August Nagro, Calvin Kranig

COM S 435

TTH 0930 – 1050

PA4

Introduction

This report will detail the methodology behind my work in PA4 as well as report upon the results of various tests.

Count Min Sketch and Count Sketch

In order to test the Count Min Sketch and Count Sketch we generated a stream of size x with n unique integers in the stream. Whilst generating the stream we created a hashmap that kept track of the actual count of each integer in the stream. After creating the stream, we created an instance of Count Min Sketch or Count Sketch respectively. We then called approximateFrequency(int x) for each unique integer from 1 to n and recorded its difference from the actual value, and whether it was outside the expected error. We then reported the following:

CMS Test

```
CMSTest with 0.010000 epsilon, 0.001000 delta, 10000000 Stream Size, 1000000 Universe Size
Total Difference: 47058741997 Average Difference 47058.741997
Worst Expected Difference: 100000.000000
Bad Estimates: 0 Expected Bad Estimates: 1000.000061
Top 10 Worst Estimates
Element:233813 Actual Count:6 Difference:47543
Element:542387 Actual Count:12 Difference:47538
Element:281890 Actual Count:15 Difference:47535
Element:13106 Actual Count:3 Difference:47504
Element:800375 Actual Count:13 Difference:47494
Element:488086 Actual Count:6 Difference:47492
Element:748583 Actual Count:8 Difference:47490
Element:398081 Actual Count:7 Difference:47476
Element:131806 Actual Count:8 Difference:47475
Element:137584 Actual Count:8 Difference:47475
Actual Heavy Hitters: 0 Aproximate Heavy Hitters: 0 Bad Heavy Hitters: 0
```

Count Sketch Test

```
CountSketch Test with epsilon: 0.010000 , delta: 0.001000, Stream Size:  10000000,
Universe Size: 1000000
Total Difference: 11858699 Average Difference 11.858699
Worst Expected Difference: 104.409601
Bad Estimates: 2 Expected Bad Estimates: 1000.000061
Top 10 Worst Estimates
Element:373096 Actual Count:13 Difference:129
Element:930805 Actual Count:9 Difference:106
Element:749430 Actual Count:7 Difference:95
Element:746588 Actual Count:11 Difference:92
Element:770378 Actual Count:7 Difference:91
Element:200763 Actual Count:7 Difference:90
Element:323964 Actual Count:11 Difference:89
Element:809452 Actual Count:13 Difference:89
Element:890153 Actual Count:6 Difference:88
Element:960816 Actual Count:15 Difference:88
```

Heavy Hitters

This test was similar to the one detailed above, but this time we used a weighted data set with 2 actual heavy hitters. We ran the test as above, and then checked the return value of approximateHH () vs. the actual heavy hitters. We also reported any values returned by approximateHH that were below N*r. As you can see the two incorrectly reported heavy hitters where also the only two elements with bad estimates.

```
CMSTest with 0.010000 epsilon, 0.001000 delta, 10000000 Stream Size, 1000000 Universe
Size
Total Difference: 45179509138 Average Difference 45179.509138
Worst Expected Difference: 100000.000000
Bad Estimates: 2 Expected Bad Estimates: 1000.000061
Top 10 Worst Estimates
Element:141228 Actual Count:9 Difference:245164
Element:277730 Actual Count:13 Difference:245127
Element:992686 Actual Count:5 Difference:45646
Element:205417 Actual Count:12 Difference:45639
Element:489094 Actual Count:4 Difference:45620
Element:234375 Actual Count:8 Difference:45616
Element:749591 Actual Count:10 Difference:45614
Element:153798 Actual Count:10 Difference:45613
Element:317311 Actual Count:5 Difference:45613
Element:228597 Actual Count:11 Difference:45613
Actual Heavy Hitters: 2 Aproximate Heavy Hitters: 4 Bad Heavy Hitters: 2
```

Count Sketch Test vs. Count Min Sketch

For this test we constructed a Count Sketch Test with the same amount of memory as a Min Count Sketch, and then checked the average errors between the two.

```
CMSTest with 0.010000 epsilon, 0.001000 delta, 10000000 Stream Size, 1000000 Universe Size
Total Difference: 47044208252 Average Difference 47044.208252

CountSketch Test with epsilon: 0.010000 , delta: 0.001000, Stream Size:  10000000, Universe Size: 1000000
Total Difference: 71025598 Average Difference 71.025598
```

Count Sketch Test and Count Min Sketch Observations

Error was highest for elements that had low counts, and lowest for elements with high actual counts. Count Sketch performed better on average than Count Min Sketch.