COMPSCI 753: Algorithms for Massive Data
Assignment 2: Data Stream Algorithms (Worth 5 Pts in Total)
Due date: 23:59 4 September 2022

Learning Objectives: The goal of this assignment is to investigate heavy hitters algorithms on a real-world dataset, which we have learnt during weeks 4-5.

General Instruction:

This core component in this assignment is to find frequent items on a data stream by investigating data stream algorithms, including **Brute Force** approach, **Misra-Gries** Summary, and **Count Sketch** algorithm. Please write a **python program** to complete the following components:

- Part I: Brute Force Approach and Performance Evaluation
- Part II: Miisra-Gries Approach and Performance Evaluation
- Part III: Count Sketch Approach and Performance Evaluation

Datasets:

Let's consider the Spotify playlists dataset, consisting of over 281,000 tracks. After loading the data (.json file), you will see that each playlist is detailed with: `<name,num_holdouts,...,tracks,num_samples>`, where `tracks` hold the details of each track/song in the playlist. You can find the data stream files, `challenge.json` on Canvas.

Submission:

Please submit a single **report** (`.pdf`) and the **source code with detailed comments**.(`.py or .ipynb or .html or .zip`) on Canvas by **23:59, Sunday 4 September 2022**. The answer file must contain your studentID, UPI and name.

Penalty Dates:

The assignment will not be accepted after the last penalty date unless there are special circumstances (e.g., sickness with medical certificate, family/personal emergencies). Penalties will be calculated as follows as a percentage of the mark for the assignment.

- By 23:59 NZST, Sunday 4 September 2022 (No penalty)
- By 23:59 NZST, Monday 5 September 2022 (25% penalty)
- By 23:59 NZST, Tuesday 6 September 2022 (50% penalty)

Part I: Brute Force Approach and Performance Evaluation [15 pts]

(a) Compute the average frequency of the tracks/songs in the data stream. **[5 pts]**

(b) Compute the frequencies of all tracks. Please report the frequencies of all tracks in descending order to see the true distributional skewness. **[10 pts]**

Part II: Miisra-Gries Approach and Performance Evaluation [40 pts]

(a) Implement Misra-Gries summary to find the most frequent tracks**.** Please report the plot of the estimated frequencies in descending order to see the approximation skewness. (Please clearly state your chosen parameters.) **[25 pts]**

(b) Compare the estimated frequency of all tracks from the generated Misra-Gries summary with their true frequencies from Part I(b). In particular, please report the plot of the relative error for all tracks with the estimated frequencies in descending order by Misra-Gries approach with the same parameters in Part II(a). (**Note**: relative error is $rel\_error_{20} = \left|1 - \frac{estimated\_freq}{true\_freq}\right|$) **[5 pts]**

(c) Run your Misra-Gries summary and report the number of decrement steps with your chosen parameter **[10 pts]**

Part III: Count Sketch Approach and Performance Evaluation [45 pts]

(a) Implement Count Sketch Algorithm to find the most frequent tracks**.** Please report the plot of the estimated frequencies in descending order to see the approximation skewness. (Please clearly state your chosen parameters.) **[30 pts]**

(b) Compare the estimated frequency of all tracks with their true frequencies from Part I(b). In particular, please report the plot of the relative error for all tracks with the estimated frequencies in descending order by Count Sketch Algorithm with the same parameters in Part III(a). **[5 pts]**

(c) Explain the impact of the size of summary (k) to the average relative error across all tracks and the runtime by Misra-Gries Approach and Count Sketch Algorithm. Please suggest how you would specify the value of k to achieve more accurate estimations and lesser estimation time, respectively? **[5 pts]**

(d) Please report the top-20 frequent tracks by Miisra-Gries, and Count Sketch Algorithm, respectively, along with (i) the track name, (ii) the estimation frequency, (iii) the true frequency, (vi) the relative error per track. **[5 pts]**