669000	Benjamin	8.0	POSITIVE Sentiment
2015-12-15	456426128959809000	Josephine	-7.0	NEGATIVE Sentiment
2016-07-11	460552602542292000	Zachary	1.0	POSITIVE Sentiment
2011-09-10	461976342977450000	Timothy	-4.0	NEGATIVE Sentiment

2011-09-10	468277315717248000	Kristin	NULL	NONE	
2016-06-07	468486546885691000	Geraldine		NULL	NONE
2016-07-11	469723244877641000	Gregory	-3.0	NEGATIVE Sentiment	
2015-12-18	472723630925884000	Stephen	-4.0	NEGATIVE Sentiment	
2011-09-10	474976644944922000	Lawrence		NULL	NONE
2016-03-20	476210282051743000	Laura	NULL	NONE	
2009-03-18	477123889205614000	Judith	4.0	POSITIVE Sentiment	
2016-07-08	477260705971257000	Ruby	NULL	NONE	
2011-09-10	479756098123419000	Joanne	NULL	NONE	
2011-09-10	479955518780594000	Heidi	-1.0	NEGATIVE Sentiment	
2009-03-18	480237525533163000	Molly	1.0	POSITIVE Sentiment	
2016-08-23	481044586874523000	Ellen	1.0	POSITIVE Sentiment	
2016-04-26	482744238810159000	Herbert	-6.0	NEGATIVE Sentiment	
2015-12-24	483358381242562000	Jackie	3.0	POSITIVE Sentiment	
2015-12-15	483916164698509000	Jesse	1.0	POSITIVE Sentiment	
2011-09-10	484096517001968000	Matthew	NULL	NONE	
2011-09-10	484175757542947000	Amy	2.0	POSITIVE Sentiment	
2011-09-10	486056666857108000	Vivian	-6.0	NEGATIVE Sentiment	
2011-09-10	486991358380482000	Sean	7.0	POSITIVE Sentiment	
2011-09-10	487304435464149000	Shirley	NULL	NONE	
2009-03-18	488845101475781000	Marion	NULL	NONE	
2016-05-21	489405407400938000	Brenda	-2.0	NEGATIVE Sentiment	
2011-09-10	489618280588232000	Bryan	NULL	NONE	
2011-09-10	489843325321979000	Michelle		2.0	POSITIVE Sentiment
2011-09-10	491901722285607000	Carlos	NULL	NONE	
2011-09-10	492630396824288000	Julie	NULL	NONE	
2011-09-10	497042084430103000	Eugene	NULL	NONE	
2016-01-19	497742885879118000	Stanley	NULL	NONE	
2011-09-10	497904421215766000	Oscar	NULL	NONE	
2011-09-10	498785734603842000	Tony	NULL	NONE	
2011-09-10	498961330469141000	Ethel	7.0	POSITIVE Sentiment	
2016-04-26	500296478047702000	Glen	NULL	NONE	
2011-09-10	502938309621066000	Joan	NULL	NONE	
2016-04-26	503520275003631000	Vernon	NULL	NONE	
2016-07-24	505380376161911000	Max	NULL	NONE	
2016-07-11	507196095788222000	Joy	NULL	NONE	
2011-09-10	508665556481819000	Emma	4.0	POSITIVE Sentiment	
2011-09-10	509056219671459000	Arthur	-3.0	NEGATIVE Sentiment	
2016-01-19	511319795634024000	Ted	NULL	NONE	
2015-12-15	513137013318944000	Jennifer		NULL	NONE
2015-12-22	515049870010582000	Neil	NULL	NONE	
2011-09-10	515550406704664000	Kathryn	NULL	NONE	
2011-09-10	516680352480685000	Erica	NULL	NONE	
2011-09-10	518775726818080000	Tina	3.0	POSITIVE Sentiment	
2016-07-08	520559468929107000	Ricky	NULL	NONE	

2016-07-29	522735212763929000	Joan	NULL	NONE	
2011-09-10	522751262894529000	Ruth	-5.0	NEGATIVE Sentiment	
2011-09-10	523670627452799000	Constance		NULL	NONE
2011-09-10	526352956630677000	Kyle	3.0	POSITIVE Sentiment	
2011-09-10	527473370935382000	Sean	NULL	NONE	
2016-07-08	528017694057397000	Lynn	3.0	POSITIVE Sentiment	
2011-09-10	530448856332303000	Andrea	NULL	NONE	
2016-06-07	532565902986021000	Bruce	2.0	POSITIVE Sentiment	
2016-07-29	534262624541602000	Dan	NULL	NONE	
2011-09-10	534806540605840000	Tina	-5.0	NEGATIVE Sentiment	
2011-09-10	540253378244347000	Terry	-3.0	NEGATIVE Sentiment	
2011-09-10	541361084395978000	Walter	-2.0	NEGATIVE Sentiment	
2016-01-19	541661596694693000	Kristin	1.0	POSITIVE Sentiment	
2016-01-29	542683216038544000	Steve	1.0	POSITIVE Sentiment	
2016-04-04	545555361130000000	Gina	NULL	NONE	
2011-09-10	545893443885142000	Dorothy	NULL	NONE	
2016-02-20	547258850791958000	Kevin	NULL	NONE	
2009-03-18	550058458689286000	Kristen	NULL	NONE	
2016-07-24	550426354302548000	Brooke	NULL	NONE	
2016-04-04	550528175483889000	Edwin	5.0	POSITIVE Sentiment	
2016-06-07	552086165505119000	Sally	-2.0	NEGATIVE Sentiment	
2011-09-10	553485622362101000	Lynda	NULL	NONE	
2016-01-29	553979538674296000	Constance		1.0	POSITIVE Sentiment
2011-09-10	554234959190646000	Walter	NULL	NONE	
2011-09-10	554464446481069000	Brooke	4.0	POSITIVE Sentiment	
2016-05-21	557148105287787000	Bonnie	-3.0	NEGATIVE Sentiment	
2016-03-20	560272229196327000	Claudia	NULL	NONE	
2016-07-11	560940286305638000	Lee	NULL	NONE	
2011-09-10	561697330497060000	Joseph	NULL	NONE	
2015-12-15	561923667941253000	Gretchen		-5.0	NEGATIVE Sentiment
2016-07-08	562903027578790000	Bob	6.0	POSITIVE Sentiment	
2016-07-24	565047664539432000	Emma	NULL	NONE	
2016-07-29	568832906644516000	Karen	NULL	NONE	
2011-09-10	569781884278506000	Earl	NULL	NONE	
2016-07-11	574007977344591000	Karen	NULL	NONE	
2016-07-11	577008061158285000	Lorraine		NULL	NONE
2011-09-10	577086515368951000	Melanie	NULL	NONE	
2016-04-04	577677076504199000	Billie	-6.0	NEGATIVE Sentiment	
2011-09-10	581496238710310000	Shirley	NULL	NONE	
2015-12-18	583475249539407000	Susan	NULL	NONE	
2011-09-10	586132851463130000	Charlie	NULL	NONE	
2016-07-29	591002608868294000	Anne	NULL	NONE	
2016-01-19	591005493005734000	Tim	2.0	POSITIVE Sentiment	
2011-09-10	591041649780096000	Kathleen		NULL	NONE
2011-09-10	598640139733146000	Lori	NULL	NONE	

……

```
2011-09-10      876517185424031000      Dorothy NULL    NONE
2011-09-10      876557870030974000      Theodore        -3.0    NEGATIVE Sentiment
2011-09-10      876852986037848000      Douglas NULL    NONE
2016-07-08      878424466087027000      Stuart  NULL    NONE
2011-09-10      878934128370149000      Wendy   3.0     POSITIVE Sentiment
2016-06-07      880050504499670000      Keith   -3.0    NEGATIVE Sentiment
2011-09-10      881919550707052000      Marie   NULL    NONE
2016-01-29      883029445050689000      John    -3.0    NEGATIVE Sentiment
2011-09-10      883318030923533000      Gilbert NULL    NONE
2015-12-22      883849857287617000      Harriet NULL    NONE
2016-01-29      886494948390315000      Kenneth NULL    NONE
2011-09-10      890745718284415000      Suzanne NULL    NONE
2011-09-10      892340354124439000      Katharine       0.0     NONE
2016-01-19      892786063807571000      Catherine       -4.0    NEGATIVE Sentiment
2015-12-18      892812529934080000      Calvin  -2.0    NEGATIVE Sentiment
2011-09-10      892993179592271000      Edgar   -8.0    NEGATIVE Sentiment
2016-03-20      893860208370497000      Greg    3.0     POSITIVE Sentiment
2011-09-10      894589045629596000      Louis   1.0     POSITIVE Sentiment
2015-12-22      897897234225893000      Penny   NULL    NONE
2011-09-10      898147844068158000      Alexandra       8.0     POSITIVE Sentiment
2016-04-26      899120405062922000      Todd    -3.0    NEGATIVE Sentiment
2016-01-25      899285588251612000      Anna    2.0     POSITIVE Sentiment
Time taken: 0.125 seconds, Fetched: 550 row(s)
```

These are the few results from the above query.

3. Propose a better solution for the sentiment analysis as compared to 1(c). Cite the source. (5 points) Note: You just need to provide the solution; you are not required to solve the problem using the solution.

Ans

- Above what I have done was I broke down the text into single words. By following this procedure, while analyzing the hash tags this might not know for what situations these hashtags were used. They don't have the ability to understand the situation.
- They just analyze based on the hashtag word and don't understand the human way of reacting to the situations using the words that is they don't cannot sense the situation whether the hashtag is used for sarcastic or ironic or mocking or etc cases.
- For this the best way is through the natural processing algorithms.
- Below were some of the links which describes the concept that might help analyze more perfectly.

  https://casmodeling.springeropen.com/articles/10.1186/s40294-016-0016-9

  https://algorithmia.com/blog/benchmarking-sentiment-analysis-algorithms

  https://github.com/cjhutto/vaderSentiment

- This analysis can also be done by implementing machine learning methodologies so as it can learn itself on basis of what words were used, how many times they were used and tries to understand the situation in which they were used.