Bashar Alhafni **INF-553**
USC ID: 9065856238 **Fall 2018 Assignment 5**

Versions used for this assignment:

Spark: 2.2.1
Scala: 2.11.0

Twitter Streaming:

To execute the code, please run the following command:

```
spark-submit --class TwitterStreaming
~alhafni/Desktop/repos/inf553_data_mining/hw5/Bashar_Alhafni/Solution/
Bashar_Alhafni_hw5.jar
```

Expected output after establishing the connection:

```
The number of the twitter from beginning: 100
Top 5 hot hashtags:
Jazz: 1
HackedData: 1
GDPR: 1
fundourlibraries: 1
Cyberspace: 1
the average length of the twitter is: 115.48


The number of the twitter from beginning: 101
Top 5 hot hashtags:
GDPR: 0
NoClue: 0
PECR: 0
Jazz: 1
HackedData: 1
the average length of the twitter is: 112.7128712871287


The number of the twitter from beginning: 102
Top 5 hot hashtags:
GDPR: 0
NoClue: 0
ASH18: 0
PECR: 0
Jazz: 1
the average length of the twitter is: 111.6078431372549
```

Bashar Alhafni **INF-553**
USC ID: 9065856238 **Fall 2018 Assignment 5**

BloomFiltering:

To execute the code, please run the following command:

```
spark-submit --class BloomFiltering
~alhafni/Desktop/repos/inf553_data_mining/hw5/Bashar_Alhafni/Solution/
Bashar_Alhafni_hw5.jar
```

Expected output after establishing the connection:

```
total unique hash tags seen so far: 8
the number of correct estimates: 1
the number of incorrect estimates: 2
the number of false positives is: 2

total unique hash tags seen so far: 12
the number of correct estimates: 2
the number of incorrect estimates: 15
the number of false positives is: 15

total unique hash tags seen so far: 14
the number of correct estimates: 2
the number of incorrect estimates: 25
the number of false positives is: 25
```

Notes:
The numbers of correct estimates and incorrect estimates are being calculated for all the tweets. Therefore, the number of correct estimates will be equal to the number of false positives.

I am finding the number of false positives in the following manner:
Once a hashtag is received and hashed to two numbers from 0 to 31 (using two different hash functions), I am checking the values in the bloom filter where the indices will be equal to these two numbers. If the values are both 1s, that means we have seen this hashtag before according to the bloom filter. If that is the case, I go and double check if we actually have seen this hashtag through the entire streaming (by checking if it exists in a global set of hashtags). If it doesn't, that means it's a false positive.