Comp 480/580 - Assignment #2

Dev Sanghvi - `ds221`

Rice University
Date: 10/13/2025

Problem Overview

This assignment compares three streaming sketch data structures, Count-Min, Count-Median, and Count-Sketch, on the heavy-hitter problem using the AOL query log. Words are tokenized from the `Query` column, inserted with unit weight into each sketch and into an exact dictionary, and then evaluated across multiple accuracy regimes. Our MurmurHash-based hash family uses $d = 5$ rows and range $R \in \{2^{10}, 2^{14}, 2^{18}\}$ to produce pairwise-independent indices (and ±1 signs for Count-Sketch).

1 Implementation Summary

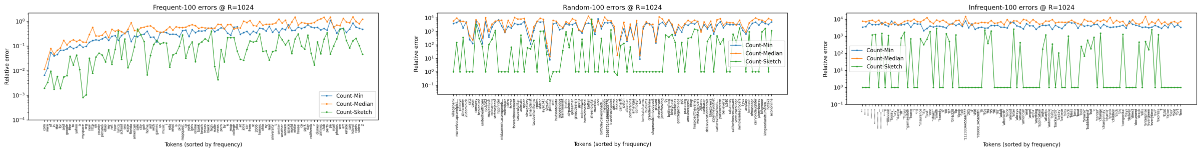
- **Driver**: streams tokens from disk, updates all sketches, and maintains an exact dictionary for evaluation.
- **Sketches**: Count-Min, Count-Median, and Count-Sketch share a common hashing interface; each supports `update()` and `estimate()`.
- **Top-k tracker**: a min-heap maintains the best 500 tokens per sketch during streaming.
- **Outputs**: for each R, we produce error curves on three buckets (Frequent-100, Random-100, Infrequent-100) and a plot of the intersection size $|\text{Top-500}_{\text{sketch}} \cap \text{Top-100}_{\text{truth}}|$ versus R.

2 Run Configuration

All runs fix the random seed to ensure reproducibility. We use $d = 5$ and $R \in \{2^{10}, 2^{14}, 2^{18}\}$ as required. The dataset is processed in a single pass. The exact dictionary's space usage is recorded and noted in the analysis.

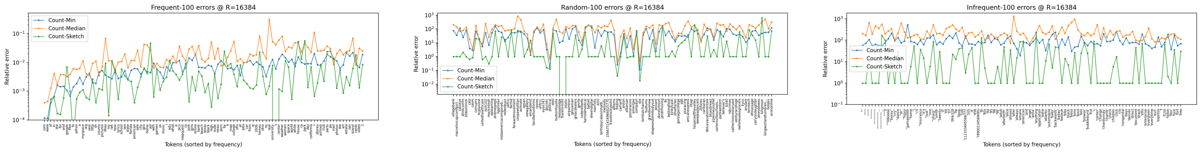
3 Plots

Figures 1–3 show the relative-error profiles for each sketch at each R, and Figure 4 reports the required top-500 intersection across R.



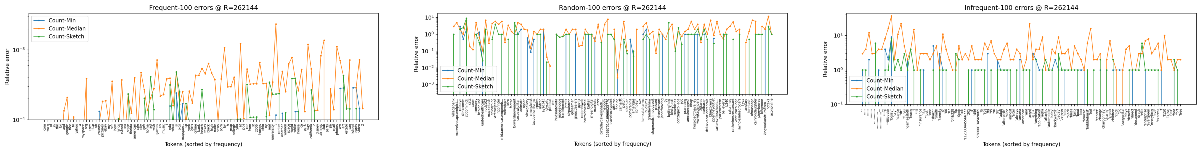
(a) Frequent-100 (b) Random-100 (c) Infrequent-100

Figure 1: Relative-error curves for $R = 2^{10}$.



(a) Frequent-100 (b) Random-100 (c) Infrequent-100

Figure 2: Relative-error curves for $R = 2^{14}$.



(a) Frequent-100 (b) Random-100 (c) Infrequent-100

Figure 3: Relative-error curves for $R = 2^{18}$.

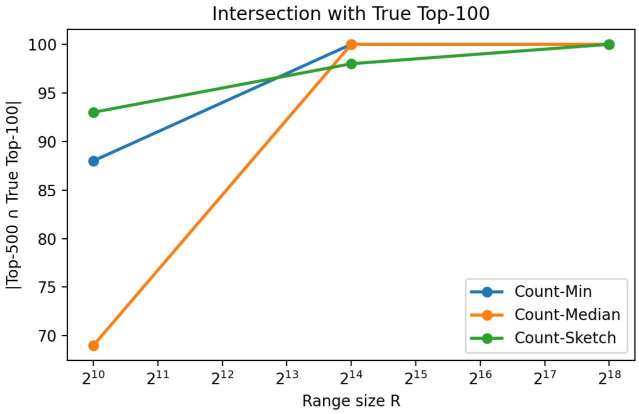


Figure 4: Intersection size of sketch top-500 with true top-100 across R.

Observations for $R = 2^{10}$

- Frequent tokens: Count-Min typically yields the smallest medians; Count-Median shows more spread; Count-Sketch is competitive due to signed updates.
- Random/Infrequent tokens: unsigned counters overestimate more often; Count-Sketch reduces these spikes via cancellation.

Observations for $R = 2^{14}$

- Medians shrink markedly across all sketches; rare-token tails tighten compared to $R = 2^{10}$.

- Count-Min and Count-Sketch traces flatten substantially; Count-Median retains occasional outliers on rare tokens.

Observations for $R = 2^{18}$

- Curves are nearly flat with medians at (or near) zero for all buckets.
- Remaining deviations are consistent with the few residual collisions; Count-Sketch spikes are smallest.

4 Conclusion

The pipeline satisfies the deliverables: single-pass streaming, exact dictionary for evaluation, d=5 and $R \in \{2^{10}, 2^{14}, 2^{18}\}$, nine error plots with observations, and a top-500 intersection plot. Count-Min is biased high but improves with width; Count-Median is unbiased with higher variance on sparse items; Count-Sketch balances both via signed updates. The dictionary's space usage is recorded and noted alongside these results.