CS 4140/6140: Data Mining HW 5

Novella Alvina - u1413401

February 2023

Part I
Streaming Algorithms

Misra-Gries and Count Min Sketch algorithm

1 Question 1.A

Run Misra-Gries Algorithm (see L11.3.1) with (k-1) = 11 counters on streams S1 and S2. Report the output of the counters at the end of the stream. In addition to each counter report the estimated ratio of each label using estimated count/m.

In each stream, use just the counters to report which characters might occur at least 25% of the time (if any), and which must occur at least 25% of the time (if any).

```
Misra-Gries Algorithm with (k-1) = 11 counters run on S1, counters:
{’a’: 772865, ’b’: 472863, ’p’: 2, ’d’: 1, ’s’: 1, ’f’: 1, ’c’: 233850,
’v’: 2, ’k’: 1, ’t’: 1, ’n’: 1}

Misra-Gries Algorithm with (k-1) = 11 counters run on S2, counters:
{’t’: 0, ’b’: 710525, ’a’: 1911225, ’h’: 0, ’j’: 0, ’s’: 0, ’c’: 311750,
’u’: 0, ’f’: 0, ’i’: 0, ’o’: 0}

The estimated ratio of each label using estimated count/m on S1:
{’a’: 0.2576216666666667, ’b’: 0.157621, ’p’: 6.666666666666667e-07,
’d’: 3.3333333333333335e-07, ’s’: 3.3333333333333335e-07,
’f’: 3.3333333333333335e-07, ’c’: 0.07795, ’v’: 6.666666666666667e-07,
’k’: 3.3333333333333335e-07, ’t’: 3.3333333333333335e-07, ’n’: 3.3333333333333335e-07}

The estimated ratio of each label using estimated count/m on S2:
{’t’: 0.0, ’b’: 0.17763125, ’a’: 0.47780625, ’h’: 0.0, ’j’: 0.0, ’s’: 0.0,
’c’: 0.0779375, ’u’: 0.0, ’f’: 0.0, ’i’: 0.0, ’o’: 0.0}
```

```
Character 25% "might" or "must" chance of occurence on S1:
{’a’: ’might’, ’b’: ’neither’, ’p’: ’neither’, ’d’: ’neither’, ’s’: ’neither’,
’f’: ’neither’, ’c’: ’neither’, ’v’: ’neither’, ’k’: ’neither’, ’t’: ’neither’,
’n’: ’neither’}

Character 25% "might" or "must" chance of occurence on S2:
{’t’: ’neither’, ’b’: ’might’, ’a’: ’must’, ’h’: ’neither’, ’j’: ’neither’,
’s’: ’neither’, ’c’: ’neither’, ’u’: ’neither’, ’f’: ’neither’, ’i’: ’neither’,
’o’: ’neither’}
```

The 25% chance of the objects to occur on the streams can be approximated by this threshold:

$$|f_q - \hat{f}_q| \leq \frac{m}{k} \tag{1}$$

where:

- m = len(Stream)
- f_q = 25% of m
- $\hat{f}_q$ = output counters from Misra-Gries Algorithm

Hence:

$$|750 \times 10^3 - \hat{f}_q| \leq \frac{3 \times 10^6}{12} \qquad \text{for S1} \tag{2}$$

$$|1 \times 10^6 - \hat{f}_q| \leq \frac{4 \times 10^6}{12} \qquad \text{for S2} \tag{3}$$

Simplified bounds:

$$-250 \times 10^3 \leq 750 \times 10^3 - \hat{f}_q \leq 250 \times 10^3 \qquad \text{for S1} \tag{4}$$

$$-250 \times 10^3 - 750 \times 10^3 \leq -\hat{f}_q \leq 250 \times 10^3 - 750 \times 10^3 \tag{5}$$

$$-1 \times 10^6 \leq -\hat{f}_q \leq -500 \times 10^3 \tag{6}$$

$$1 \times 10^6 \geq -\hat{f}_q \geq 500 \times 10^3 \tag{7}$$

$$500 \times 10^3 \leq \hat{f}_q \leq 1 \times 10^6 \tag{8}$$

$$-333333 \leq 1 \times 10^6 - \hat{f}_q \leq 333333 \qquad \text{for S2} \tag{9}$$

$$-333333 - 1 \times 10^6 \leq -\hat{f}_q \leq 333333 - 1 \times 10^6 \tag{10}$$

$$-1333333 \leq -\hat{f}_q \leq -666667 \tag{11}$$

$$1333333 \geq \hat{f}_q \geq 666667 \tag{12}$$

$$666667 \leq \hat{f}_q \leq 1333333 \tag{13}$$

Or for positive result from $f_q - \hat{f}_q$:

$$\hat{f}_q \geq 5 \times 10^5 \qquad \text{for S1} \tag{14}$$
$$\hat{f}_q \geq 666667 \qquad \text{for S2} \tag{15}$$

Thus, those character counters' that fall within the range might occur at least 25% of the time in the streams and those above the upper bound must occur at least 25% of the time in the streams and those below the lower bound neither might nor must occur 25% of the time in the streams.

Alternatively, from the lecture slides this is also considered the bound:

$$\hat{f}_q \leq f_q \leq \hat{f}_q + \frac{m}{k} \tag{16}$$

By using this bound and the same principle of categorisation applied, it can be conclude that S1 has character 'a' must occur at least 25% of the time in the streams and the rest is neither might nor must occur 25% of the time in the streams. Similar case with S2.

2 Question 1.B

Build a Count-Min Sketch (see L12.1.1) with k = 12 counters using t = 6 hash functions. Run it on streams S1 and S2.

For both streams, report the estimated counts for characters a, b, and c. In addition to each counter report the estimated ratio of each of these labels using the estimated count/m. Just from the output of the sketch, with probably 1- δ = 63/64 (that is assuming the randomness in the algorithm does not do something bad), which objects might occur at least 25% of the time (if any), and which objects must occur at least 25% of the time (if any).

```
the estimated counts for characters a, b, and c in S1:
{'a': 899566, 'b': 599564, 'c': 480289}

the estimated counts for characters a, b, and c in S2:
{'a': 2000100, 'b': 799400, 'c': 480623}

the estimated ratio of each of these labels using the estimated count/m in S1:
{'a': 0.2998553333333333, 'b': 0.19985466666666668, 'c': 0.16009633333333334}

the estimated ratio of each of these labels using the estimated count/m in S2:
{'a': 0.500025, 'b': 0.19985, 'c': 0.12015575}

Objects 25% "might" or "must" chance of occurence in S1:
{'a': 'might', 'b': 'neither', 'c': 'neither'}

Objects 25% "might" or "must" chance of occurence in S2:
{'a': 'must', 'b': 'neither', 'c': 'neither'}
```

The 25% chance of the objects to occur on the streams is approximated by the PAC bound:

$$f_q \leq \hat{f}_q \leq f_q + \varepsilon F_1 \tag{17}$$

where:

- $\varepsilon = 2/k$
- $F_1 = m = \text{len(Stream)}$
- $f_q = 25\% \text{ of m}$
- $\hat{f}_q = \text{output counters from CountMinSketch algorithm}$

Hence:

$$750 \times 10^3 \leq \hat{f}_q \leq 750 \times 10^3 + \left(\frac{2}{12} * 3 \times 10^6\right) \qquad \text{for S1} \tag{18}$$

$$1 \times 10^6 \leq \hat{f}_q \leq 1 \times 10^6 + \left(\frac{2}{12} * 4 \times 10^6\right) \qquad \text{for S2} \tag{19}$$

Simplified:

$$750 \times 10^3 \leq \hat{f}_q \leq 1,250,000 \qquad \text{for S1} \tag{20}$$

$$1 \times 10^6 \leq \hat{f}_q \leq 1,666,666 \qquad \text{for S2} \tag{21}$$

Thus, those objects counters' that fall within the range might occur at least 25% of the time in the streams and those above the upper bound must occur at least 25% of the time in the streams and those below the lower bound neither might nor must occur 25% of the time in the streams.

3 Question 1.C

How would your implementation of these algorithms need to change (to answer the same questions) if each object of the stream was a "word" seen on Twitter, and the stream contained all tweets concatenated together?

The implementation of the algorithm can be implemented similarly, however the text would have to be pre-processed differently for making the object. Instead of separating by character like what is applied here, separate the text sentences by word with vocabulary.

4 Question 1.D

Describe one advantage of the Count-Min Sketch over the Misra-Gries Algorithm.

Count-Min Sketch guarantees that the count of each element remains positive whenever it either increase or decrease by one from corpus, which can be considered as turnstile model. Whereas Misra-Gries does not guarantee this.