Homework 3

CS 301 (004), Spring 2021, Introduction to Data Science

Due Date: April 21, 11:59 PM (EST)

This homework will focus on the understanding and implementation of stream algorithms we have discussed in the class.

1 Majority Item Problem. The dataset we will use is hw3a.txt, which consists of a series of integral numbers. Please treat it as a data stream. For a given value j, let $f_j \doteq |\{1 \leq t \leq n : x_t = j\}|$, *i.e.,* the frequency of value j in the stream. In the class, we have discussed in detail an algorithm that can output all values j satisfying $f_j > n/k$, where k is a given parameter (known in advance), and n is the size of the data stream (unknown to the algorithm).

1.1 Please run the algorithm as shown in the class and output all values j with $f_j > n/k$ where $k = 40$. Please attach your codes as part of your solution in addition to the final output. **(30 points)**

1.2 Please run the modified version of the algorithm after swapping the two "Ifs", which is first to check if there is some empty box and then check if there is some existing box storing the current arriving value. Compare the output here with that in 1.1. Is there any difference between the two outputs? If yes, do we miss some feasible candidate value j with $f_j > n/k$? **(10 points)**

2 Reservoir Sampling. We will use the file hw3b.txt as the data stream. For both questions 2.1 and 2.2, please attach your codes as part of your solution in addition to the final outputs.

2.1 In the class, we talk about the original definition of a random sample of a given size without replacement from a data stream. Please follow the definition and output a random sample of size 100 without replacement. **(20 points)**

2.2 Implement the reservoir sampling algorithm discussed in class to output a random sample of size 100 without replacement. **(20 points)**

2.3 Run the above two methods each for 1000 times and use the results to calculate the probability of each number being included in the final sample. Show that the results of the two methods are approximately the same. **(20 points)**