Scalable Algorithms for Data Analysis - CS6713

Coding Assignment

Datasets: Please use `T10I4D100K` and `T40I10D100K`, and `kosarak` datasets mentioned at the link below[1]. The dataset consists of streams of integers. You may consider all the numbers mentioned in the file as one stream. Also, please consider the universe as the set of positive integers. Please use the same dataset to empirically verify on the algorithms for the following problems.

You are expected to implement everything from scratch and are not expected to use any predefined functions/libraries. Each questions consist of 5 marks.

1. Implement the Tidemark algorithm for estimating the number of distinct elements. Test it for the stream consisting of all the numbers in the file, windows of 50000 numbers each, compare it with the ground truth and plot this information.

2. Write a code to test whether there is a number that appears at least $m/10$ times in the stream, where m is the length of the stream. If so, what is the frequency of that number. That is implememnt the heavy hitters algorithm where $k = 10$.

3. Implement `Bloom filter` with the following values of the sketch size $50, 70, 100, 150, 500, 1000, 2000$. Please use the appropriate values of the hash function as per the sketch size and number of items in the stream. Consider the first 5% of elements as your test datasets (don't include the test dataset while creating bloom filter), and report the confusion matrix corresponding to each datasets, on various values of the sketch size mentioned above.

4. Implement `Count-min-sketch` algorithm with the following values of $(t, k) = \{(50, 50), (25, 100), (250, 10), (500, 5)\}$ [2]. Consider the first 5% of elements as your test datasets (consist of query items), and report the RMSE bar charts on these values of (t, k). The RMSE is defined as follows – for each query item, compute the difference of its ground truth frequency and its estimation from the sketch, square all these values, add them up, and compute the mean. Note that smaller RMSE is an indication of better performance.

[1] http://fimi.uantwerpen.be/data/

[2] Recall that k denote the sketch dimension, and t denote the repetition

5. Repeat the above for the `Count-Sketch algorithm`. In the bar-chart, put the bar-chart results of Count-sketch and Count-min-sketch side-by-side for comparison.

6. Implement `AMS-sketch` for estimating the ℓ_2 norm of the frequency vector using `medians-of-means` estimates with the following values of $(t, k) = \{(50, 50), (25, 100), (250, 10), (500, 5)$ [3]. Compute the difference of estimated quantity and the ground truth ℓ_2 norm, and report it in a bar-chart.

Note: Kindly submit a jupyter notebook file. Please copy the question in a cell, and in the following cell write its code. The code should be well commented and self explanatory. In your code, please set the path of datasets (preferably) to the desktop location.

[3] Recall that in AMS, we need to compute the mean of k observations, and then need to compute the median of t mean observations.