CSL7630: Assignment #2

Suman Kundu
suman@iitj.ac.in

CSE, IIT Jodhpur — March 17, 2022

Due Date — March 28, 2022

Instructions

This is the second assignment of the course. All the assignments are from the module 'Sketching and Streaming Algorithms'. There are 5 programming assignment totaling 100 marks.

Submit the code files (in original text file, e.g., for python submit .py file). Do not submit notebook file. No marks will be given if code is submitted as jupyter notebook or pdf.

Warning: Please do by yourself and remember the policy of 'no tolerance on plagiarism'. Visit the course website for penalties.

1 Dataset Generation

Use the following code written in python (or you can write similar code in your favorite programming language) to generate a large stream of integer. Size of the data will be more than 3 GB. Perform the following task.

```
with open("numbers", 'w') as f:
    no_of_element = rnd.randint(900000000, 1000000000)
    for i in range(no_of_element):
        r = rnd.randint(1000, 10000)
        f.write("{0}\n".format(rnd.randint(1, r)))
```

Question 1 *(Building Statistics of the Data (10 Marks))*

Find the following exact summary of the data.

- Total number of integers in the file.
- Total number of unique integers in the file.
- Frequency of each unique integers in the file.

2 Sketching

In our class we have learned several Sketching algorithms to approximately counting on integer streams. This section lists several programming tasks on integer lists generated in the previous section.

Approximate count of the event can be obtained by Morris and its variants. Please note that counting events is to count the number of items (integer) in the file. We don't care about the uniqueness here. For example, exact count of 1,2,3,2,1,5,3 is 7.

Question 2 *(Sketching: Approximate Count (20 Marks))*

Find the approximate count of the data using Morris+ algorithm. Report values of ϵ, δ and any other parameter you have considered. Side-by-side report the actual number of integer (from Question 1) and approximate count (output of this program). What is the value of the counter X?

Identifying the unique no of event is also important and the idealized version is the FM algorithm. Here we would like to implement non-idealized version. There are two algorithms we have discussed in our class. Anyone can be used for this. Note that here we are interested only the count of unique events viewed so far. For example, exact distinct or unique count of 1,2,3,2,1,5,3 is 4.

Question 3 *(Sketching: Approximate Distinct Count (20 Marks))*

Find the approximate distinct count using any of the practical algorithm we studied in the class. Report the parameters such as ϵ and δ. Side-by-side report the actual number of distinct element and the approximate distinct count. What is the value of the counter(s)?

In the frequency calculation, we are interested to find the number of appearance of each unique numbers in the file. For example, in case of the data 1,2,3,2,1,5,3 if we query about 3 the exact output will be 2 and if we query about 5 the output will be 1.

Question 4 *(Sketching: Frequency Query (20 Marks))*

Create a CountMin Sketch data structure for calculating the frequency of any number. Report side-by-side the actual frequency of the number and estimated frequency of number by CountMin and Count Sketch for 15 random numbers of the file.

3 Streaming

The above sketching algorithms will pass the whole data and produce the desired results as per the questions. However, here we are interested to convert the same code into the streaming version. That is the program should take user input/query anytime during the execution and answer the user query using the stream seen so far only.

Question 5 *(Building the Streaming version of the Sketches (30 Marks))*

Make all the above problem into a streaming one. That is a person can query at anytime during the execution and the system should show the estimated values at that moment.