Data Mining
Assignment 1

-Dewangee Agrawal (2016034)

Question 1 :

To form a sample of size s from N number of tweets, such that all tweets can occur in the sample with an equal probability, we can choose from either of two sampling techniques-

- Sampling with replacement -
    - For a specific number of iterations, some tweets are selected from the the N tweets at random to appear in the sample with equal probability.
    - The same tweets can be included multiple times in the sample as repetition is allowed in this case.
- Sampling without replacement -
    - For a specific number of iterations, some tweets are selected from the the N tweets at random to appear in the sample with equal probability.
    - The same tweets cannot be included multiple times in the sample as repetition isn't allowed in this case. Thus, all tweets in the sample are distinct.

Question 2:

Part a :

- According to my assumption, the contents of the tweet are integers.
- For eg, for N=100, the elements are integers from 0 to 99.
- A stream of size n=20 is created for each value of N = {100, 500, 1000, 10000} by selecting random numbers from the initial set without repetition.
- The getNextStream(int n) function creates this stream.

Part b :

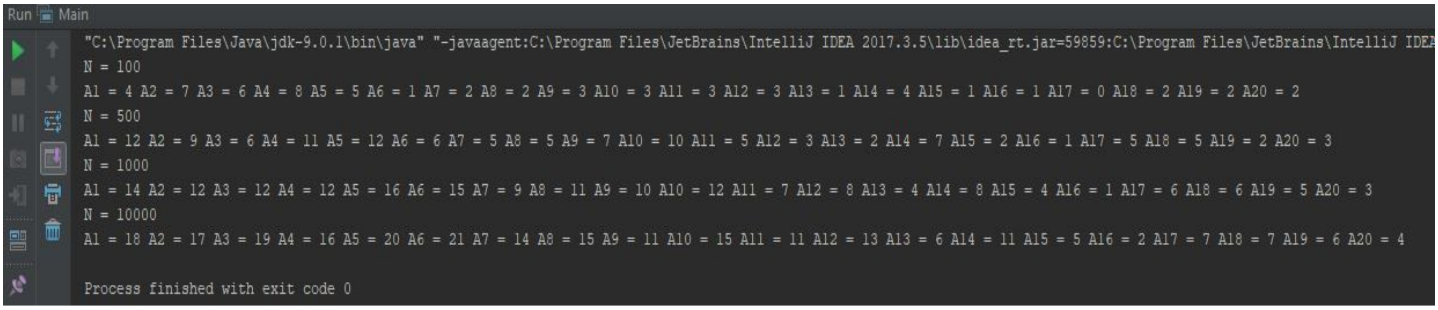
- Out of the stream of size 20, a sample of size 5 is created by running 100 iterations on the 20 numbers.
- The numbers are added to the sample according to the given conditions.
- When the size of the sample becomes 5, a previous value is removed before any new value is added.
- The updateSample (Integer streamItem, int itemNumber) creates this sample.

Part c :

- Called in the main function.

Part d :

On repeating the sampling process for 100 iterations, the following observation is obtained -



Part e :

- The sampling process described above is unbiased. This is because the streams (of size n) are chosen from the sets of tweets at random. The probability of all the items of the stream to form the sample remains the same. This can be seen by the values of the count obtained via the above code.

Part f :

Given : The length of the stream = n
The size of the sample = s (n>s)

RTP : After n stream points have arrived, the probability of any stream point being included in the sample of size s is the same and equal to s/n.

Proof : We use the principle of mathematical induction to produce the following proof.

Assumption : The probability of each stream point to get selected in a stream of size n is s/n.
To Prove : The probability of each stream point to get selected in a stream of size n+1 is s/(n+1).

When an element is added to the stream, another element is removed. Since this is done for randomly for each element with equal probability, the probability is (1/s).

Now, when an element from the sample created from the stream of size n, this might be replaced by that from the stream of size n+1. The probability for this is $(s/(n+1))*(1/s) = 1/(n+1)$.

The probability of it not being replaced is $1-(1/(n+1)) = n/(n+1)$.

So, the probability that the sample created from the stream of size n+1, which was present in the sample created from the stream of size n and not replaced is $= (s/n)*(n/(n+1)) = s/(n+1)$.

Thus, proved.
The probability of any stream point being included in the sample of size s in a stream of size n = s/n.