Assignment 5 and 6

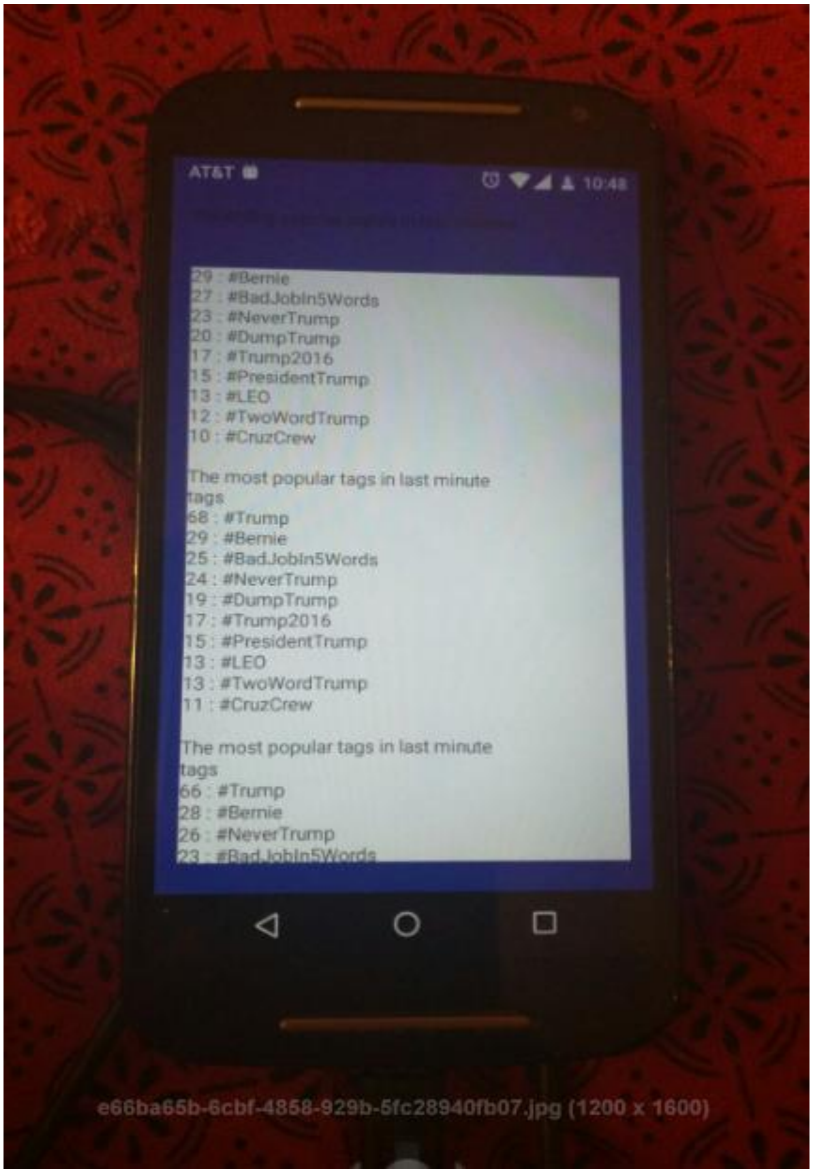
Q-1

I have collected the twitter tweet for by filtering it with following keywords as Oscar award and primary presidential election were there.

```
"oscar","leo","Primary","election","clinton","cruz","trump","sanders"
```

I kept interval of 1 Minute and collecting the popular hashtags generated during last minute. I am sending this hash tag and its count result to the smart phone application.

Below is the output of how it looks to the mobile app screen. It keeps appending on the screen and when we scroll down we will get the latest tweets. First part is number of count and second part is hashtag.



```
29 : #Bernie
27 : #BadJobIn5Words
23 : #NeverTrump
20 : #DumpTrump
17 : #Trump2016
15 : #PresidentTrump
13 : #LEO
12 : #TwoWordTrump
10 : #CruzCrew

The most popular tags in last minute
tags
68 : #Trump
29 : #Bernie
25 : #BadJobIn5Words
24 : #NeverTrump
19 : #DumpTrump
17 : #Trump2016
15 : #PresidentTrump
13 : #LEO
13 : #TwoWordTrump
11 : #CruzCrew

The most popular tags in last minute
tags
66 : #Trump
28 : #Bernie
26 : #NeverTrump
23 : #BadJobIn5Words
```

Q-2 : For this task I have divided it into 3 steps. First step is of collecting the dataset, second step is of examining tweet and train the model and the third and last step is to apply the model and predict the data. Below is the brief explanation.

Collect the data:

This stage collects the streaming data and stores it to the file. It collects some initial number of tweets in regular interval and store it to the file. Program has variable set-up to control the number of tweets, regular interval and the number of output file for individual intervals to be generated as per RDD partition.

Train a model:

After collecting and analyzing some random tweet, program will use spark machine learning library to perform k-means clustering to cluster the tweets in different set. Here the Number of cluster and iteration are configurable.

First it will featurize the tweet text using HashingTF Spark MLIB class. Then k-means creates the cluster and runs same process for defined number of iterations.

Apply and predict data

Spark streaming is used to filter the data and take only those data classified as defined cluster.

modelDirectory - This the directory where the model that was trained in part 2 was persisted.
clusterNumber - This is the cluster you want to select from part 2. Only tweets that match this language cluster will be printed out.

Algorithm runs in 4 steps

1. Load up a Spark Streaming Context.
2. Create a Twitter DStream and map it to grab the text.
3. Load up the K-Means model that was trained in step 2.
4. Apply the model on the tweets, filtering out only those that match the specified cluster, and print the matching tweets.

Reference: https://databricks.gitbooks.io/databricks-spark-reference-applications/content/twitter_classifier