Steps for running

For Task1:

$SPARK_HOME/bin/spark-submit --class TwitterStreaming Devansh_Sharma_hw5.jar

```
Devanshs-MacBook-Pro:Streaming_jar devansh$ $SPARK_HOME/bin/spark-submit --class TwitterStreaming Devansh
_Sharma_hw5.jar
```

For Task2:

$SPARK_HOME/bin/spark-submit --class BloomFiltering Devansh_Sharma_hw5.jar

```
Devanshs-MacBook-Pro:Streaming_jar devansh$ $SPARK_HOME/bin/spark-submit --class BloomFiltering Devansh_S
harma_hw5.jar
```

Versions Used

Spark version: 2.2.1

Scala version: 2.11.7

Screenshots:

Task1:

```
The number of the twitter from beginning: 125
Top 5 hot hashtags:
GoogleForWonderfulIndonesia:3
data:2
databreach:1
1:شخص_لا_ترفض_طلبه
gradstudent:1
The average length of the twitter is: 145.99

The number of the twitter from beginning: 126
Top 5 hot hashtags:
GoogleForWonderfulIndonesia:3
data:2
databreach:1
1:شخص_لا_ترفض_طلبه
gradstudent:1
The average length of the twitter is: 145.64
```

Task2:

```
---------- 10 Seconds Batch Processing Starts ----------
Current Batch Correct Count: 21
Total Correct Count: 53
Current Batch Incorrect Count: 3
Total False positive count from starting: 4

Current Batch Correct hashtags are: Trailhead  Trailblazer  TopTrailblazers  Salesforce  Ohana  Marriott
 business  cybersecurity  Mars  Data  Prepare  fight  IoT  maker  stemlc  swiftsafe  cybersecurity  vulne
rabilityassessment  databreach  uber  penetrationtest
Current Batch Incorrect Hashtags are: reInvent  customerengagement  datasecurity
False Positives from beginning are: Industry40  reInvent  customerengagement  datasecurity
---------- Current 10 Seconds Batch Processing Ends ----------


---------- 10 Seconds Batch Processing Starts ----------
Current Batch Correct Count: 5
Total Correct Count: 58
Current Batch Incorrect Count: 0
Total False positive count from starting: 4

Current Batch Correct hashtags are: JustinFungDelusion  vanre  kissofdeath  caregivers  Mars
Current Batch Incorrect Hashtags are:
False Positives from beginning are: Industry40  reInvent  customerengagement  datasecurity
---------- Current 10 Seconds Batch Processing Ends ----------
```

Method Used:

Task 1:

I have implemented the Reservoir Sampling Algorithm as given in the assignment

Task 2:

- As mentioned in the question, I have used the Bloom Filtering algorithm
- I am using 2 hash functions: string_ascii and string_fold as h1 and h2 respectively
- Bloom Filtering can have false positives, i. e., it can declare a hashtag as seen before even when it hasn't
- But it can have no false negatives, i. e., if it says it has never seen a hashtag, then it is truly new
- If for a hashtag both hash functions return 1 in the bloom array, but the tag is not in previously seen hashtags, it means it is a false positive hashtag
- Since it isn't specified, based on empirical analysis and amount of twitter streaming data I was getting, I am using Bloom filter of size 235
- I am displaying the following for each batch of the twitter stream:

  Current Batch Correct Count:, Total Correct Count:, Current Batch Incorrect Count: , Total False positive count from starting: , Current Batch Correct hashtags are: , Current Batch Incorrect Hashtags are: , False Positives from beginning are: