Total Points: 100 (5%)

Assignment 2

COMPSCI 753: Algorithms for Massive Data

Due: 3 Sep at 11:59 pm 2023

- **General Instruction:** The objective of this assignment is to explore heavy hitters' algorithms using a real-world dataset, as covered in weeks 4-5 of the course. The assignment is divided into three parts. Your task is to develop a Python program that accomplished the following components:

  1. Brute Force Approach and Performance Evaluation
  2. Misra-Gries Approach and Performance Evaluation
  3. Count Sketch Approach and Performance Evaluation

- **Datasets:** Please feel free to use the processed news article stream dataset (**news_stream.csv**), containing 209,527 news articles from 2012-01-28 to 2022-09-23. You can utilize this dataset to construct the Misra-Gries summary and the Count Sketch Summary. Each arriving news in the file is formatted as a comma-separated line with three columns:

  ```
  news_id <,> news_category <,> date
  ```

  For the original news article stream, you may refer to the URL[1].

- **Submission:** Please submit a **single report** (**.pdf or .html**) and the **source code with detailed comments** (**.py or .ipynb**) on Canvas by 23:59, Sunday 3 September 2023. It is important to include your student ID, UPI, and name in the submission file.

- **Penalty Policy:** Please note that the assignment will not be accepted after the final deadline unless special circumstances such as sickness with a medical certificate or family/personal emergencies are present. In such cases, exceptions may be made. Penalties for late submissions will be calculated as a percentage of the assignment's mark.

  1. By 23:59, Sunday 3 September 2023 (No penalty)
  2. By 23:59, Monday 4 September 2023 (25% penalty)
  3. By 23:59, Tuesday 5 September 2023 (50% penalty)

[1] https://www.kaggle.com/datasets/rmisra/news-category-dataset

1 Brute Force Approach and Performance Evaluation [10 marks]

(A) Compute the average frequency of the news categories in the news stream. **[5 marks]**

(B) Compute the true frequencies of all categories. Please report the observed category distribution using a bar chart with frequencies in descending order. (**Note**: x-axis as categories, and y-axis as true frequency.) **[5 marks]**

2 Misra-Gries Approach and Performance Evaluation [45 marks]

(A) Implement Misra-Gries summary to find the most frequent categories. Please generate a plot of the estimated frequencies in descending order to observe the approximation skewness with a summary size of $k = 20$. **[20 marks]**

(B) Compare the estimated frequency of all categories from the generated Misra-Gries summary with their true frequencies from Q1(B). In particular, please provide a bar chart for all categories, displaying (1) the estimated frequencies by the Misra-Gries approach in descending order (with $k = 20$), and (2) their corresponding true frequencies. (**Note**: x-axis as categories, y-axis as frequency, and two distributions.) **[10 marks]**

(C) Run your Misra-Gries summary and report the actual number of decrement steps calculated by your Misra-Gries with $k = 20$. **[5 marks]**

(D) Investigate the impact of the size of summary $k \in \{10, 20, 30, 40\}$ on the average relative error across all categories by Misra-Gries Approach. Please provide curve plot across varying summary size k, with k as the x-axis and average absolute error of each news category (c_i) as the y-axis.

$$\text{Absolute Error } (c_i) = \tilde{f}(c_i) - f(c_i). \tag{1}$$

where $\tilde{f}(c_i)$ is the estimated frequency and $f(c_i)$ is the true frequency of category c_i.

[5 marks]

(E) Investigate the impact of the size of summary $k \in \{10, 20, 30, 40\}$ on the run-time by Misra-Gries Approach. Please provide curve plot across varying summary size k, with k as the x-axis and run-time as the y-axis and comment how you would specify the value of k to achieve more accurate estimations with lesser run-time. **[5 marks]**

3 Count Sketch Approach and Performance Evaluation [45 marks]

(A) Implement Count Sketch Algorithm to find the most frequent categories. Please report the plot of the estimated frequencies in descending order to observe the approximation skewness with a summary size of $(w = 20, d = 4)$. **[25 marks]**

(B) Compare the estimated frequency of all categories with their true frequencies from Q1(B). In particular, please provide a curve plot for all categories, displaying (1) the estimated frequencies by the Count Sketch Algorithm in descending order with $(w = 20, d = 4)$, and (2) their corresponding true frequencies. (**Note**: x-axis as categories, y-axis as frequency, and two distributions.) **[10 marks]**

(C) Investigate the impact of the bucket size $w \in \{10, 20, 30, 40\}$ to the absolute error across all categories by the Count Sketch Algorithm. Please provide curve plot across varying bucket size w, with w as the x-axis and average absolute error of each news category (c_i) as the y-axis (Eq.1). Please comment how you would specify the value of w to achieve more accurate estimations. **[5 marks]**

(D) Investigate the impact of the number of hash functions $d \in \{2, 4, 8, 16\}$ to the absolute error across all categories by the Count Sketch Algorithm. Please provide curve plot across varying number of hash functions d, with d as the x-axis and average absolute error of each news category (c_i) as the y-axis (Eq.1). Please comment how you would specify the value of d to achieve more accurate estimations. **[5 marks]**