Twitter Data Analysis
Assignment 2

- Darshil Sanghvi (drs170530)

- Prompt hive about the jar file for reading json files with serde method

```
ADD JAR Desktop/json-serde-1.3-jar-with-dependencies.jar;
```

-Creating table in hive

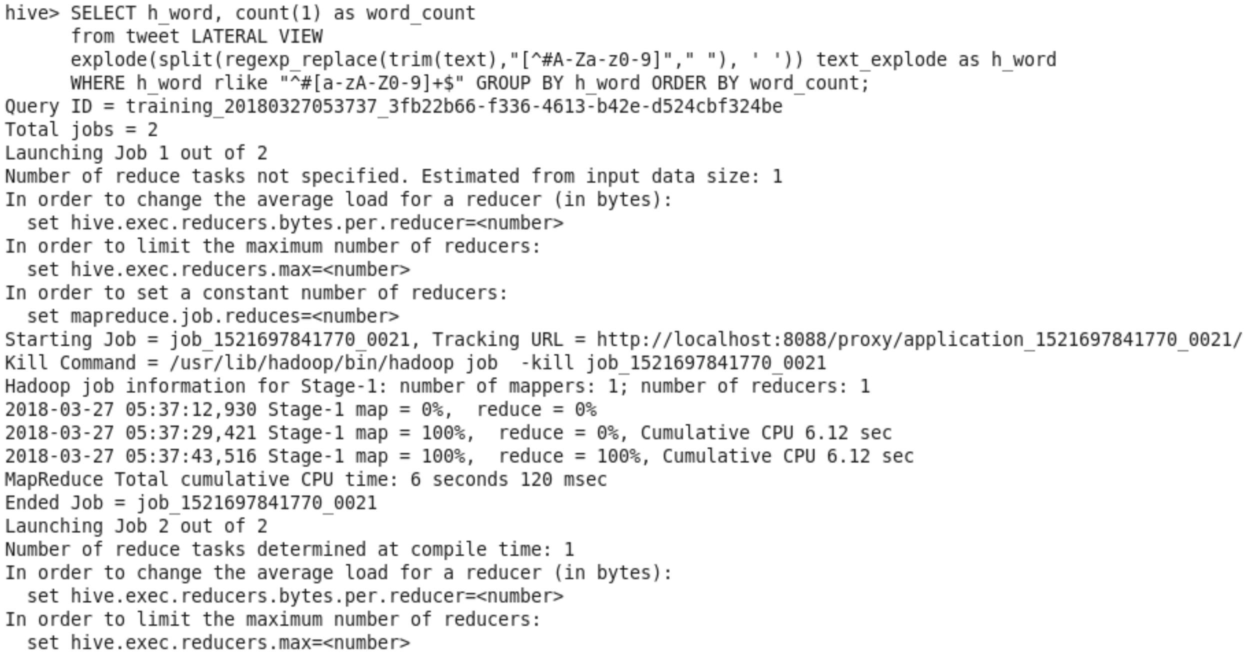
```
create table tweet
(
id bigint,
text string,
created_at string,
retweet_count int,
user struct<location:string,id_str:bigint,name:string,created_at:string,screen_name:string,followers_count:int>,
quoted_status struct< user : struct<location:string,id: bigint,followers_count:int,name:string> >

)ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe';
```

Loading data from local path into the created table

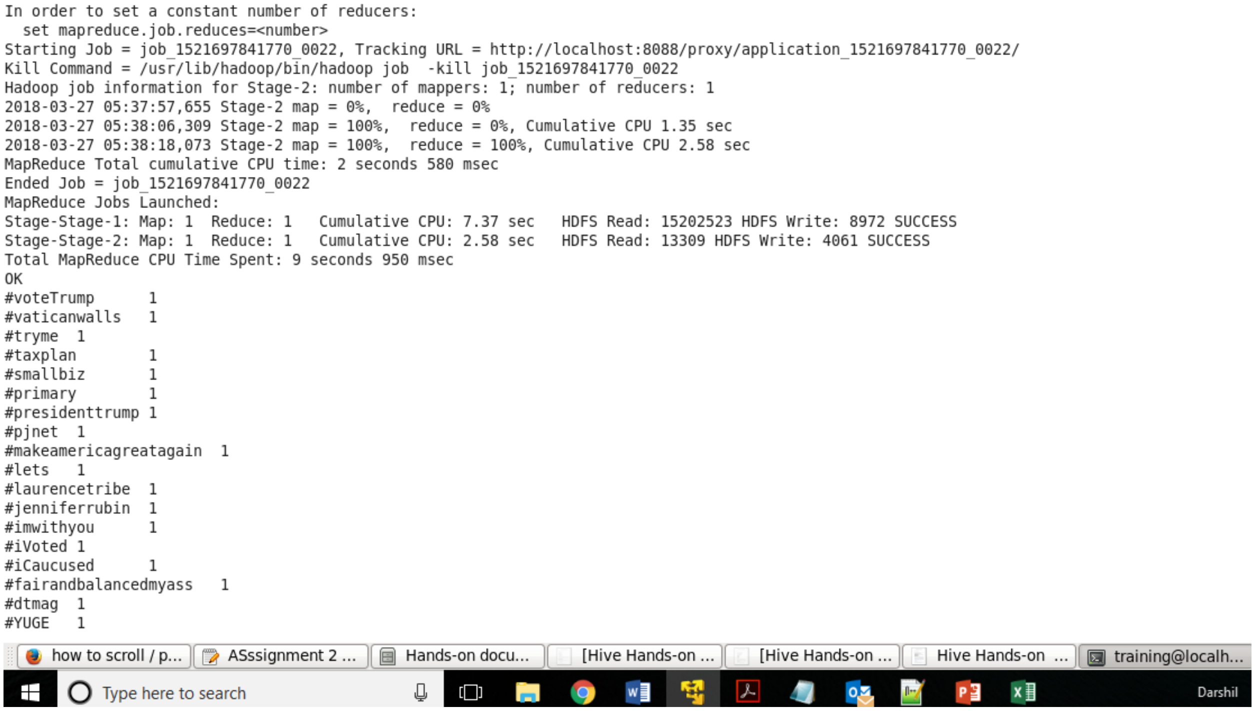
```
load data local inpath 'Desktop/tweet_final.json' overwrite into table tweet;
```

Question 1.a. What are the hashtags used and how many times each are used?

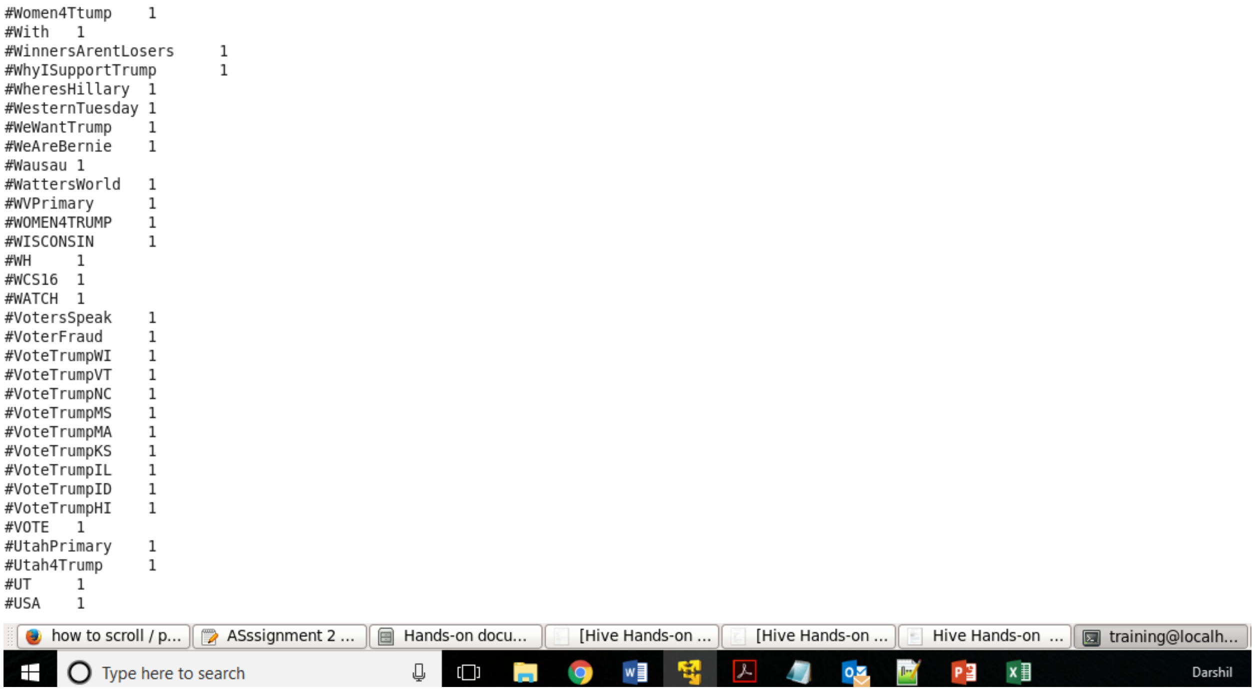
SELECT h_word, count(1) as word_count
from tweet LATERAL VIEW
explode(split(regexp_replace(trim(text),"[^#A-Za-z0-9]"," "), ' ')) text_explode as h_word
WHERE h_word rlike "^#[a-zA-Z0-9]+$" GROUP BY h_word ORDER BY word_count;



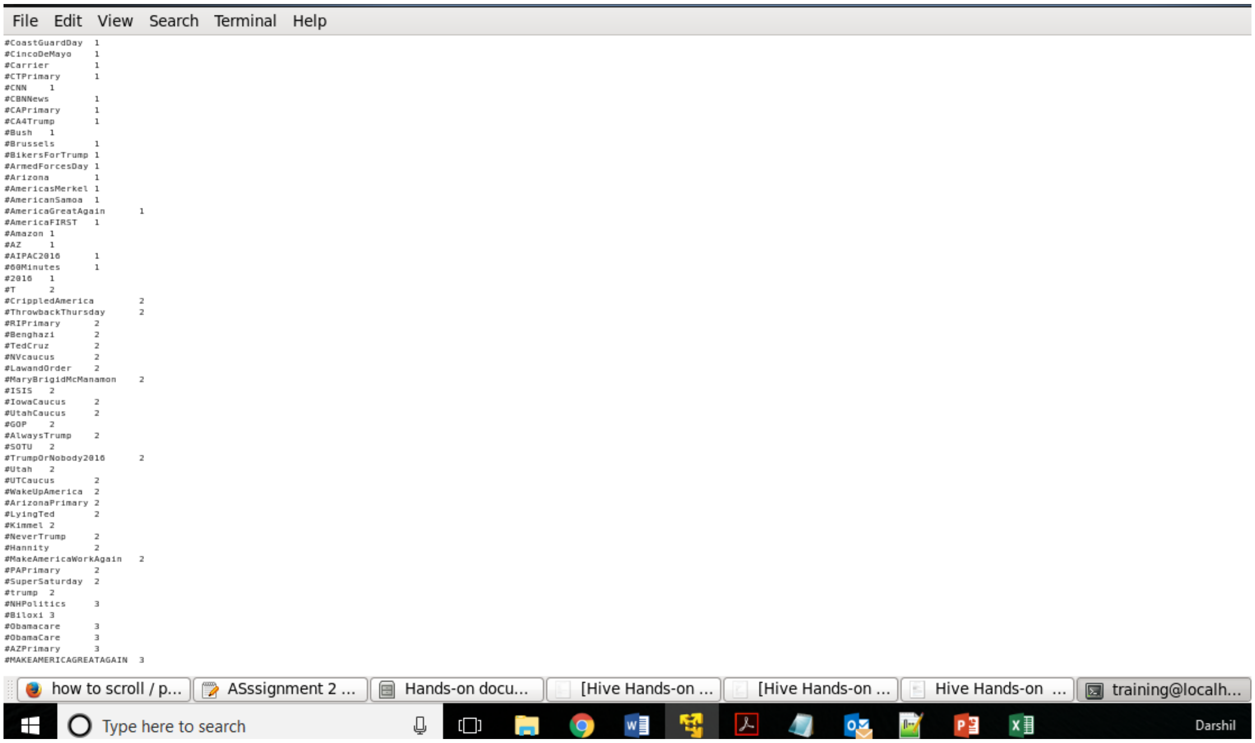
```
hive> SELECT h_word, count(1) as word_count
      from tweet LATERAL VIEW
      explode(split(regexp_replace(trim(text),"[^#A-Za-z0-9]"," "), ' ')) text_explode as h_word
      WHERE h_word rlike "^#[a-zA-Z0-9]+$" GROUP BY h_word ORDER BY word_count;
Query ID = training_20180327053737_3fb22b66-f336-4613-b42e-d524cbf324be
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_1521697841770_0021, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 05:37:12,930 Stage-1 map = 0%,  reduce = 0%
2018-03-27 05:37:29,421 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.12 sec
2018-03-27 05:37:43,516 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.12 sec
MapReduce Total cumulative CPU time: 6 seconds 120 msec
Ended Job = job_1521697841770_0021
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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>
```



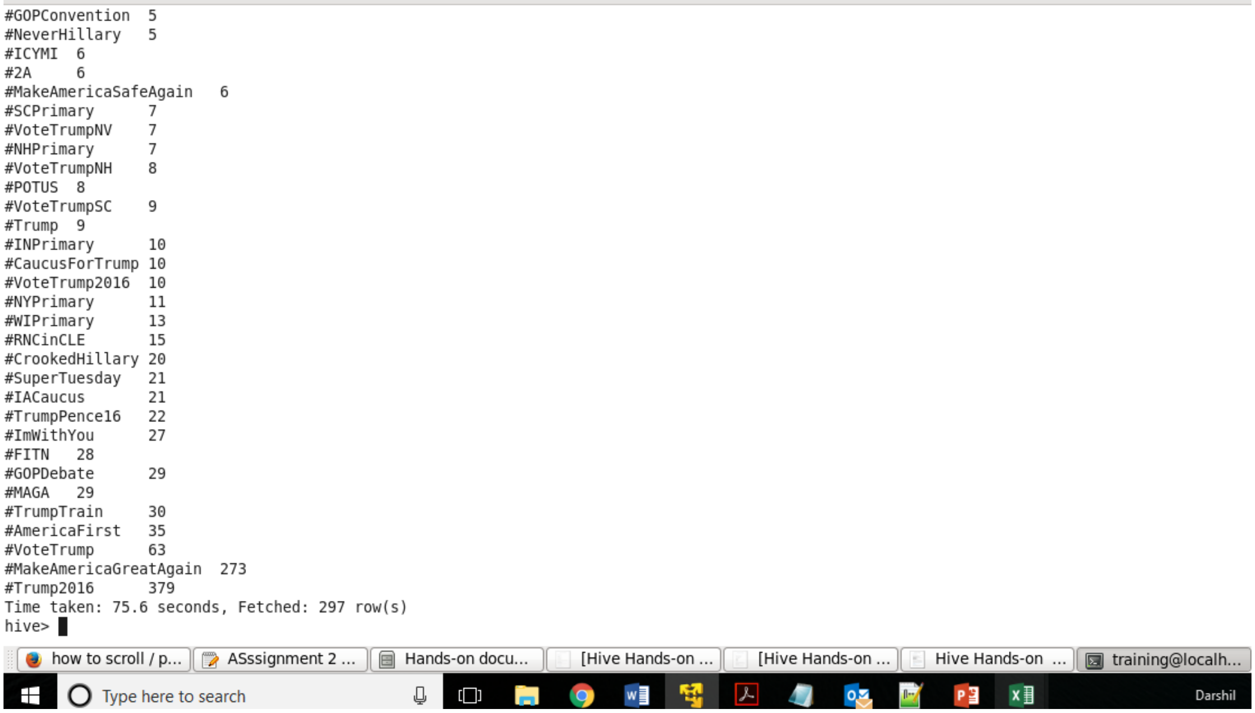
```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1521697841770_0022, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0022
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-27 05:37:57,655 Stage-2 map = 0%,  reduce = 0%
2018-03-27 05:38:06,309 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.35 sec
2018-03-27 05:38:18,073 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.58 sec
MapReduce Total cumulative CPU time: 2 seconds 580 msec
Ended Job = job_1521697841770_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.37 sec   HDFS Read: 15202523 HDFS Write: 8972 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.58 sec   HDFS Read: 13309 HDFS Write: 4061 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 950 msec
OK
#voteTrump      1
#vaticanwalls   1
#tryme  1
#taxplan        1
#smallbiz       1
#primary        1
#presidenttrump 1
#pjnet  1
#makeamericagreatagain  1
#lets   1
#laurencetribe  1
#jenniferrubin  1
#imwithyou      1
#iVoted 1
#iCaucused      1
#fairandbalancedmyass   1
#dtmag  1
#YUGE   1
```



```
#Women4Ttump    1
#With   1
#WinnersArentLosers     1
#WhyISupportTrump       1
#WheresHillary  1
#WesternTuesday 1
#WeWantTrump    1
#WeAreBernie    1
#Wausau 1
#WattersWorld   1
#WVPrimary      1
#WOMEN4TRUMP    1
#WISCONSIN      1
#WH     1
#WCS16  1
#WATCH  1
#VotersSpeak    1
#VoterFraud     1
#VoteTrumpWI    1
#VoteTrumpVT    1
#VoteTrumpNC    1
#VoteTrumpMS    1
#VoteTrumpMA    1
#VoteTrumpKS    1
#VoteTrumpIL    1
#VoteTrumpID    1
#VoteTrumpHI    1
#VOTE   1
#UtahPrimary    1
#Utah4Trump     1
#UT     1
#USA    1
```



```
#CoastGuardDay  1
#CincoDeMayo    1
#Carrier        1
#CTPrimary      1
#CNN    1
#CBNNews        1
#CAPrimary      1
#CA4Trump       1
#Bush   1
#Brussels       1
#BikersForTrump 1
#ArmedForcesDay 1
#Arizona        1
#AmericasMerkel 1
#AmericanSamoa  1
#AmericaGreatAgain      1
#AmericaFIRST   1
#Amazon 1
#AZ     1
#AIPAC2016      1
#60Minutes      1
#2016   1
#T      2
#CrippledAmerica        2
#ThrowbackThursday      2
#RIPrimary      2
#Benghazi       2
#TedCruz        2
#NVcaucus       2
#LawandOrder    2
#MaryBrigidMcManamon    2
#ISIS   2
#IowaCaucus     2
#UtahCaucus     2
#GOP    2
#AlwaysTrump    2
#SOTU   2
#TrumpOrNobody2016      2
#Utah   2
#UTCaucus       2
#WakeUpAmerica  2
#ArizonaPrimary 2
#LyingTed       2
#Kimmel 2
#NeverTrump     2
#Hannity        2
#MakeAmericaWorkAgain   2
#PAPrimary      2
#SuperSaturday  2
#trump  2
#NHPolitics     3
#Biloxi 3
#Obamacare      3
#ObamaCare      3
#AZPrimary      3
#MAKEAMERICAGREATAGAIN  3
```



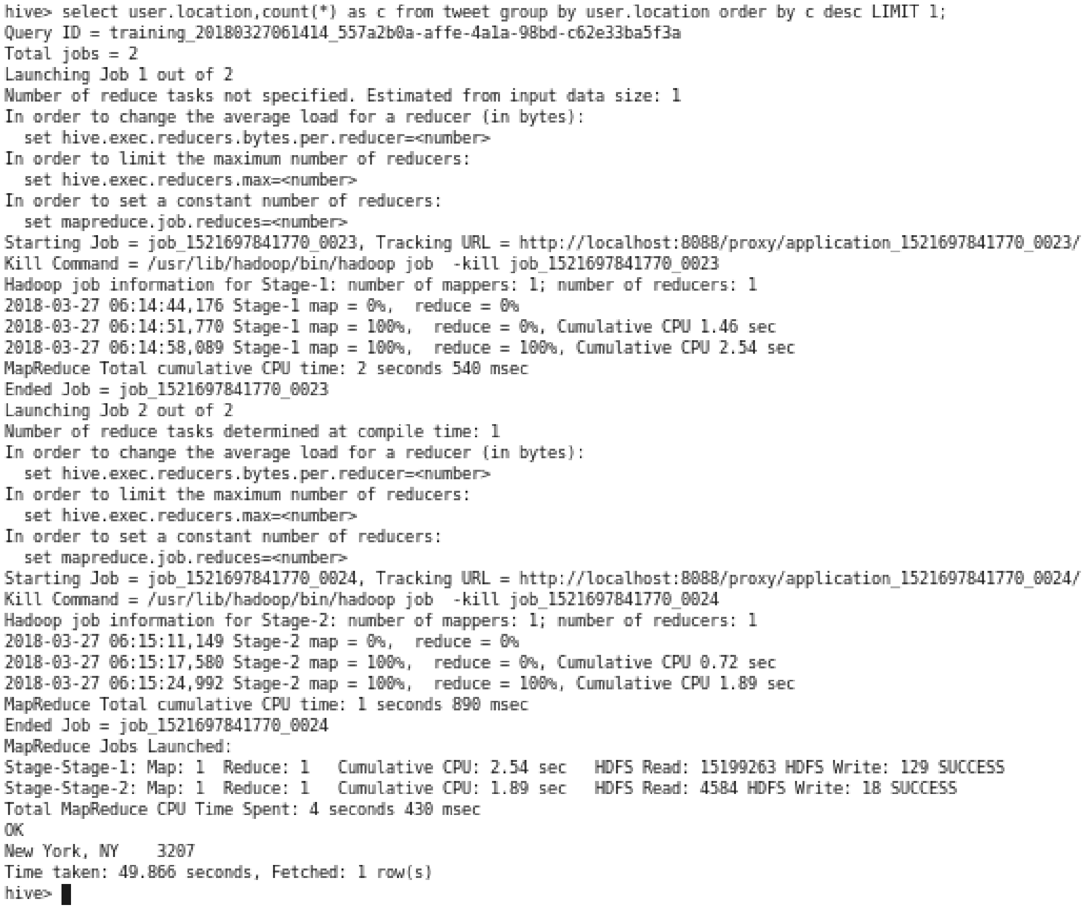
```
#GOPConvention  5
#NeverHillary   5
#ICYMI  6
#2A     6
#MakeAmericaSafeAgain   6
#SCPrimary      7
#VoteTrumpNV    7
#NHPrimary      7
#VoteTrumpNH    8
#POTUS  8
#VoteTrumpSC    9
#Trump  9
#INPrimary      10
#CaucusForTrump 10
#VoteTrump2016  10
#NYPrimary      11
#WIPrimary      13
#RNCinCLE       15
#CrookedHillary 20
#SuperTuesday   21
#IACaucus       21
#TrumpPence16   22
#ImWithYou      27
#FITN   28
#GOPDebate      29
#MAGA   29
#TrumpTrain     30
#AmericaFirst   35
#VoteTrump      63
#MakeAmericaGreatAgain  273
#Trump2016      379
Time taken: 75.6 seconds, Fetched: 297 row(s)
hive>
```

Question 1.b. Which State have the most active users and how many tweets are posted by State?

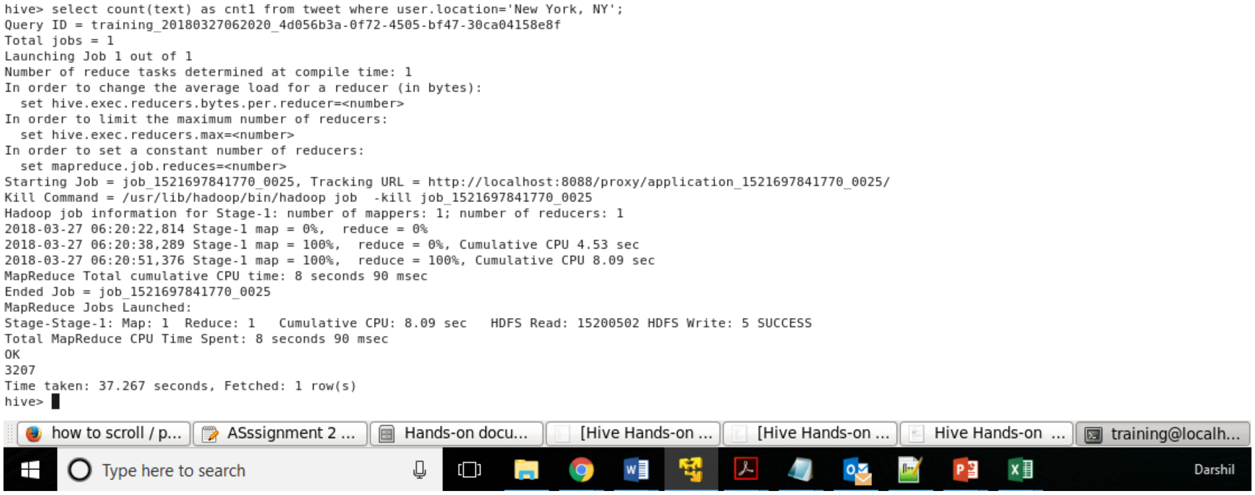
select user.location,count(*) as c from tweet group by user.location order by c desc LIMIT 1;

select count(text) as cnt1 from tweet where user.location='New York, NY';

New York has the most active users



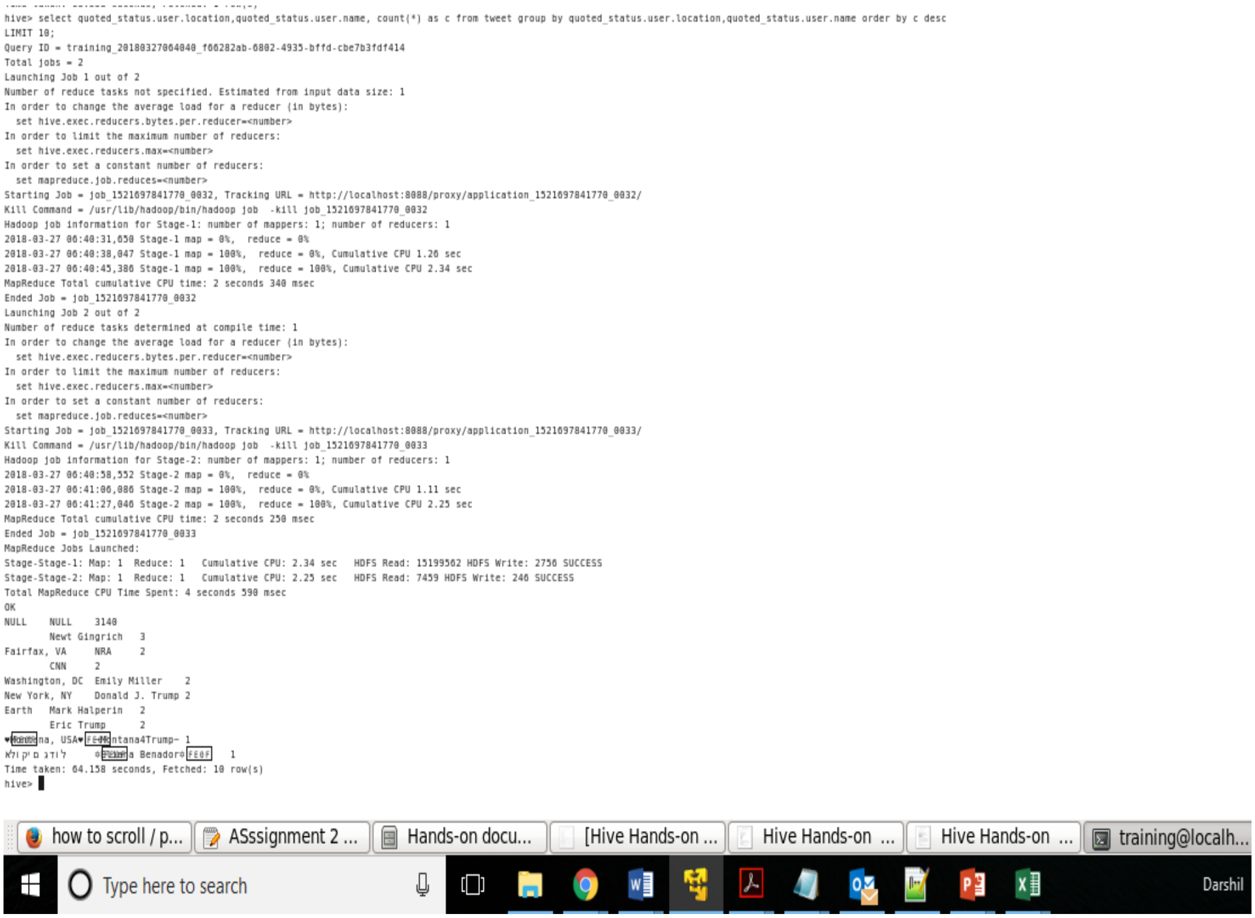
```
hive> select user.location,count(*) as c from tweet group by user.location order by c desc LIMIT 1;
Query ID = training_20180327061414_557a2b0a-affe-4a1a-98bd-c62e33ba5f3a
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change 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_1521697841770_0023, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 06:14:44,176 Stage-1 map = 0%,  reduce = 0%
2018-03-27 06:14:51,770 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.46 sec
2018-03-27 06:14:58,089 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.54 sec
MapReduce Total cumulative CPU time: 2 seconds 540 msec
Ended Job = job_1521697841770_0023
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_1521697841770_0024, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0024/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0024
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-27 06:15:11,149 Stage-2 map = 0%,  reduce = 0%
2018-03-27 06:15:17,580 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.72 sec
2018-03-27 06:15:24,992 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.89 sec
MapReduce Total cumulative CPU time: 1 seconds 890 msec
Ended Job = job_1521697841770_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.54 sec   HDFS Read: 15199263 HDFS Write: 129 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 1.89 sec   HDFS Read: 4584 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 430 msec
OK
New York, NY    3207
Time taken: 49.866 seconds, Fetched: 1 row(s)
hive> 
```



```
hive> select count(text) as cntl from tweet where user.location='New York, NY';
Query ID = training_20180327062020_4d056b3a-0f72-4505-bf47-30ca04158e8f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1521697841770_0025, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 06:20:22,814 Stage-1 map = 0%,  reduce = 0%
2018-03-27 06:20:38,289 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.53 sec
2018-03-27 06:20:51,376 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.09 sec
MapReduce Total cumulative CPU time: 8 seconds 90 msec
Ended Job = job_1521697841770_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.09 sec   HDFS Read: 15200502 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 90 msec
OK
3207
Time taken: 37.267 seconds, Fetched: 1 row(s)
hive>
```

Users from other States can be found via the user within the quoted_status in the json file:

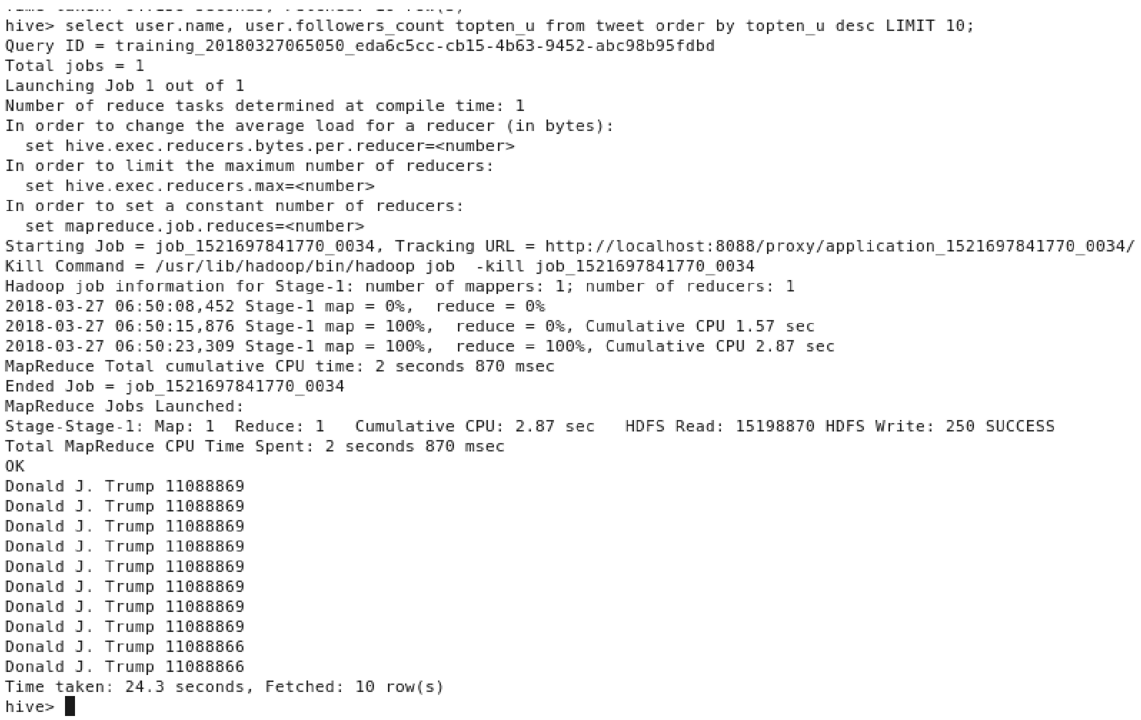
select quoted_status.user.location,count(*) as c from tweet group by quoted_status.user.location order by c desc LIMIT 1;



```
hive> select quoted_status.user.location,quoted_status.user.name, count(*) as c from tweet group by quoted_status.user.location,quoted_status.user.name order by c desc
LIMIT 10;
Query ID = training_20180327064040_f66282ab-6802-4935-bffd-cbe7b3fdf414
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_1521697841770_0032, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0032/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0032
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 06:40:31,650 Stage-1 map = 0%,  reduce = 0%
2018-03-27 06:40:38,047 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2018-03-27 06:40:45,386 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.34 sec
MapReduce Total cumulative CPU time: 2 seconds 340 msec
Ended Job = job_1521697841770_0032
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_1521697841770_0033, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0033/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0033
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-27 06:40:58,552 Stage-2 map = 0%,  reduce = 0%
2018-03-27 06:41:06,086 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.11 sec
2018-03-27 06:41:27,046 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.25 sec
MapReduce Total cumulative CPU time: 2 seconds 250 msec
Ended Job = job_1521697841770_0033
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.34 sec   HDFS Read: 15199562 HDFS Write: 2756 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.25 sec   HDFS Read: 7459 HDFS Write: 246 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 590 msec
OK
NULL    NULL    3140
        Newt Gingrich   3
Fairfax, VA     NRA     2
        CNN     2
Washington, DC  Emily Miller    2
New York, NY    Donald J. Trump 2
Earth   Mark Halperin   2
        Eric Trump      2
♥[illegible]na, USA♥FE0FMontana4Trump- 1
לודג םיקולא     ♦[illegible]a Benador♦FE0F   1
Time taken: 64.158 seconds, Fetched: 10 row(s)
hive>
```

Question 1.c.: Based on the user's followers count, who are the top ten users who have tweeted?

select user.name, user.followers_count topten_u from tweet order by topten_u desc LIMIT 10;



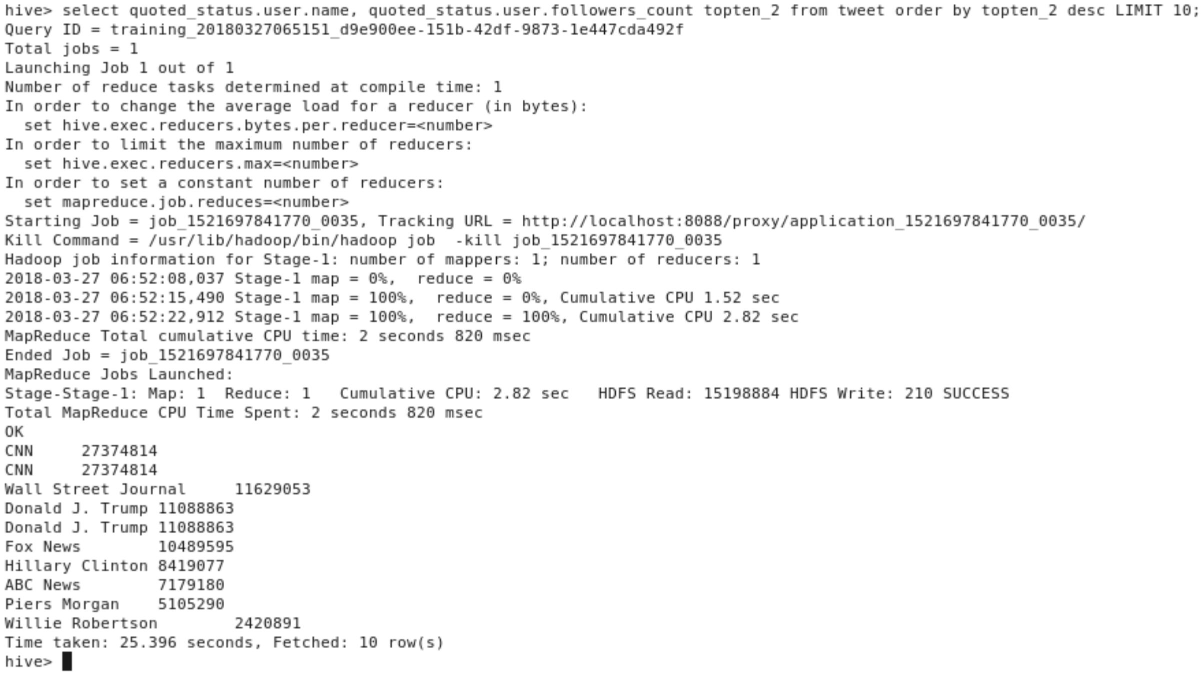
```
hive> select user.name, user.followers_count topten_u from tweet order by topten_u desc LIMIT 10;
Query ID = training_20180327065050_eda6c5cc-cb15-4b63-9452-abc98b95fdbd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load 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_1521697841770_0034, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0034/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0034
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 06:50:08,452 Stage-1 map = 0%,  reduce = 0%
2018-03-27 06:50:15,876 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.57 sec
2018-03-27 06:50:23,309 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.87 sec
MapReduce Total cumulative CPU time: 2 seconds 870 msec
Ended Job = job_1521697841770_0034
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.87 sec   HDFS Read: 15198870 HDFS Write: 250 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 870 msec
OK
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088869
Donald J. Trump 11088866
Donald J. Trump 11088866
Time taken: 24.3 seconds, Fetched: 10 row(s)
hive> █
```

Donald Trump is repeated all top ten times in terms of **followers count.**

If we consider the username in quoted_status section,

select quoted_status.user.name, quoted_status.user.followers_count topten_2 from tweet order by topten_2 desc LIMIT 10;

We get some other users other than Trump now:



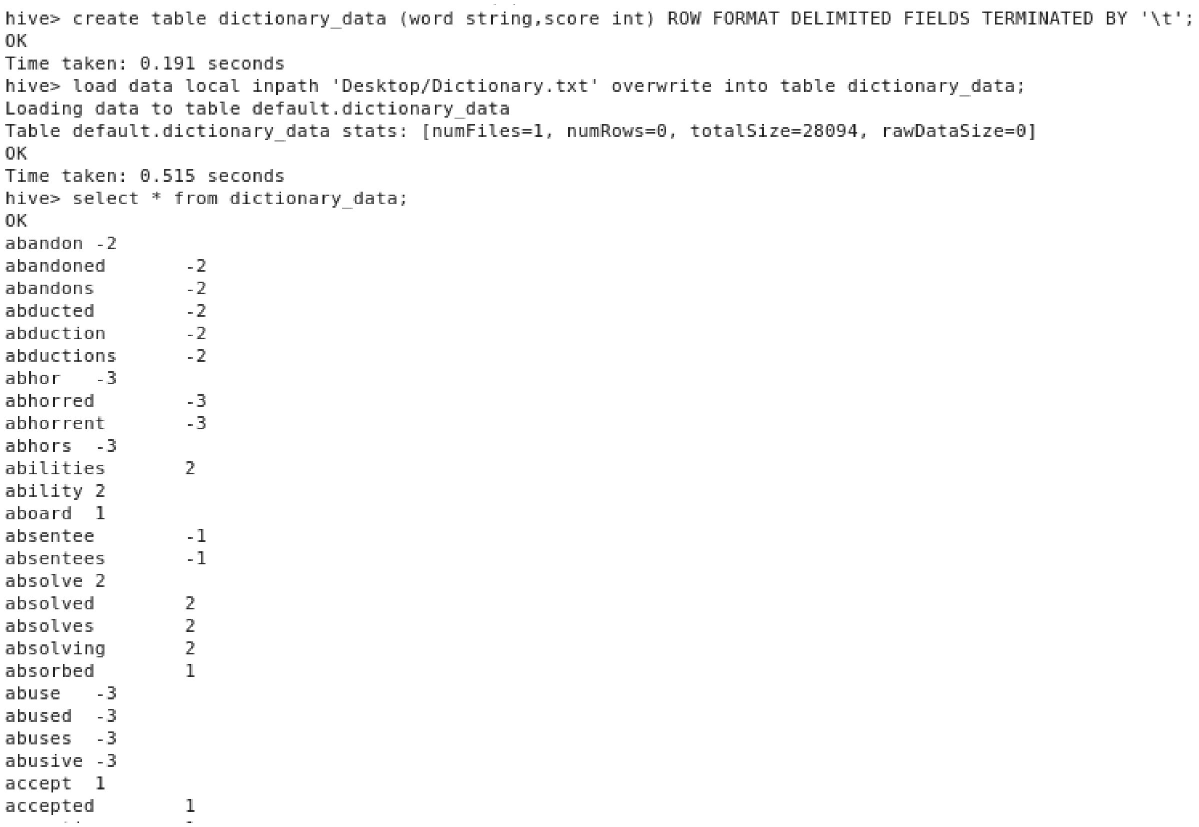
```
hive> select quoted_status.user.name, quoted_status.user.followers_count topten_2 from tweet order by topten_2 desc LIMIT 10;
Query ID = training_20180327065151_d9e900ee-151b-42df-9873-1e447cda492f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order 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_1521697841770_0035, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0035/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0035
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 06:52:08,037 Stage-1 map = 0%,  reduce = 0%
2018-03-27 06:52:15,490 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.52 sec
2018-03-27 06:52:22,912 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.82 sec
MapReduce Total cumulative CPU time: 2 seconds 820 msec
Ended Job = job_1521697841770_0035
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.82 sec   HDFS Read: 15198884 HDFS Write: 210 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 820 msec
OK
CNN     27374814
CNN     27374814
Wall Street Journal     11629053
Donald J. Trump 11088863
Donald J. Trump 11088863
Fox News        10489595
Hillary Clinton 8419077
ABC News        7179180
Piers Morgan    5105290
Willie Robertson        2420891
Time taken: 25.396 seconds, Fetched: 10 row(s)
hive> █
```

Hence, it completely depends on the scope of the data and which fields are considered significant for analysis.

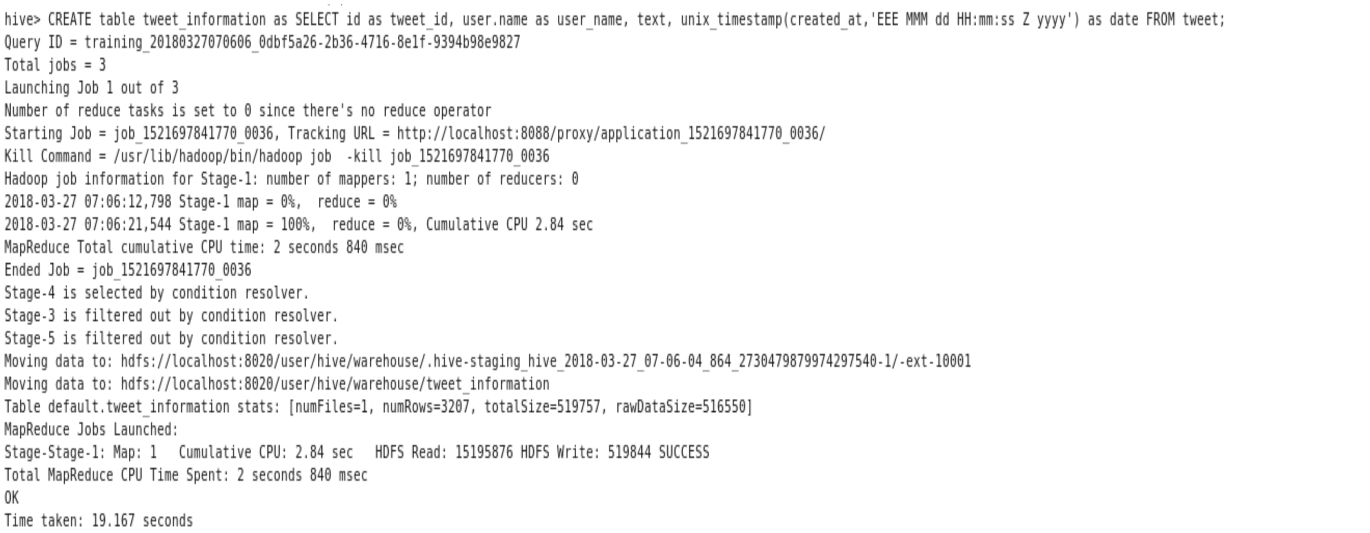
Q.1.d. What is the polarity score for each tweet that was posted? Does the tweet have a positive or negative sentiment?

```
create table dictionary_data (word string,score int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

load data local inpath 'Desktop/Dictionary.txt' overwrite into table dictionary_data;


CREATE table tweet_information as SELECT id as tweet_id, user.name as user_name, text, unix_timestamp(created_at,'EEE MMM dd HH:mm:ss Z yyyy') as date FROM tweet;


create table tweet_explode as select tweet_id, user_name, from_unixtime(date,'yyyy-MM-dd') as date, word from tweet_information LATERAL VIEW explode(split(regexp_replace(lower(text),"[^#A-Za-z0-9]"," "), ' ')) text_x as word;
```

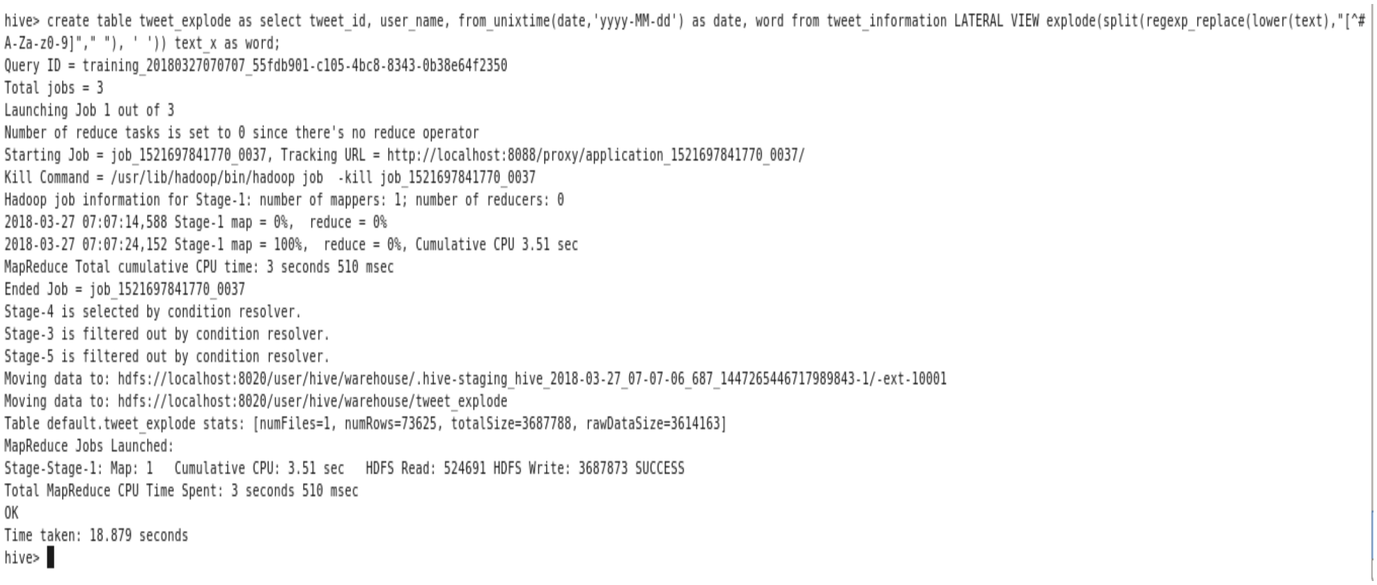
```
CREATE table mapping as SELECT t.tweet_id, t.user_name, t.date, t.word, d.score FROM
dictionary_data d RIGHT OUTER JOIN tweet_explode t on (t.word = d.word);


CREATE table tweet_score as SELECT tweet_id, user_name, date, SUM(score) as tweet_score FROM
mapping GROUP BY tweet_id, user_name, date;


CREATE table sentiment as SELECT tweet_id, user_name, date, CASE WHEN tweet_score > 0 THEN
'Positive' WHEN tweet_score < 0 THEN 'Negative' ELSE 'None' END as sentiment FROM tweet_score
where tweet_score is not null;


Select * from sentiment;
```



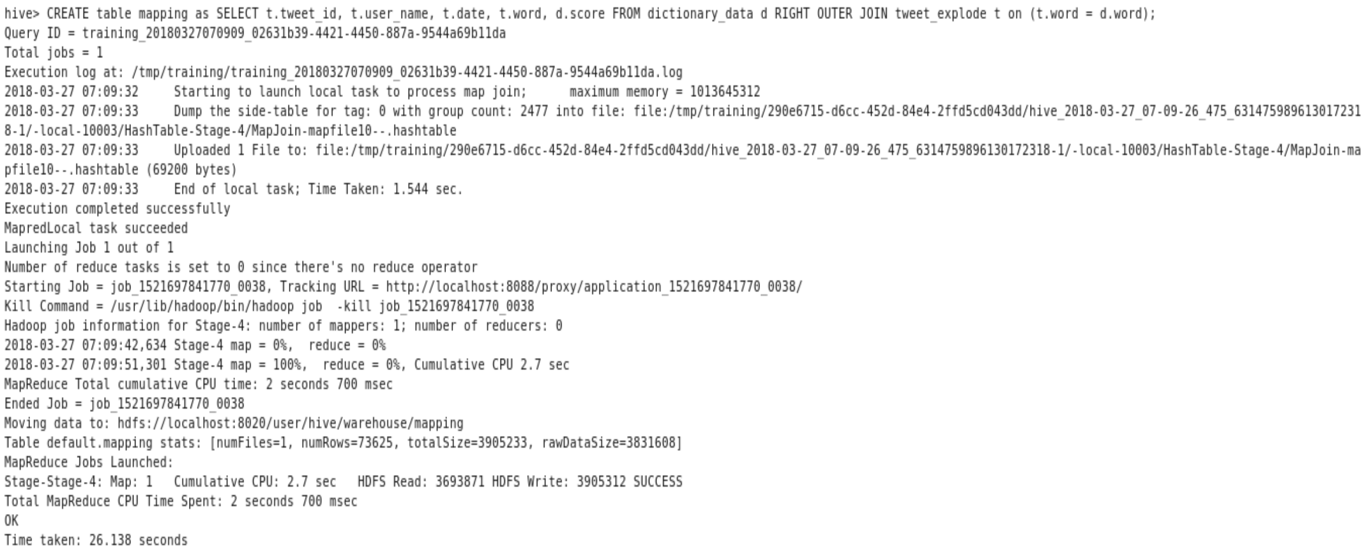
```
hive> create table dictionary_data (word string,score int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.191 seconds
hive> load data local inpath 'Desktop/Dictionary.txt' overwrite into table dictionary_data;
Loading data to table default.dictionary_data
Table default.dictionary_data stats: [numFiles=1, numRows=0, totalSize=28094, rawDataSize=0]
OK
Time taken: 0.515 seconds
hive> select * from dictionary_data;
OK
abandon -2
abandoned       -2
abandons        -2
abducted        -2
abduction       -2
abductions      -2
abhor   -3
abhorred        -3
abhorrent       -3
abhors  -3
abilities       2
ability 2
aboard  1
absentee        -1
absentees       -1
absolve 2
absolved        2
absolves        2
absolving       2
absorbed        1
abuse   -3
abused  -3
abuses  -3
abusive -3
accept  1
accepted        1
```



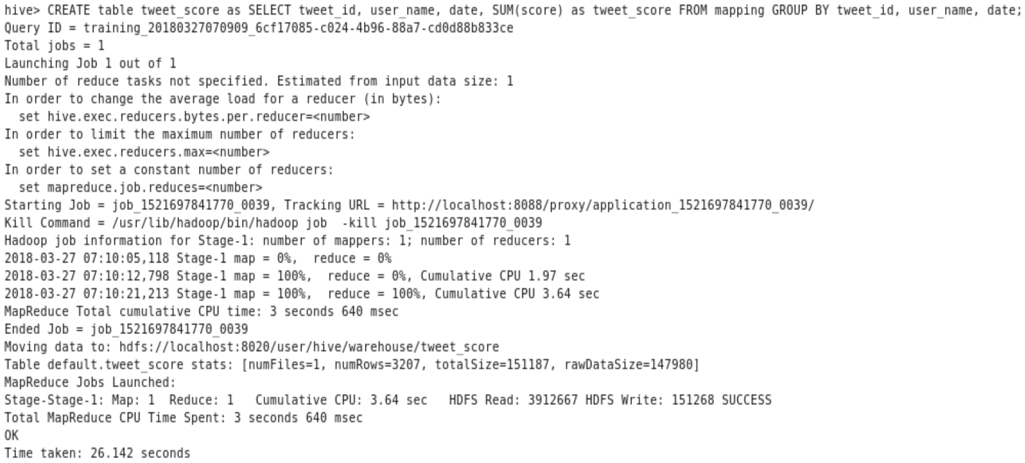
```
hive> CREATE table tweet_information as SELECT id as tweet_id, user.name as user_name, text, unix_timestamp(created_at,'EEE MMM dd HH:mm:ss Z yyyy') as date FROM tweet;
Query ID = training_20180327070606_0dbf5a26-2b36-4716-8e1f-9394b98e9827
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1521697841770_0036, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0036/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0036
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 07:06:12,798 Stage-1 map = 0%,  reduce = 0%
2018-03-27 07:06:21,544 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.84 sec
MapReduce Total cumulative CPU time: 2 seconds 840 msec
Ended Job = job_1521697841770_0036
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/.hive-staging_hive_2018-03-27_07-06-04_864_2730479879974297540-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweet_information
Table default.tweet_information stats: [numFiles=1, numRows=3207, totalSize=519757, rawDataSize=516550]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.84 sec   HDFS Read: 15195876 HDFS Write: 519844 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 840 msec
OK
Time taken: 19.167 seconds
```



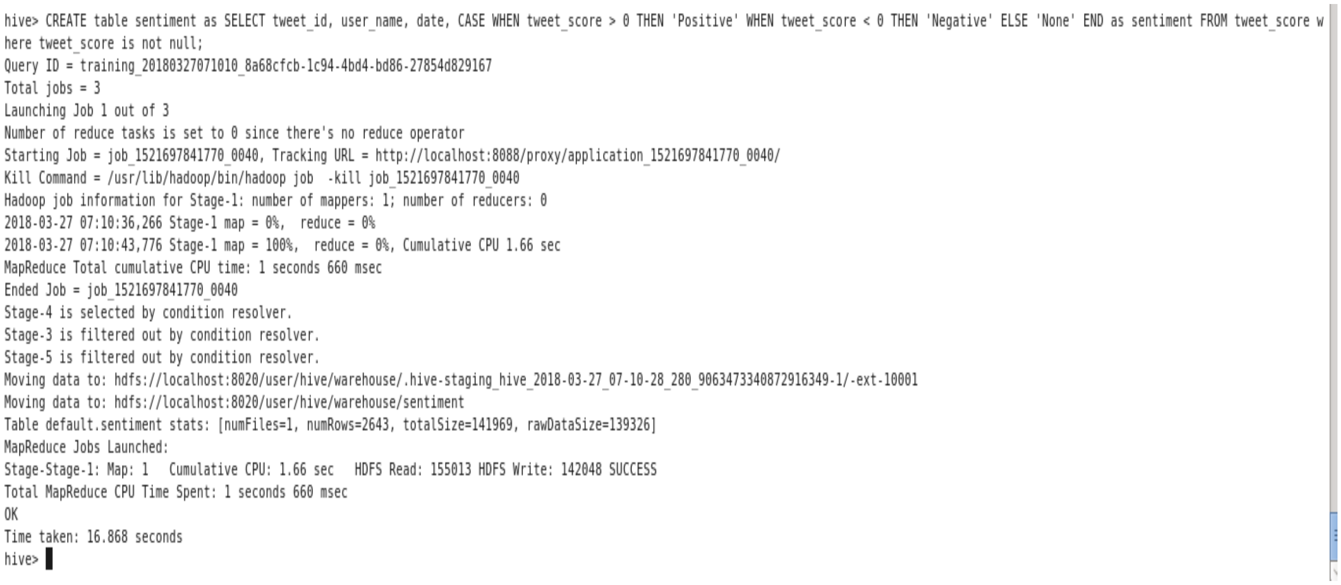
```
hive> create table tweet_explode as select tweet_id, user_name, from_unixtime(date,'yyyy-MM-dd') as date, word from tweet_information LATERAL VIEW explode(split(regexp_replace(lower(text),"[^#
A-Za-z0-9]"," "), ' ')) text_x as word;
Query ID = training_20180327070707_55fdb901-c105-4bc8-8343-0b38e64f2350
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_1521697841770_0037, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0037/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0037
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 07:07:14,588 Stage-1 map = 0%,  reduce = 0%
2018-03-27 07:07:24,152 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.51 sec
MapReduce Total cumulative CPU time: 3 seconds 510 msec
Ended Job = job_1521697841770_0037
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/.hive-staging_hive_2018-03-27_07-07-06_687_1447265446717989843-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweet_explode
Table default.tweet_explode stats: [numFiles=1, numRows=73625, totalSize=3687788, rawDataSize=3614163]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.51 sec   HDFS Read: 524691 HDFS Write: 3687873 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 510 msec
OK
Time taken: 18.879 seconds
hive> █
```



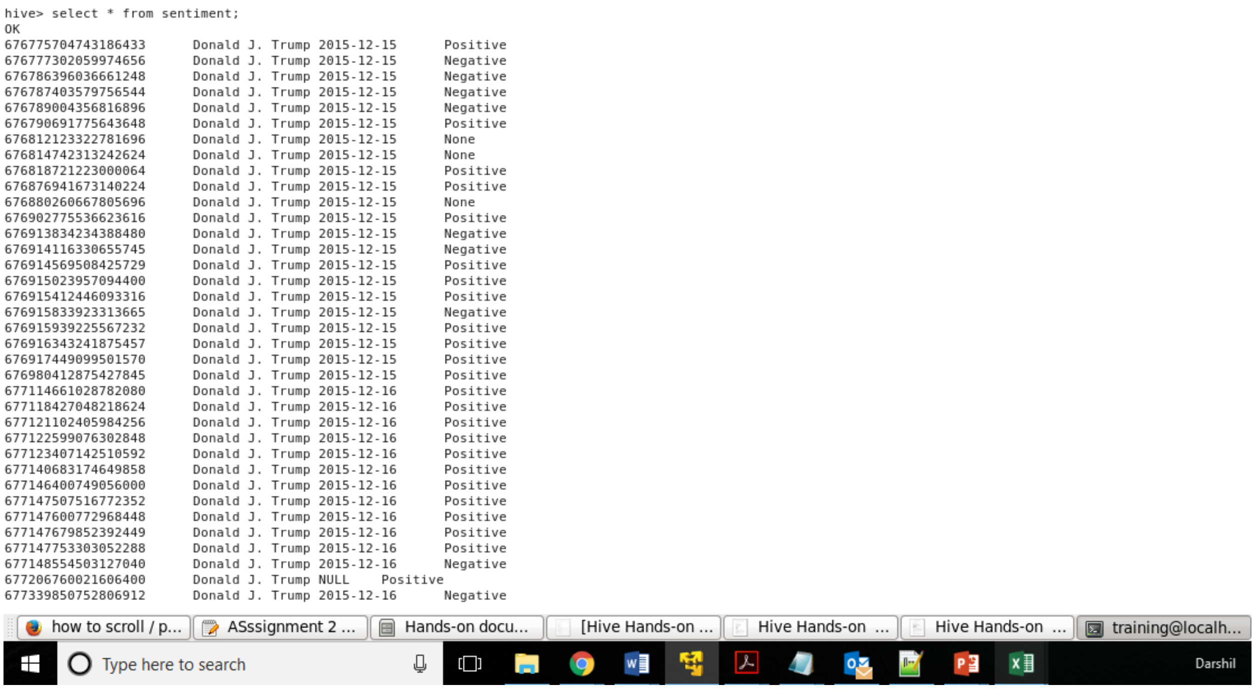
```
hive> CREATE table mapping as SELECT t.tweet_id, t.user_name, t.date, t.word, d.score FROM dictionary_data d RIGHT OUTER JOIN tweet_explode t on (t.word = d.word);
Query ID = training_20180327070909_02631b39-4421-4450-887a-9544a69b11da
Total jobs = 1
Execution log at: /tmp/training/training_20180327070909_02631b39-4421-4450-887a-9544a69b11da.log
2018-03-27 07:09:32     Starting to launch local task to process map join;      maximum memory = 1013645312
2018-03-27 07:09:33     Dump the side-table for tag: 0 with group count: 2477 into file: file:/tmp/training/290e6715-d6cc-452d-84e4-2ffd5cd043dd/hive_2018-03-27_07-09-26_475_631475989613017231
8-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile10--.hashtable
2018-03-27 07:09:33     Uploaded 1 File to: file:/tmp/training/290e6715-d6cc-452d-84e4-2ffd5cd043dd/hive_2018-03-27_07-09-26_475_6314759896130172318-1/-local-10003/HashTable-Stage-4/MapJoin-ma
pfile10--.hashtable (69200 bytes)
2018-03-27 07:09:33     End of local task; Time Taken: 1.544 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_1521697841770_0038, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0038/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0038
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2018-03-27 07:09:42,634 Stage-4 map = 0%,  reduce = 0%
2018-03-27 07:09:51,301 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.7 sec
MapReduce Total cumulative CPU time: 2 seconds 700 msec
Ended Job = job_1521697841770_0038
Moving data to: hdfs://localhost:8020/user/hive/warehouse/mapping
Table default.mapping stats: [numFiles=1, numRows=73625, totalSize=3905233, rawDataSize=3831608]
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1   Cumulative CPU: 2.7 sec   HDFS Read: 3693871 HDFS Write: 3905312 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 700 msec
OK
Time taken: 26.138 seconds
```



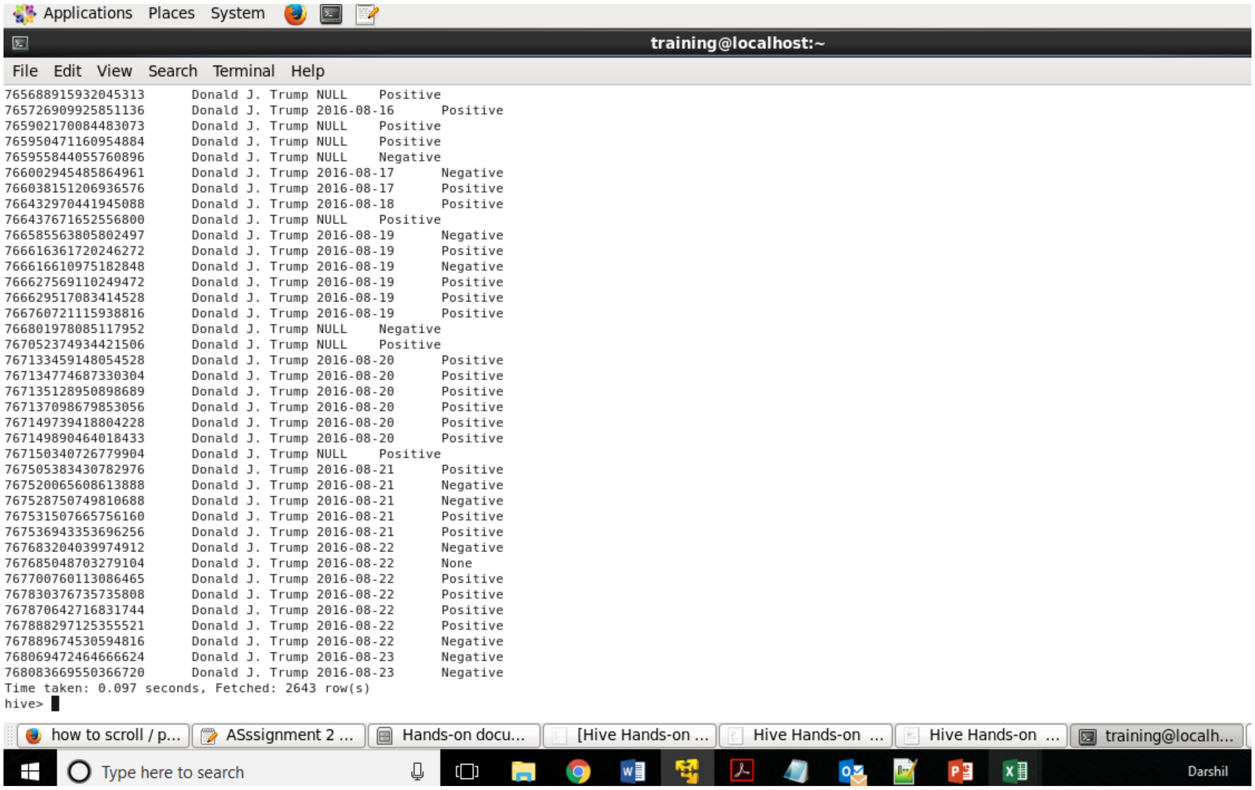
```
hive> CREATE table tweet_score as SELECT tweet_id, user_name, date, SUM(score) as tweet_score FROM mapping GROUP BY tweet_id, user_name, date;
Query ID = training_20180327070909_6cf17085-c024-4b96-88a7-cd0d88b833ce
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated 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_1521697841770_0039, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0039/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0039
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 07:10:05,118 Stage-1 map = 0%,  reduce = 0%
2018-03-27 07:10:12,798 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.97 sec
2018-03-27 07:10:21,213 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.64 sec
MapReduce Total cumulative CPU time: 3 seconds 640 msec
Ended Job = job_1521697841770_0039
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweet_score
Table default.tweet_score stats: [numFiles=1, numRows=3207, totalSize=151187, rawDataSize=147980]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.64 sec   HDFS Read: 3912667 HDFS Write: 151268 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 640 msec
OK
Time taken: 26.142 seconds
```



```
hive> CREATE table sentiment as SELECT tweet_id, user_name, date, CASE WHEN tweet_score > 0 THEN 'Positive' WHEN tweet_score < 0 THEN 'Negative' ELSE 'None' END as sentiment FROM tweet_score w
here tweet_score is not null;
Query ID = training_20180327071010_8a68cfcb-1c94-4bd4-bd86-27854d829167
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_1521697841770_0040, Tracking URL = http://localhost:8088/proxy/application_1521697841770_0040/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1521697841770_0040
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 07:10:36,266 Stage-1 map = 0%,  reduce = 0%
2018-03-27 07:10:43,776 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.66 sec
MapReduce Total cumulative CPU time: 1 seconds 660 msec
Ended Job = job_1521697841770_0040
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/.hive-staging_hive_2018-03-27_07-10-28_280_9063473340872916349-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/sentiment
Table default.sentiment stats: [numFiles=1, numRows=2643, totalSize=141969, rawDataSize=139326]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.66 sec   HDFS Read: 155013 HDFS Write: 142048 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 660 msec
OK
Time taken: 16.868 seconds
hive> 
```



```
hive> select * from sentiment;
OK
676775704743186433      Donald J. Trump 2015-12-15      Positive
676777302059974656      Donald J. Trump 2015-12-15      Negative
676786396036661248      Donald J. Trump 2015-12-15      Negative
676787403579756544      Donald J. Trump 2015-12-15      Negative
676789004356816896      Donald J. Trump 2015-12-15      Negative
676790691775643648      Donald J. Trump 2015-12-15      Positive
676812123322781696      Donald J. Trump 2015-12-15      None
676814742313242624      Donald J. Trump 2015-12-15      None
676818721223000064      Donald J. Trump 2015-12-15      Positive
676876941673140224      Donald J. Trump 2015-12-15      Positive
676880260667805696      Donald J. Trump 2015-12-15      None
676902775536623616      Donald J. Trump 2015-12-15      Positive
676913834234388480      Donald J. Trump 2015-12-15      Negative
676914116330655745      Donald J. Trump 2015-12-15      Negative
676914569508425729      Donald J. Trump 2015-12-15      Positive
676915023957094400      Donald J. Trump 2015-12-15      Positive
676915412446093316      Donald J. Trump 2015-12-15      Positive
676915833923313665      Donald J. Trump 2015-12-15      Negative
676915939225567232      Donald J. Trump 2015-12-15      Positive
676916343241875457      Donald J. Trump 2015-12-15      Positive
676917449099501570      Donald J. Trump 2015-12-15      Positive
676980412875427845      Donald J. Trump 2015-12-15      Positive
677114661028782080      Donald J. Trump 2015-12-16      Positive
677118427048218624      Donald J. Trump 2015-12-16      Positive
677121102405984256      Donald J. Trump 2015-12-16      Positive
677122599076302848      Donald J. Trump 2015-12-16      Positive
677123407142510592      Donald J. Trump 2015-12-16      Positive
677140683174649858      Donald J. Trump 2015-12-16      Positive
677146400749056000      Donald J. Trump 2015-12-16      Positive
677147507516772352      Donald J. Trump 2015-12-16      Positive
677147600772968448      Donald J. Trump 2015-12-16      Positive
677147679852392449      Donald J. Trump 2015-12-16      Positive
677147753303052288      Donald J. Trump 2015-12-16      Positive
677148554503127040      Donald J. Trump 2015-12-16      Negative
677206760021606400      Donald J. Trump NULL    Positive
677339850752806912      Donald J. Trump 2015-12-16      Negative
```



```
765688915932045313      Donald J. Trump NULL    Positive
765726909925851136      Donald J. Trump 2016-08-16      Positive
765902170084483073      Donald J. Trump NULL    Positive
765950471160954884      Donald J. Trump NULL    Positive
765955844055760896      Donald J. Trump NULL    Negative
766002945485864961      Donald J. Trump 2016-08-17      Negative
766038151206936576      Donald J. Trump 2016-08-17      Positive
766432970441945088      Donald J. Trump 2016-08-18      Positive
766437671652556800      Donald J. Trump NULL    Positive
766585563805802497      Donald J. Trump 2016-08-19      Negative
766616361720246272      Donald J. Trump 2016-08-19      Positive
766616610975182848      Donald J. Trump 2016-08-19      Negative
766627569110249472      Donald J. Trump 2016-08-19      Positive
766629517083414528      Donald J. Trump 2016-08-19      Positive
766760721115938816      Donald J. Trump 2016-08-19      Positive
766801978085117952      Donald J. Trump NULL    Negative
767052374934421506      Donald J. Trump NULL    Positive
767133459148054528      Donald J. Trump 2016-08-20      Positive
767134774687330304      Donald J. Trump 2016-08-20      Positive
767135128950898689      Donald J. Trump 2016-08-20      Positive
767137098679853056      Donald J. Trump 2016-08-20      Positive
767149739418804228      Donald J. Trump 2016-08-20      Positive
767149890464018433      Donald J. Trump 2016-08-20      Positive
767150340726779904      Donald J. Trump NULL    Positive
767505383430782976      Donald J. Trump 2016-08-21      Positive
767520065608613888      Donald J. Trump 2016-08-21      Negative
767528750749810688      Donald J. Trump 2016-08-21      Negative
767531507665756160      Donald J. Trump 2016-08-21      Positive
767536943353696256      Donald J. Trump 2016-08-21      Positive
767683204039974912      Donald J. Trump 2016-08-22      Negative
767685048703279104      Donald J. Trump 2016-08-22      None
767700760113086465      Donald J. Trump 2016-08-22      Positive
767830376735735808      Donald J. Trump 2016-08-22      Positive
767870642716831744      Donald J. Trump 2016-08-22      Positive
767888297125355521      Donald J. Trump 2016-08-22      Positive
767889674530594816      Donald J. Trump 2016-08-22      Negative
768069472464666624      Donald J. Trump 2016-08-23      Negative
768083669550366720      Donald J. Trump 2016-08-23      Negative
Time taken: 0.097 seconds, Fetched: 2643 row(s)
hive> 
```

Q.2. Do you find any problem in the way sentiment analysis was performed in the previous question? If so, how will you improve it?

- The size of the dictionary should be larger to cover all the words. Because of the scarcity of words, there is a NULL mapping issue while joining it with the tweet file.
- There are problems in handling comparison. Ex. “My bag is better than yours”. It can be classified as positive for both you and me. However, it understand the comparison.
- Problems in recognizing an entity: e.g. “I loathe Walmart.com, but I love Amazon.com”. A simple approach will label it as neutral, however, it carries a specific sentiment for both the entities present in the statement.
- The rating given to the words have no criteria. The analysis will be biased as one person’s positive opinion like ‘good’ might be more positive than other’s ‘best’.
- Machine learning algorithms can be used for sentiment analysis. Techniques like Naïve Bayes can help system learn the emotional aspect of the sentences.
- Keep revising the dictionary and keep ranges of scores rather than discrete numbers in the dictionary.