

Lab 5 and 6 report

Task 1

There were two tasks to be performed this time. The first task was to create an Android client and interact with Spark engine to perform word count on Twitter streaming data.

To achieve this task the following was done.

- Created an Android client.
- Implemented a Socket server at the client end.
- On press of the button the server would start and would be ready to listen to data.
- Created Twitter API key and secret.
- Utilized the Twitter Stream to get Tweets and filter the hashtags out of it.
- Then performed count of Tweets for 5 second window to get the the count of each hashtag in descending order.
- Send the top 5 count of hashtags to Android through Socket and display the result.

The screen shots for the task are given below.

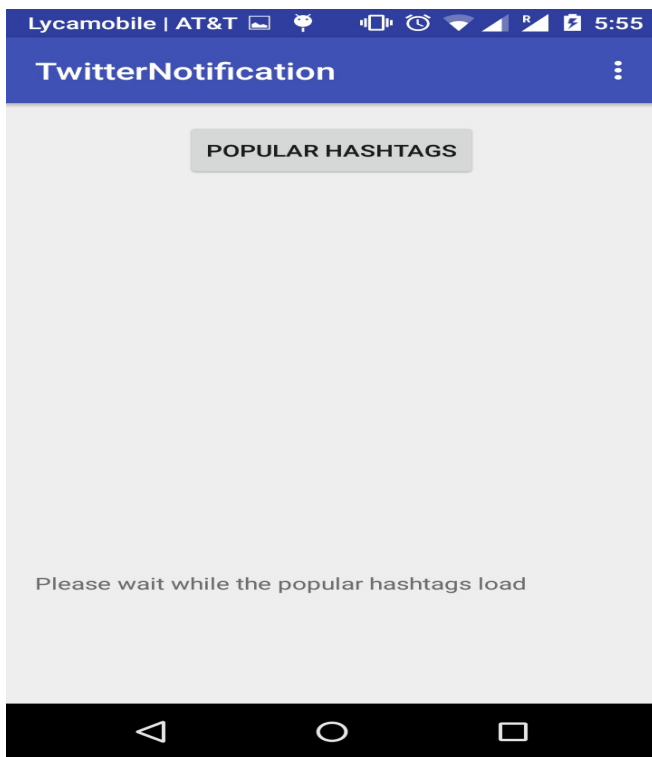


Fig 1 : Inital screen when the server starts

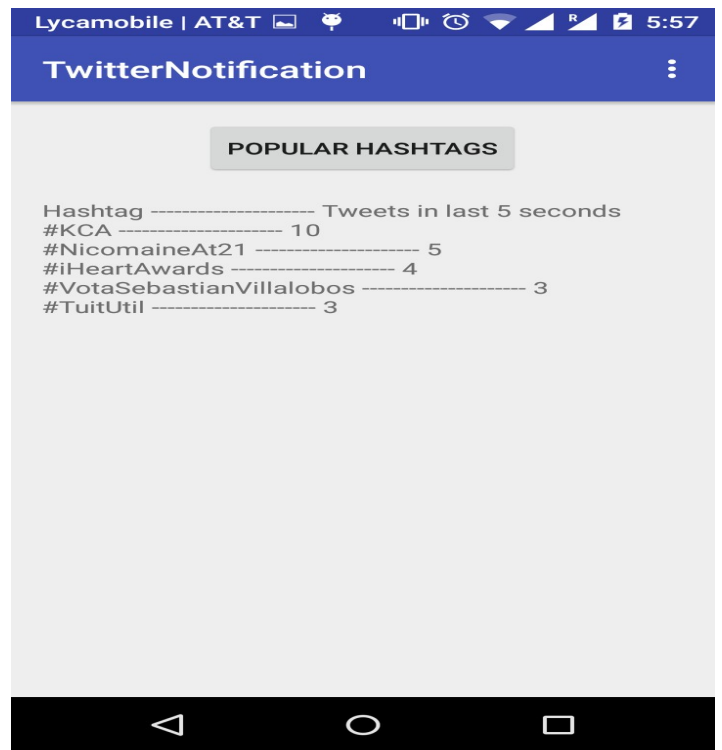


Fig 2 : The screen showing the top 5 trending hashtags with the count of Tweets.

Task 2

This task was to create a classification system that can categorize Tweets to predefined classes. In my case I had the training data which was set of Tweets on the following categories

1. #hillary
2. #uselections
3. #DonaldTrump

I collected a subset of live Tweets and allocated it as testing data.

Then I processed the testing to the FeatureVector algorithm to get it to predict the class to which the Tweet belongs to.

PFB the screenshot for the final output which classifies a particular set of Tweets to one class.

```
package edu.umkc.fv

import ...

/**
 * Created by Raghu on 8/3/2016
 */
object FeatureVector1 {

  Run FeatureVector1
  16/03/03 19:44:50 INFO DAGScheduler: Missing parents: List()
  16/03/03 19:44:50 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90), which has no missing parents
  16/03/03 19:44:50 INFO MemoryStore: ensureFreeSpace(6000) called with curMem=70713995, maxMem=990621204
  16/03/03 19:44:50 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 5.9 KB, free 877.3 MB)
  16/03/03 19:44:50 INFO MemoryStore: ensureFreeSpace(3592) called with curMem=70719995, maxMem=990621204
  16/03/03 19:44:50 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.5 KB, free 877.3 MB)
  16/03/03 19:44:50 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:45088 (size: 3.5 KB, free: 941.6 MB)
  16/03/03 19:44:50 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:861
  16/03/03 19:44:50 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90)
  16/03/03 19:44:50 INFO TaskSchedulerImpl: Adding task set 5.0 with 2 tasks
  16/03/03 19:44:50 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 10, localhost, PROCESS_LOCAL, 2372 bytes)
  16/03/03 19:44:50 INFO TaskSetManager: Starting task 1.0 in stage 5.0 (TID 11, localhost, PROCESS_LOCAL, 2372 bytes)
  16/03/03 19:44:50 INFO Executor: Running task 0.0 in stage 5.0 (TID 10)
  16/03/03 19:44:50 INFO Executor: Running task 1.0 in stage 5.0 (TID 11)
  16/03/03 19:44:50 INFO BlockManager: Found block rdd_12_0 locally
  16/03/03 19:44:50 INFO BlockManager: Found block rdd_12_0 locally
  16/03/03 19:44:50 INFO BlockManager: Found block rdd_12_1 locally
  16/03/03 19:44:50 INFO BlockManager: Found block rdd_12_1 locally
  #hillary
  #hillary
  16/03/03 19:44:50 INFO Executor: Finished task 0.0 in stage 5.0 (TID 10). 2044 bytes result sent to driver
  16/03/03 19:44:50 INFO Executor: Finished task 1.0 in stage 5.0 (TID 11). 2044 bytes result sent to driver
  16/03/03 19:44:50 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 10) in 30 ms on localhost (1/2)
  16/03/03 19:44:50 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 11) in 30 ms on localhost (2/2)
  16/03/03 19:44:50 INFO DAGScheduler: ResultStage 5 (foreach at FeatureVector1.scala:44) finished in 0.032 s
  16/03/03 19:44:50 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
  16/03/03 19:44:50 INFO DAGScheduler: Job 4 finished: foreach at FeatureVector1.scala:44, took 0.050626 s
  16/03/03 19:44:50 INFO SparkContext: Invoking stop() from shutdown hook
  16/03/03 19:44:50 INFO SparkUI: Stopped Spark web UI at http://10.99.2.233:4040
  16/03/03 19:44:50 INFO DAGScheduler: Stopping DAGScheduler
  16/03/03 19:44:50 INFO ...

Terminal SBT Console Java Enterprise Run TODO Event Log
Compilation completed successfully in 5s 335ms (a minute ago) 10:35 CRLF+ UTF-8+
```