

TWITTER ANALYSIS

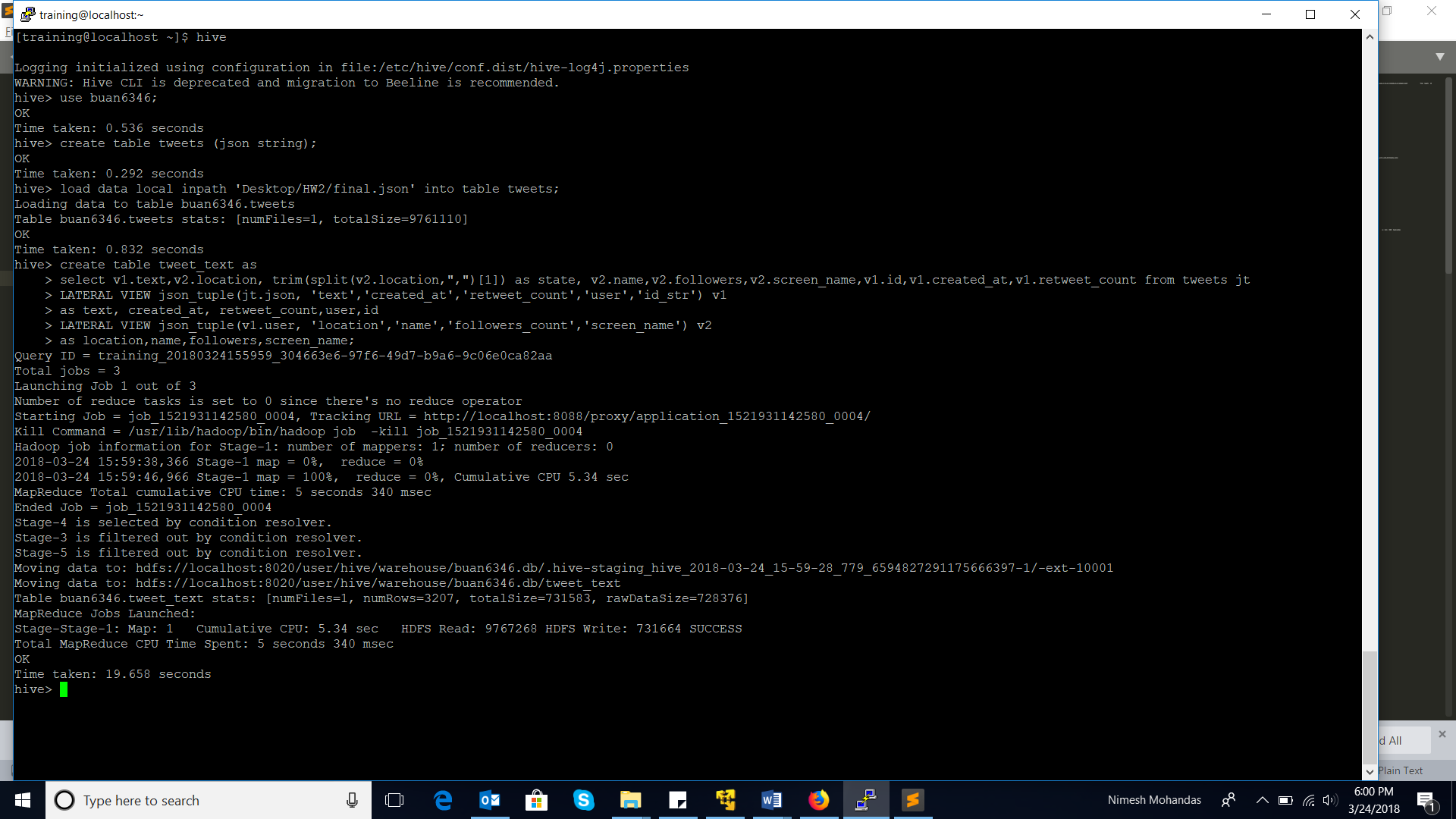
BUAN6346.002 HW2



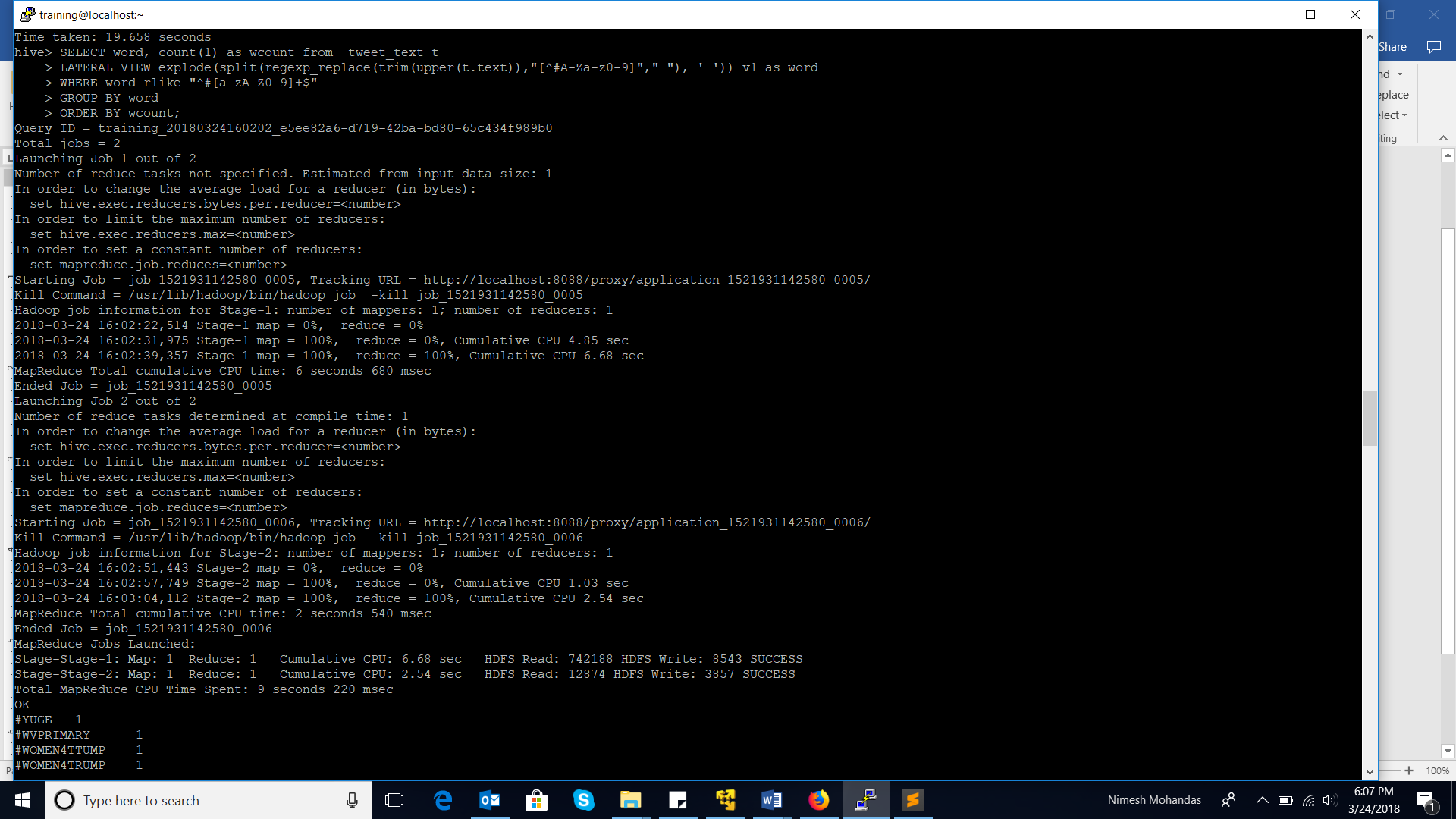
Nimesh Mohandas

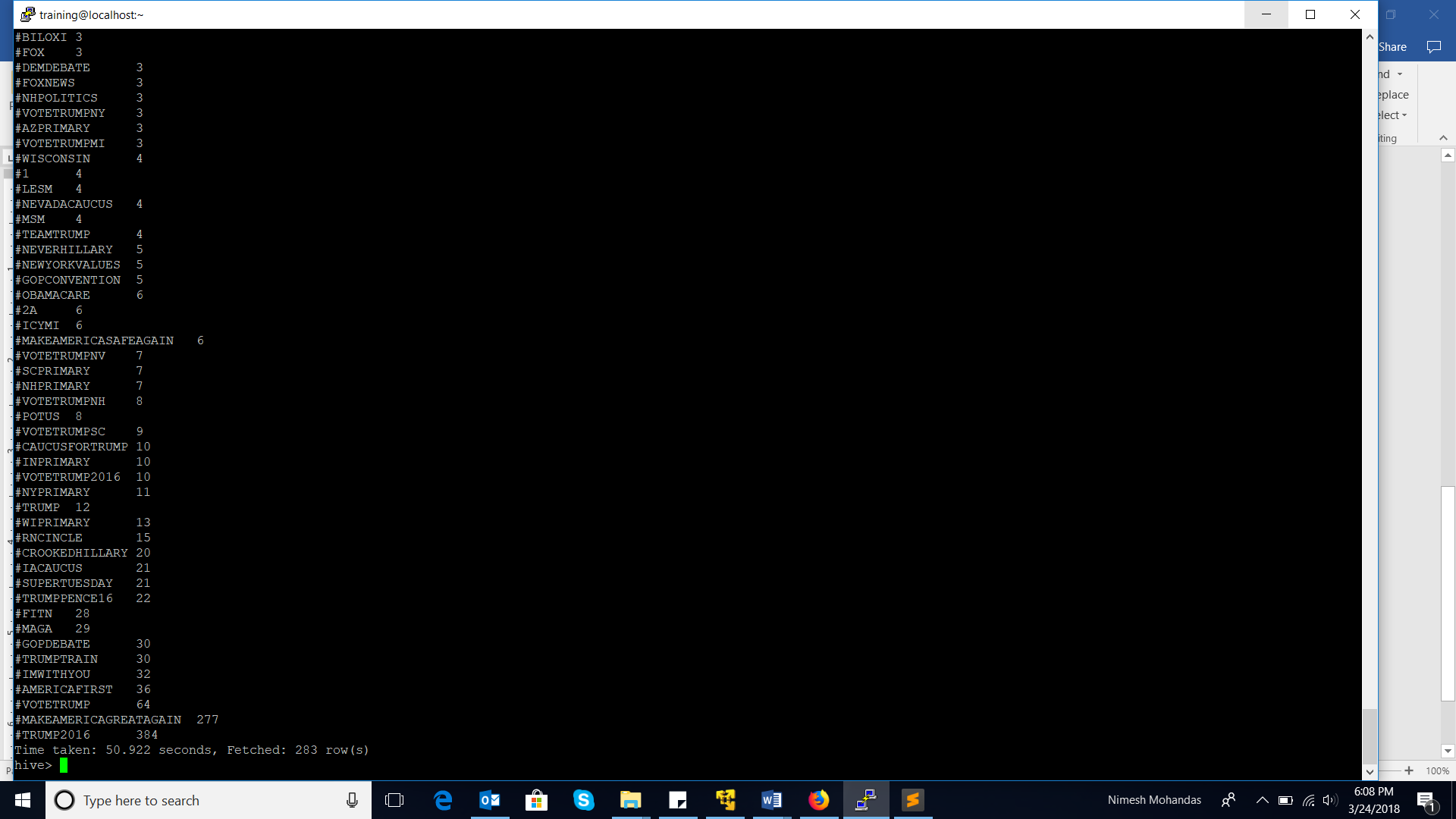
Nxm172730

Load the Twitter data into a hive table:



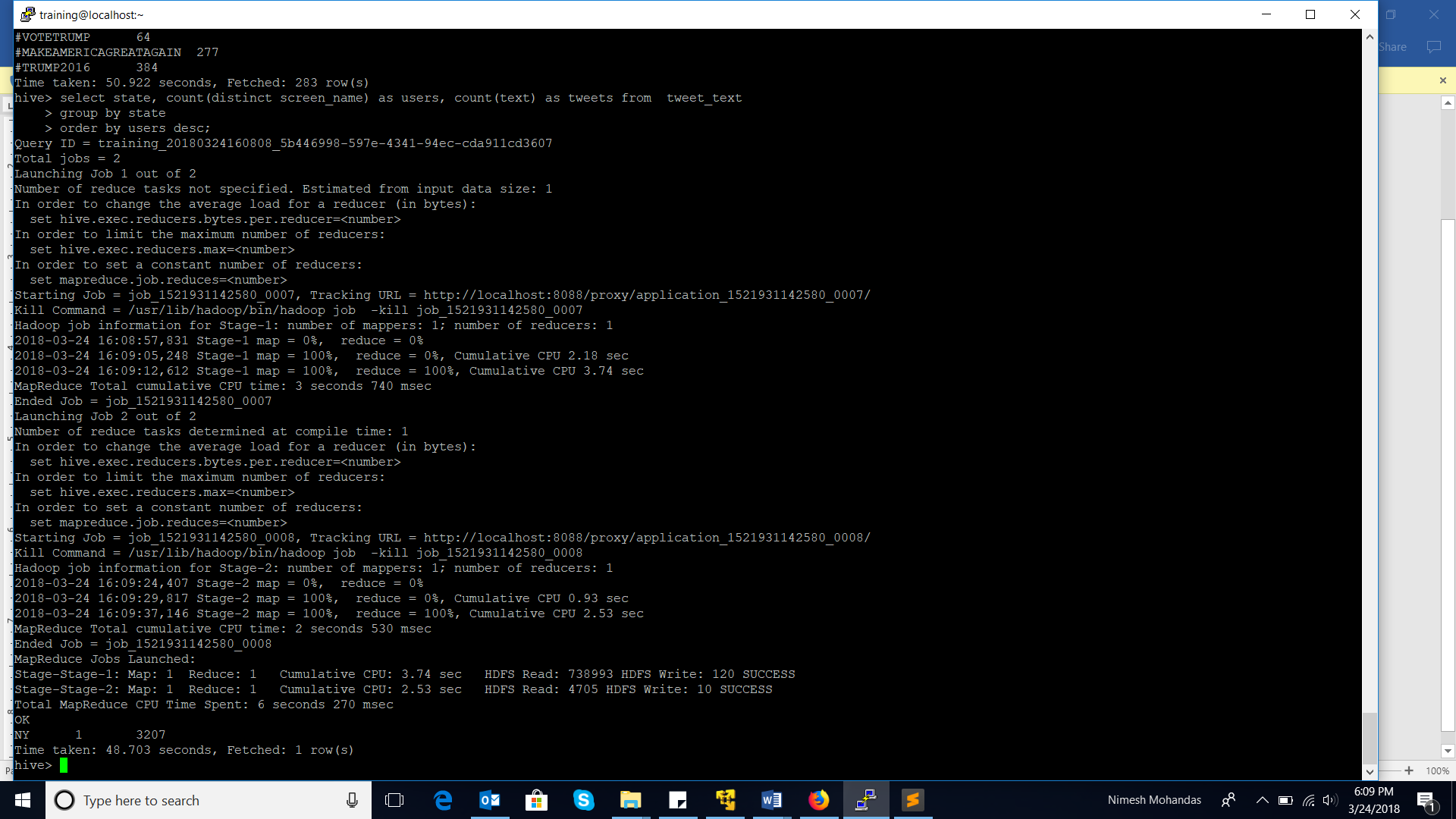
1.a) What are the hashtags used and how many times each hashtag is used?





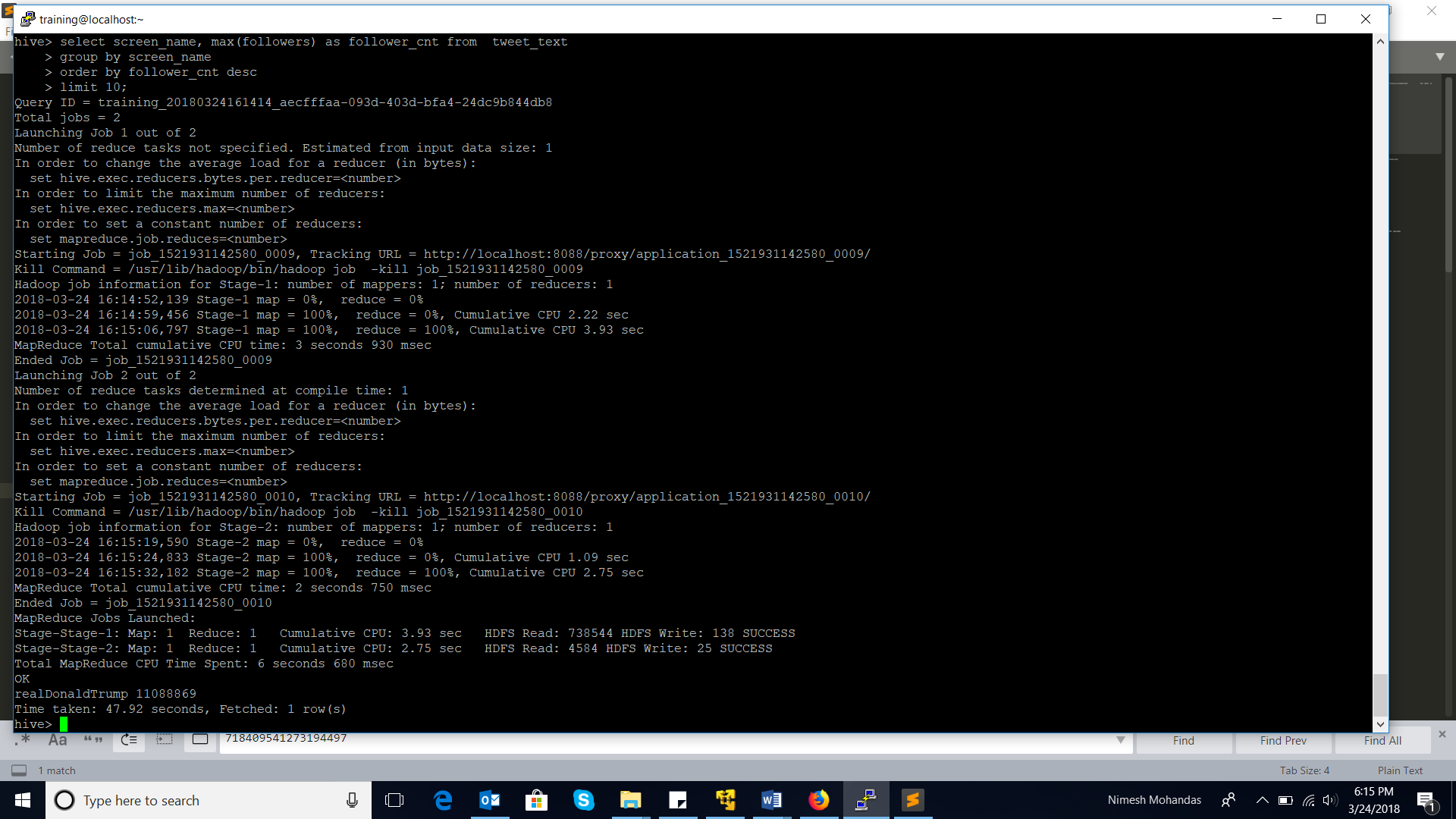
There are 283 distinct (#tags are in lowercase and uppercase, taking both as same) hash tags. #TRUMP2016 being used maximum – 384 times

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



The data provided has only one state (NY, New York) and only one user from the state – realDonaldTrump.

1.c) Based on the user’s followers count, who are the top ten users who have tweeted?

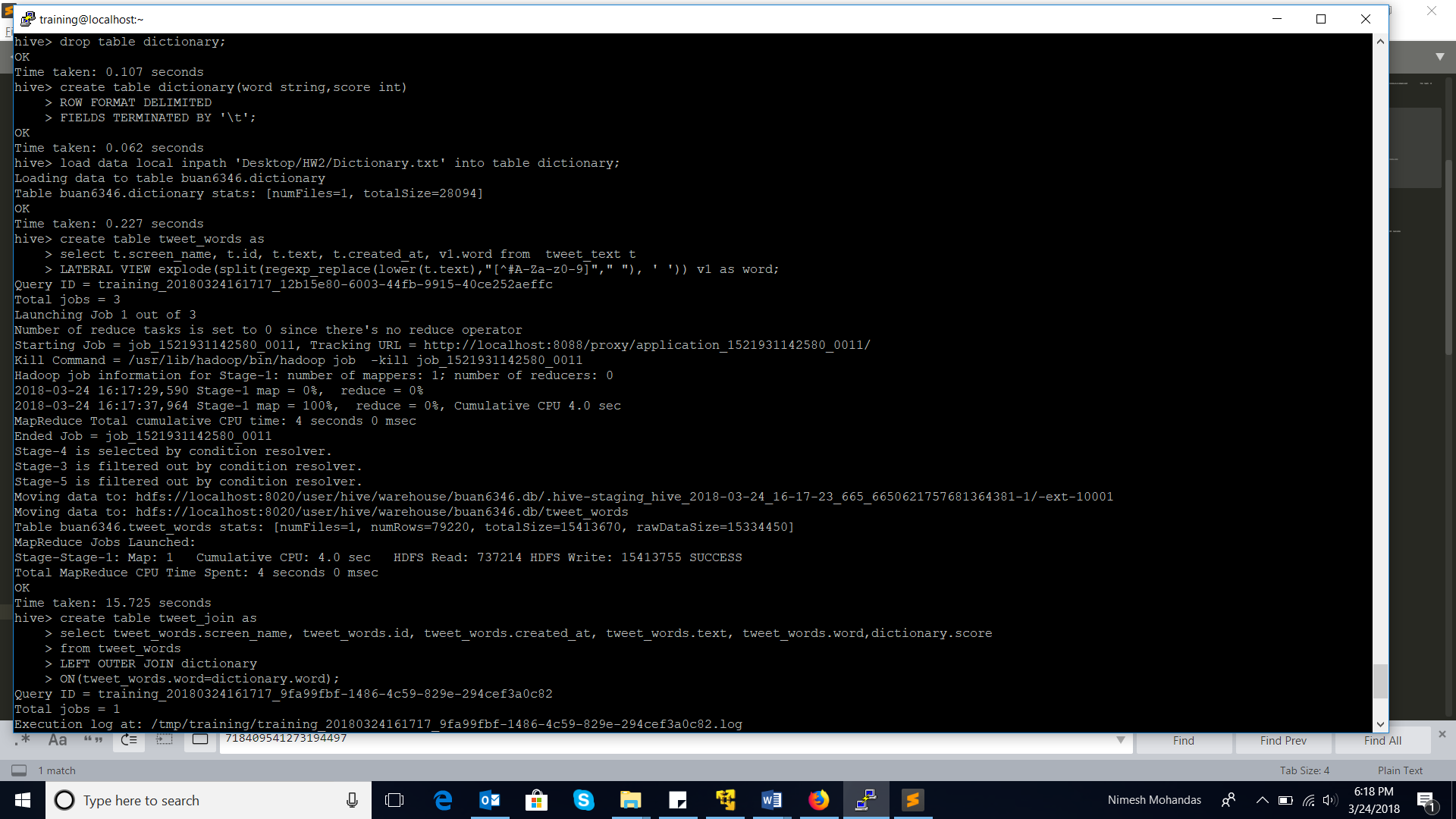


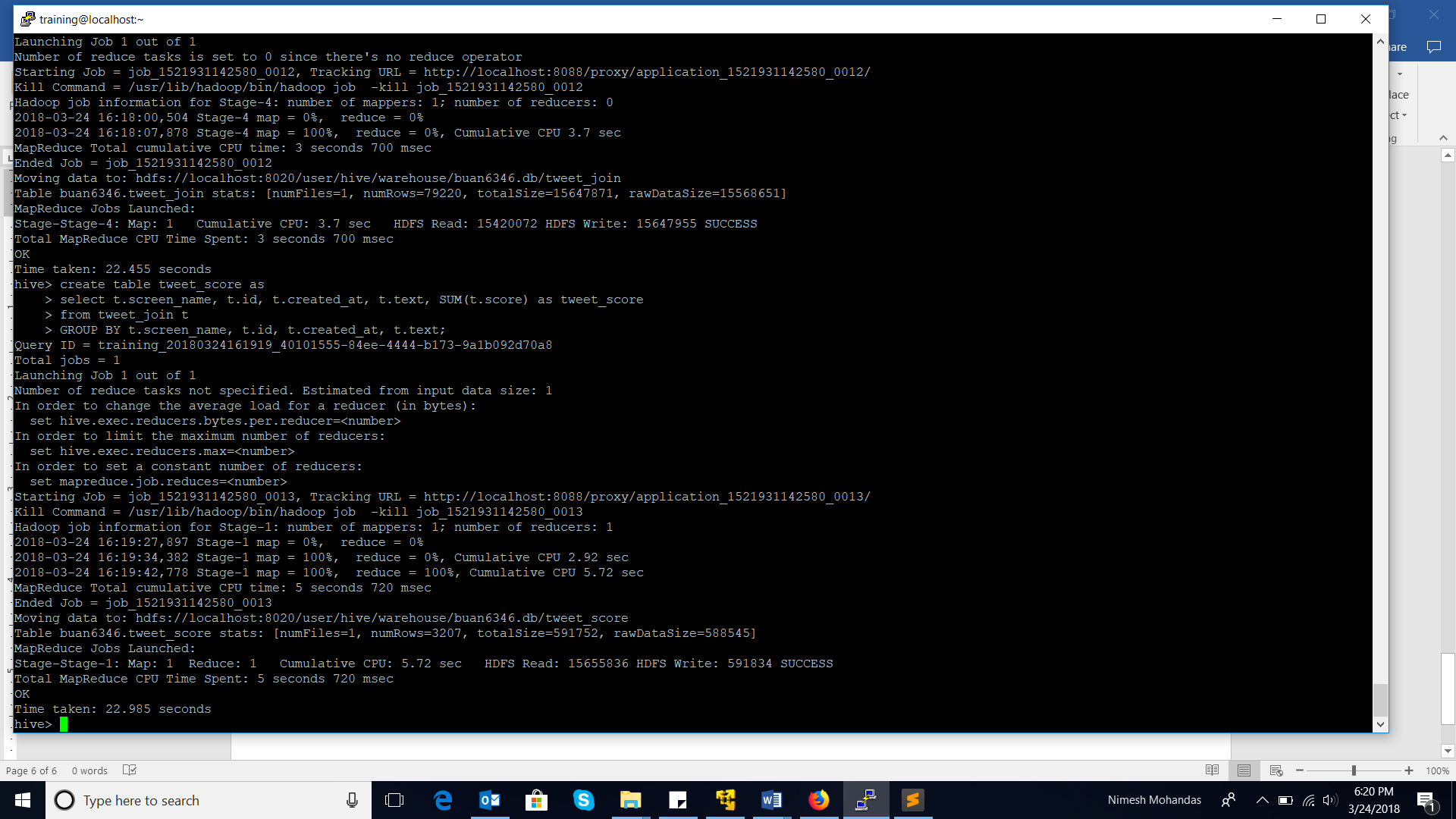
The data had only one user – realDonaldTrump, and he has 11,088,869 followers.

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

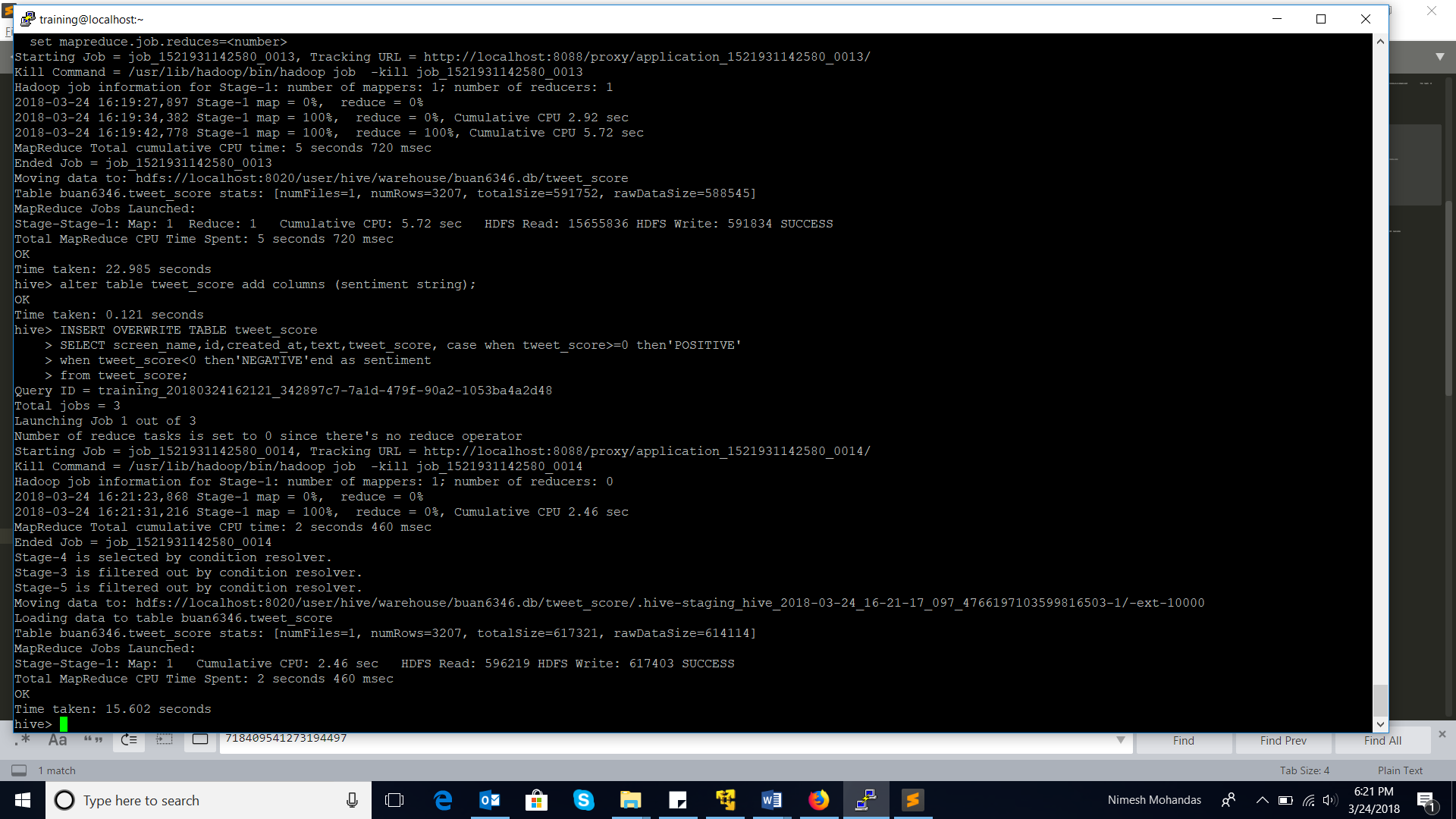
Load the dictionary data.

Explode the tweet and assign score for each word in the tweet, as per the dictionary provided. Sum up the scores for all the words in the tweet to get a total Tweet-score. Based on the sign of the score, we can decide the polarity of the tweet.

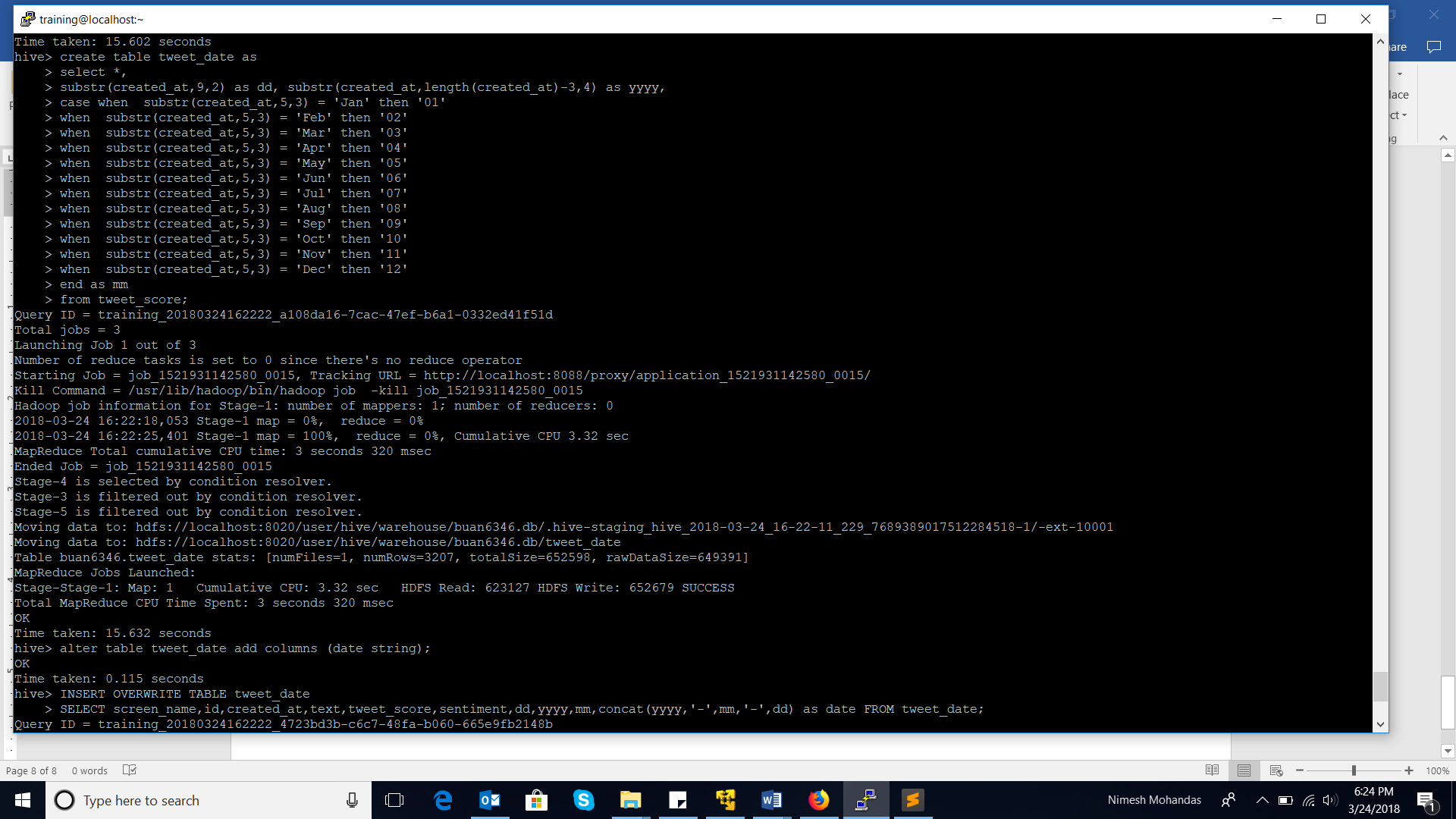


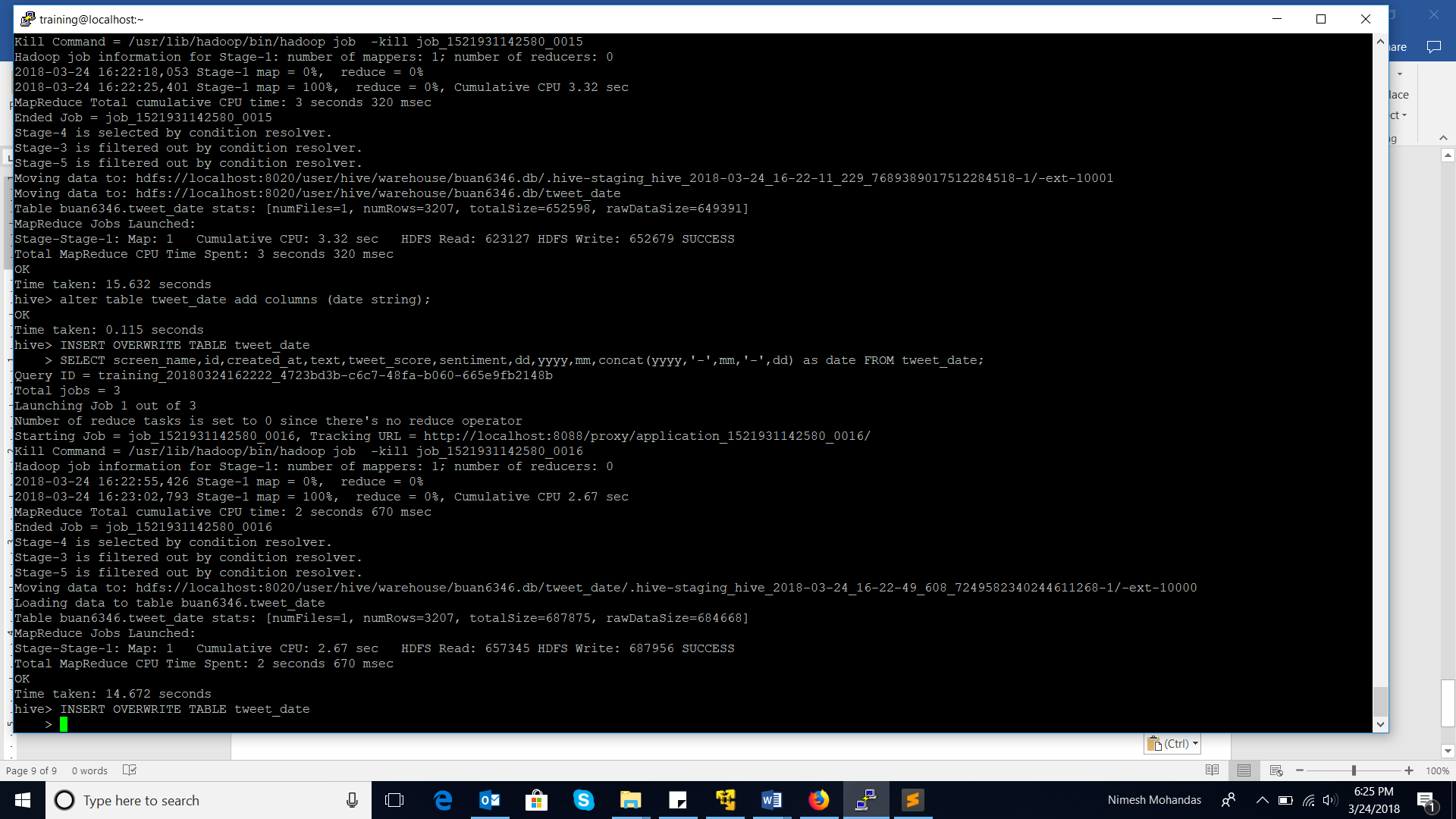


If the total Tweet score is negative then the tweet is tagged as a negative tweet, and if the tweet-score is positive we tag the tweet as a positive tweet.

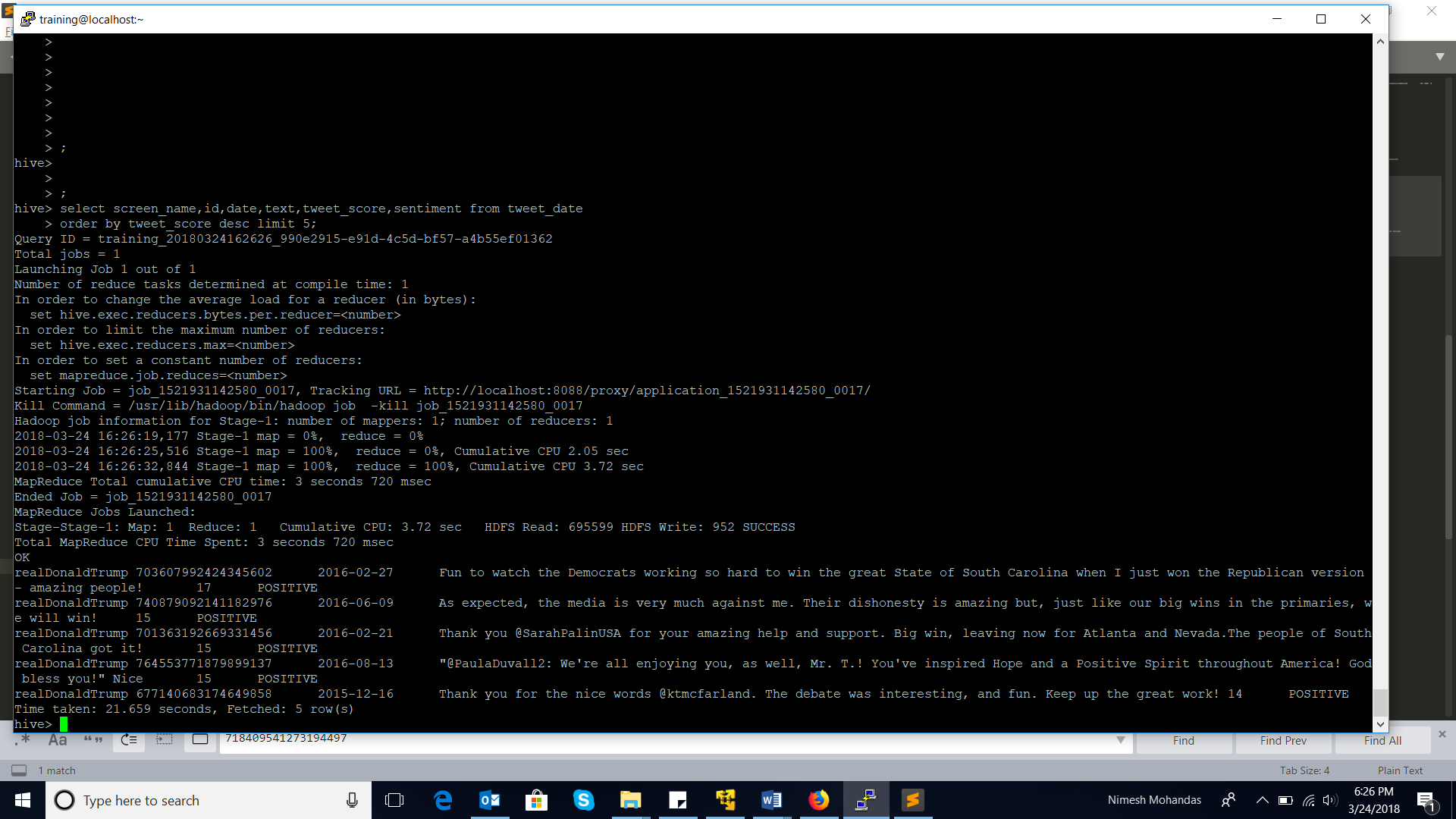


Make the required change for the required date-format.





Output final table with tweet, tweet-score, and sentiment.



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 way we are analyzing the sentiment of a tweet, is a very crude form. The accuracy of this method is questionable, as in the current process we:

* Splitting the tweet into its words (ignoring all the special characters)
* Assigning a score to each word based on a dictionary (negative words scored negative and positive words scored positive)
* Sum-up the scores of the words in the tweet to get a tweet-score
* Based on the polarity of the tweet-score, we decide the sentiment of the Tweet

Following are the gaps in the current method that I have figured out:

* The dictionary is not exhaustive, and it is difficult to make it exhaustive. People do not necessarily use formal language in the tweets. People use slangs, acronyms, and write other language using English scripts
* People write their feelings and thoughts on twitter, i.e. context is required before we make a judgement about a sentence. Many words in English mean different and polarity differ according to the situation in which they are used. A positively scored word can be used in a very negative way. Presence of certain words and in combination with other words can totally change the polarity of the sentence
* Sometimes in the tweet we mention others by their screen-name or nicknames. These do not have any score in the dictionary, making the score of the words null, and when we add scores of other words to NULL, we get NULL, making the whole tweet-score equal to NULL