Programming Assignment-3, BDA

We are going to analyze the tweets related to Coronovirus in this assignment. We shall implement the AMS algorithm and find out the surprise number for a window of 30 minutes.

1. Crawl the tweets with hashtag #Covid19, #Coronavirus and similar hashtags and create a sample of 10000 user-ids with reservoir sampling algorithm.
    a. Initially a buffer of 10000 users will be created with all the initial user-ids. We shall maintain the count of of tweets for each user in the buffer.
    b. As per the reservoir sampling algorithm, replace an existing user with the new user-id and start counting the tweets for the new user
2. Implement AMS algorithm and compute the surprise number (second moment) to get an idea of tweeting behavior of users, i.e., how uneven the users are in terms of posting tweets.
3. Check Apache Spark Storm APIs and identify APIs, if any, for selecting the samples with uniform probability as computation of 2^{nd} moment. Try to implement the above task using Apache storm.
4. Make your code modular and ensure that any function definition should not have more than 50 lines of code.
5. You may find the code on Internet, please do not use that.
6. Any programming language is fine.

What to submit: A zip file containing:

a) Code files, and a power point file
b) The power point file should describe what did you learn, how much time did you take, the challenges you faced along with your results. The result is a graph of surprise number for posting behavior for different time intervals, i.e., 10 minutes, 20 minutes, 30 minutes, 40 minutes. For each interval say, 10 minutes, run your experiment 15 times and report the average of the surprise numbers.