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';**

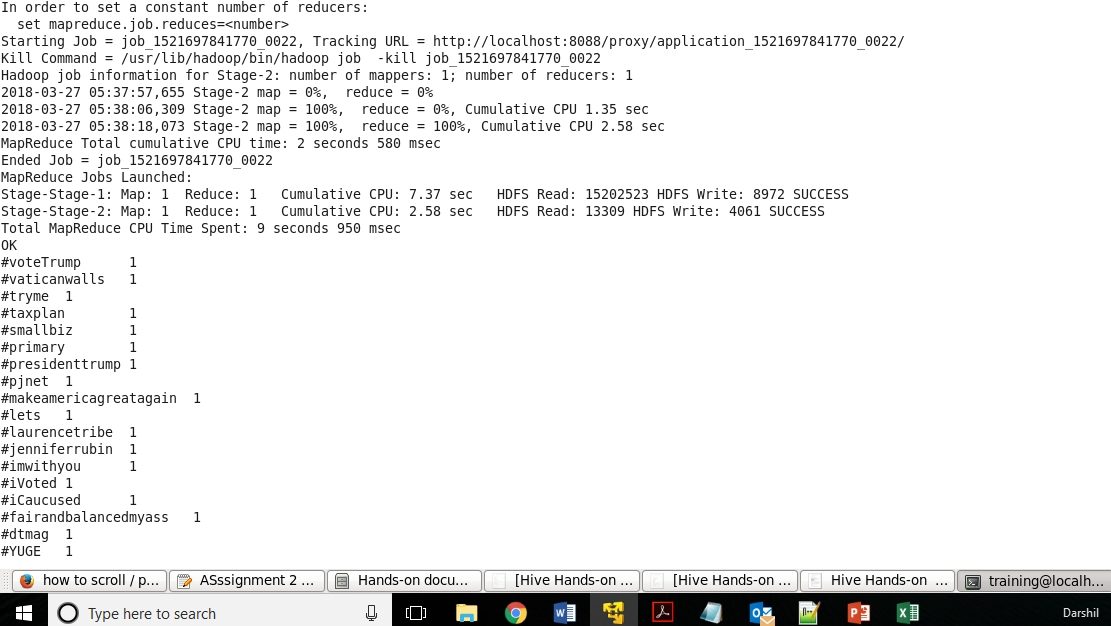
**Loading data from local path into the created table**

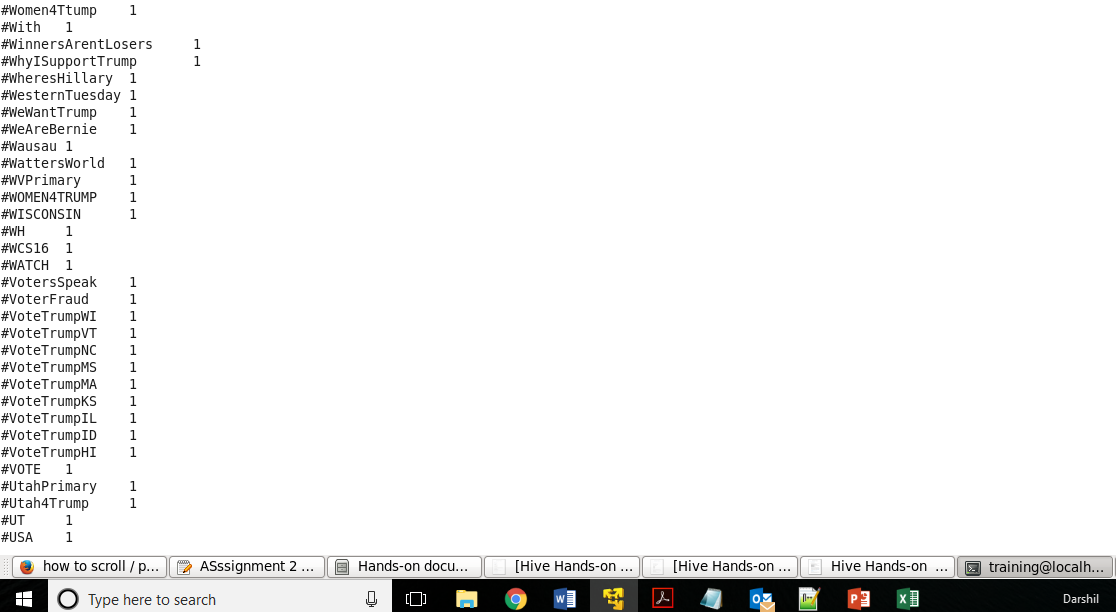
**load data local inpath 'Desktop/tweet\_final.json' overwrite into table tweet;**

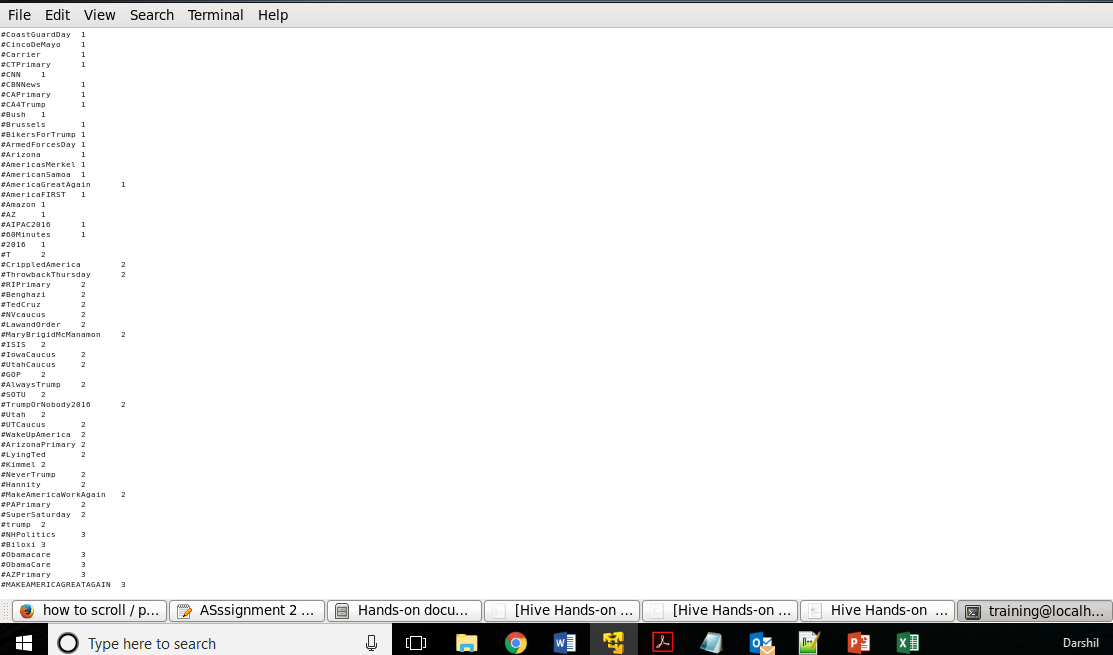
**Question 1.a. What are the hashtags used and how many times each are used?**

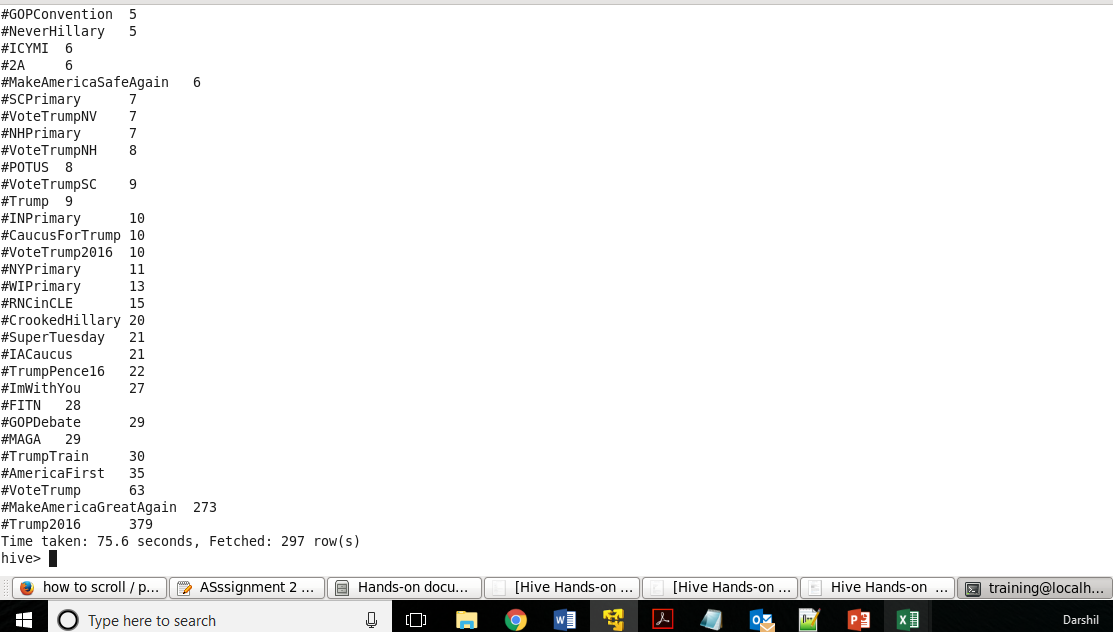
**SELECT h\_word, count(1) as word\_count from tweet LATERAL VIEW explode(split(regexp\_replace(trim(text),"[^#A-Za-z0-9]"," "), ' ')) text\_explode as h\_word WHERE h\_word rlike "^#[a-zA-Z0-9]+$" GROUP BY h\_word ORDER BY word\_count;**

****

****

****

****

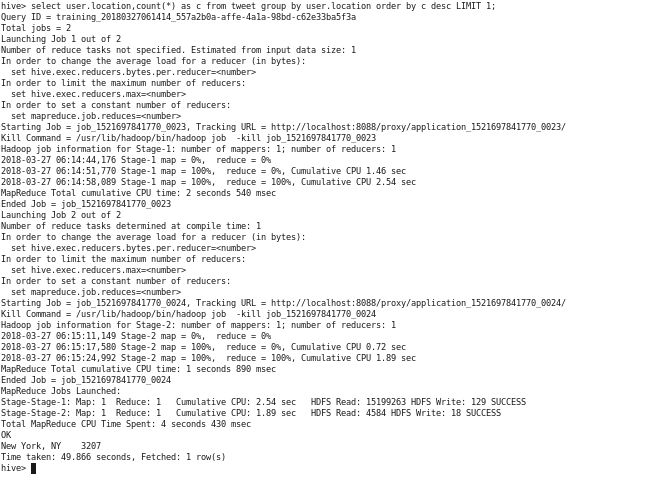
****

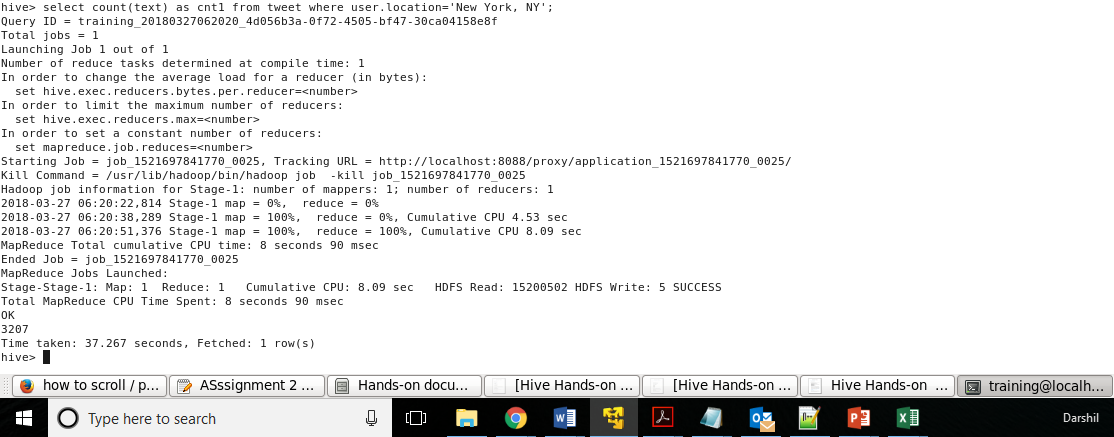
**Question 1.b. Which State have the most active users and how many tweets are posted by State?**

**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';**

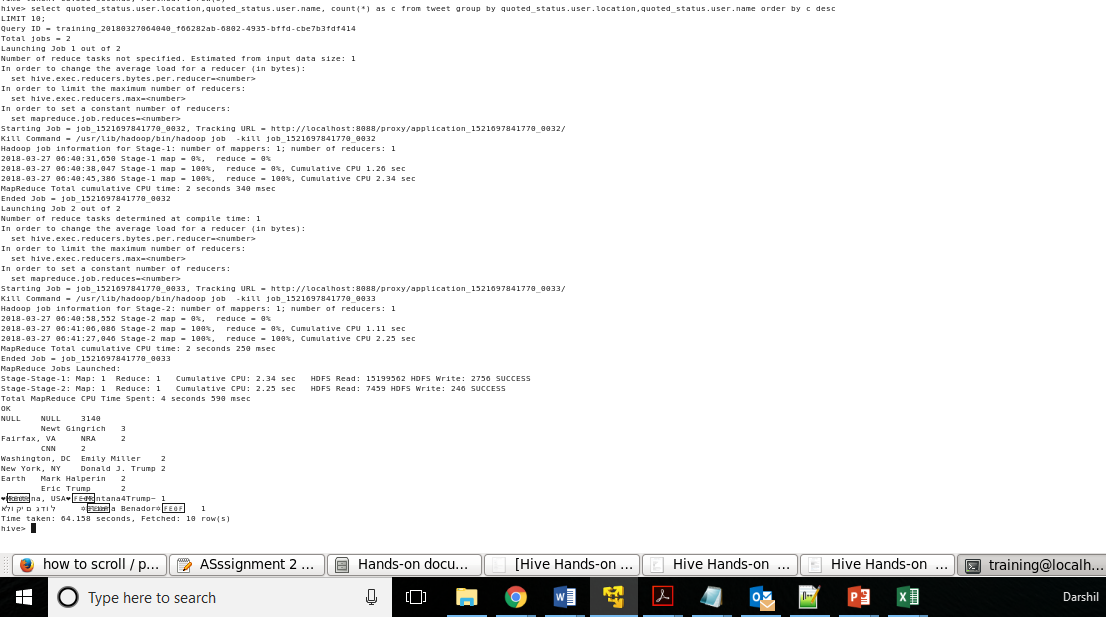
New York has the most active users

****



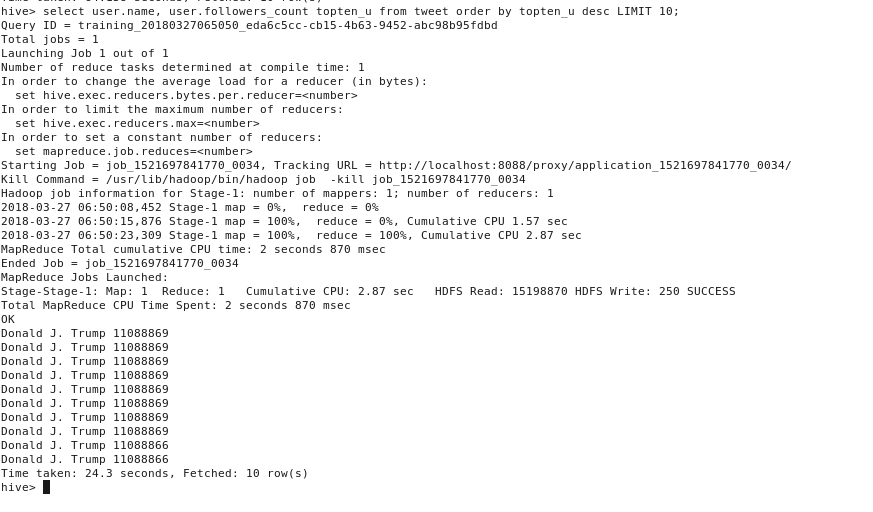
**Users from other States can be found via the user within the quoted\_status in the json file:**

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

****

**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;**

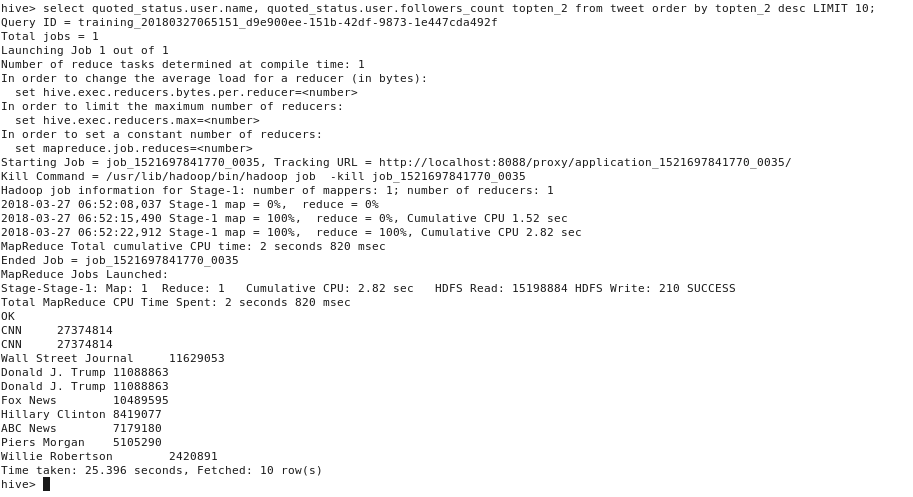


Donald Trump is repeated all top ten times in terms of **followers count.**

**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:**

****

**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;**

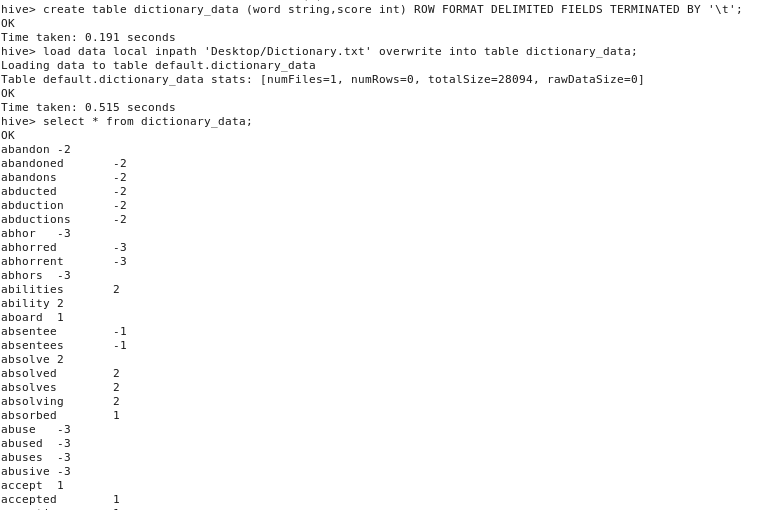
**create table tweet\_explode as select tweet\_id, user\_name, from\_unixtime(date,'yyyy-MM-dd') as date, word from tweet\_information LATERAL VIEW explode(split(regexp\_replace(lower(text),"[^#A-Za-z0-9]"," "), ' ')) text\_x as word;**

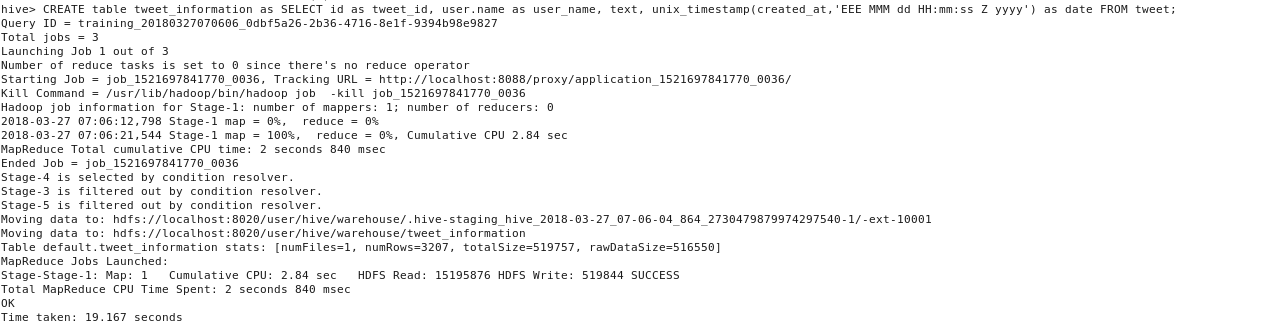
**CREATE table mapping as SELECT t.tweet\_id, t.user\_name, t.date, t.word, d.score FROM dictionary\_data d RIGHT OUTER JOIN tweet\_explode t on (t.word = d.word);**

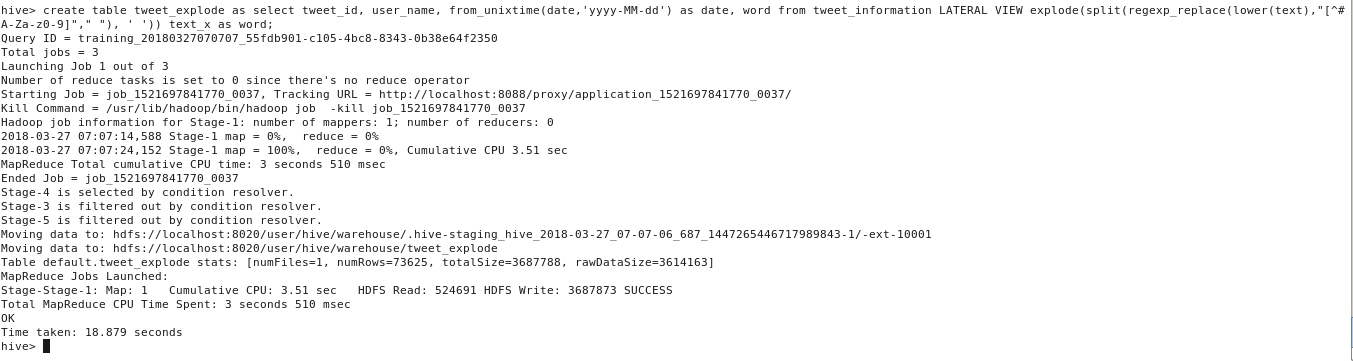
**CREATE table tweet\_score as SELECT tweet\_id, user\_name, date, SUM(score) as tweet\_score FROM mapping GROUP BY tweet\_id, user\_name, date;**

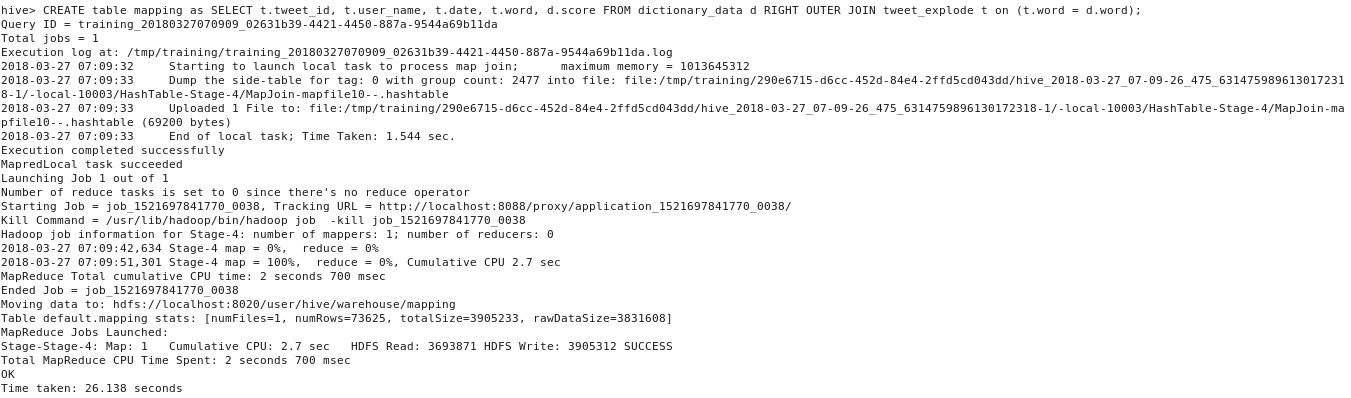
**CREATE table sentiment as SELECT tweet\_id, user\_name, date, CASE WHEN tweet\_score > 0 THEN 'Positive' WHEN tweet\_score < 0 THEN 'Negative' ELSE 'None' END as sentiment FROM tweet\_score where tweet\_score is not null;**

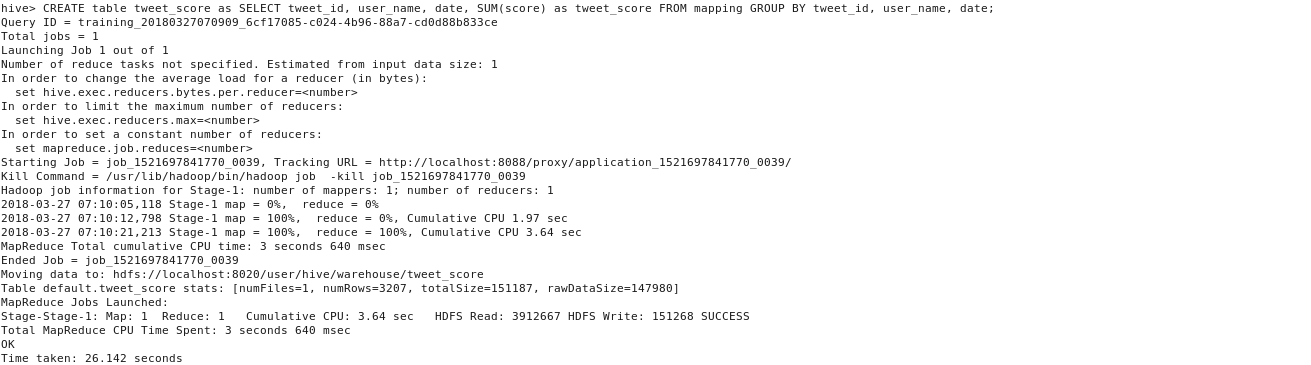
**Select \* from sentiment;**

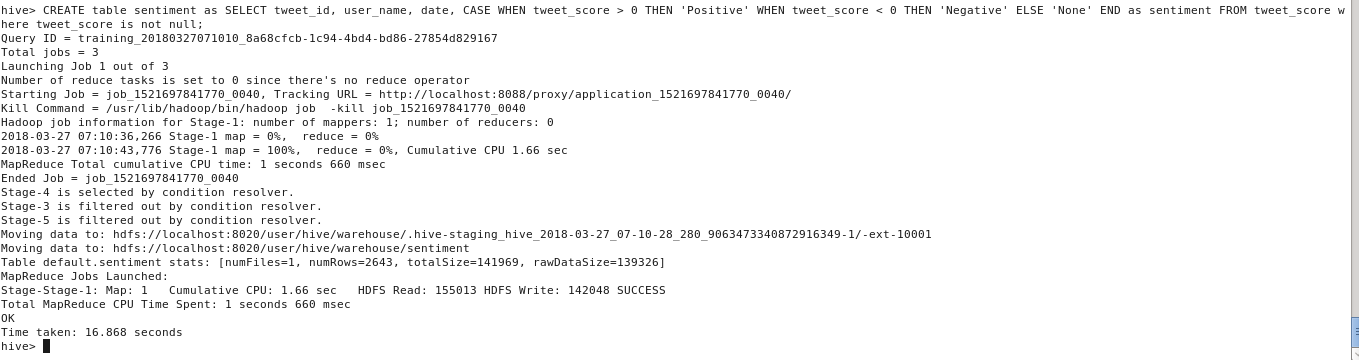
****

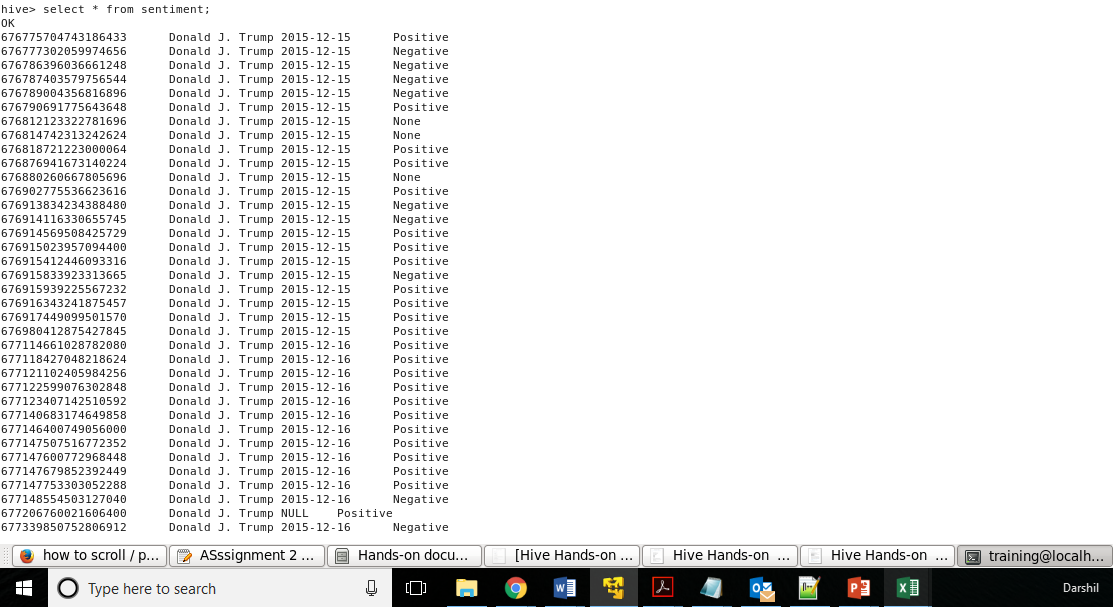
****

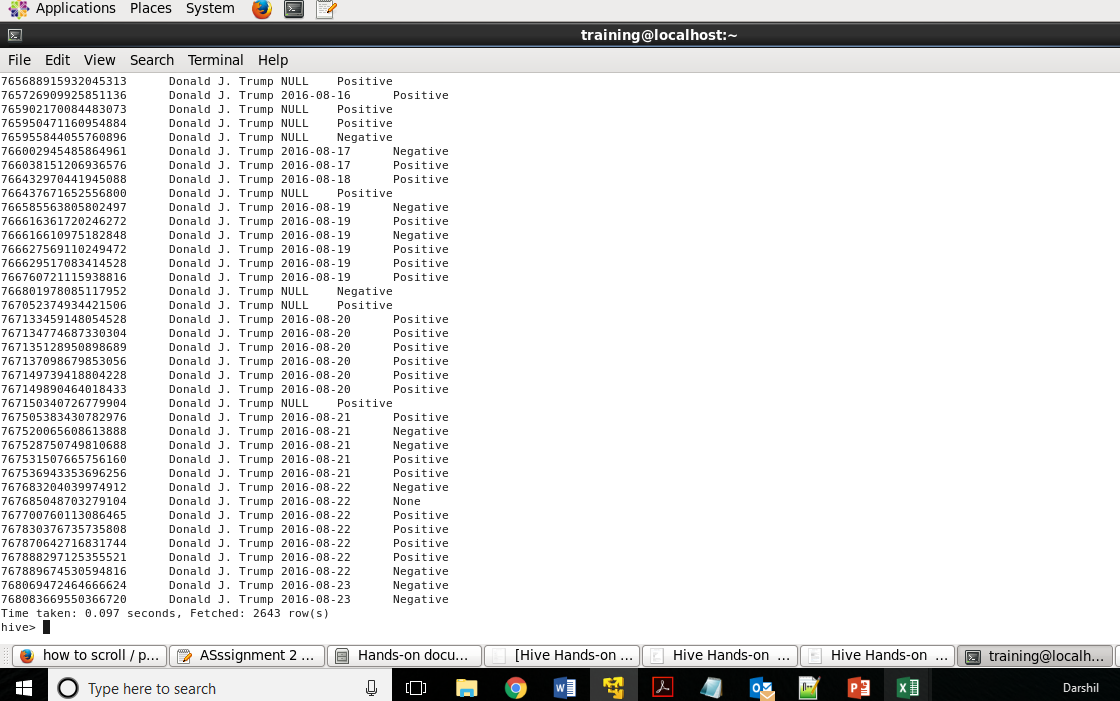
****

****

****

****

****

****

**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.