Versions:

Spark : 2.2.1

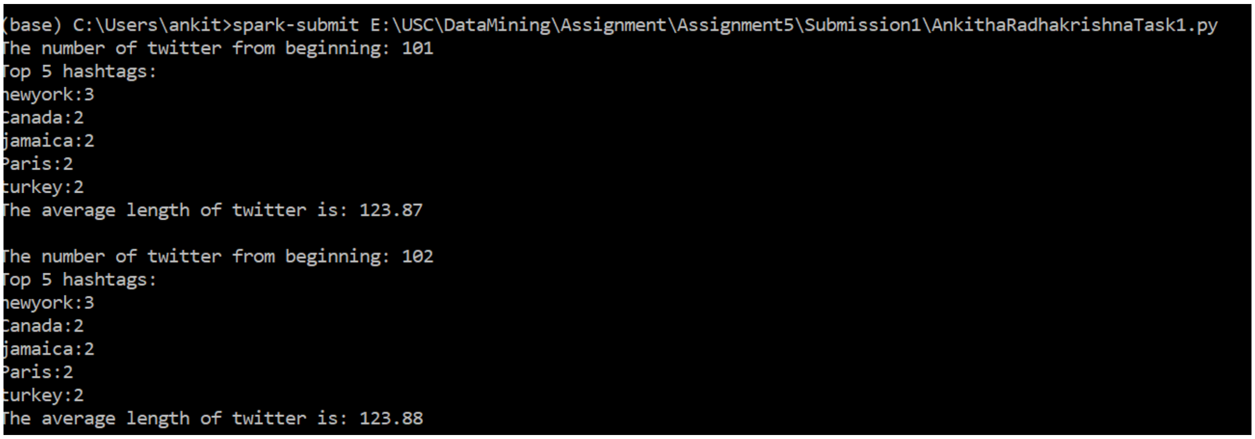
Python : 2.7

Scala : 2.11

Task 1

bin/spark-submit Ankitha_Radhakrishna_TwitterStreaming.py

Note: It takes around a minute to populate the first 100 twitter messages information in a list and then the output is seen.



```
(base) C:\Users\ankit>spark-submit E:\USC\DataMining\Assignment\Assignment5\Submission1\AnkithaRadhakrishnaTask1.py
The number of twitter from beginning: 101
Top 5 hashtags:
newyork:3
Canada:2
jamaica:2
Paris:2
turkey:2
The average length of twitter is: 123.87

The number of twitter from beginning: 102
Top 5 hashtags:
newyork:3
Canada:2
jamaica:2
Paris:2
turkey:2
The average length of twitter is: 123.88
```

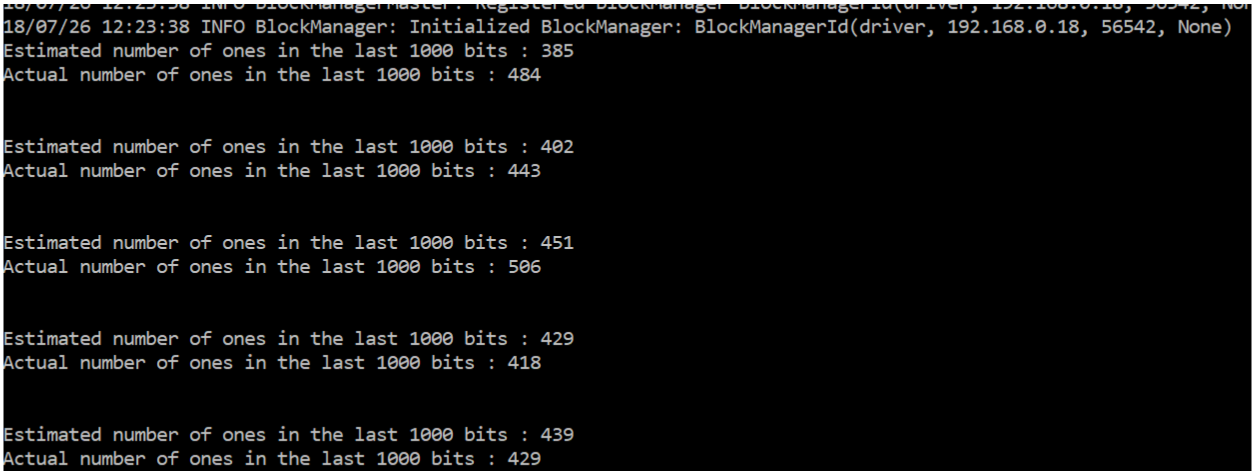
Task 2

bin/spark-submit Ankitha_Radhakrishna_DGIMAlgorithm.py

bin/spark-submit –class DGIMAlgorithm Ankitha_Radhakrishna_hw5.jar

Note: Configurations have been made to get batches every 10 seconds. Each RDD takes around a minute to get processed, so output is seen every minute.

In the beginning, if an RDD has less than 1000 records, the window is populated with the bits obtained and the next batch of bits are waited for. In such cases, the first output may take longer time to appear on the screen.



```
18/07/26 12:23:38 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.0.18, 56542, Non
18/07/26 12:23:38 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.0.18, 56542, None)
Estimated number of ones in the last 1000 bits : 385
Actual number of ones in the last 1000 bits : 484


Estimated number of ones in the last 1000 bits : 402
Actual number of ones in the last 1000 bits : 443


Estimated number of ones in the last 1000 bits : 451
Actual number of ones in the last 1000 bits : 506


Estimated number of ones in the last 1000 bits : 429
Actual number of ones in the last 1000 bits : 418


Estimated number of ones in the last 1000 bits : 439
Actual number of ones in the last 1000 bits : 429
```