Asmt 5: Frequent Items

A (40 points): Run the Misra-Gries Algorithm (see L11.3.1) with $(k-1) = 9$ counters on streams S1 and S2. Report the output of the counters at the end of the stream.

```
Misra-Gries Algorithm S1
Charactor: a Counter: 355715
Charactor: c Counter: 475715
Charactor: u Counter: 1
Charactor: b Counter: 625715
Charactor: m Counter: 1
Charactor: i Counter: 1
Charactor: z Counter: 1
Charactor: g Counter: 1
Charactor: r Counter: 0
```

a = 355715
b = 625715
c = 475715

```
Misra-Gries Algorithm S2
Charactor: p Counter: 1
Charactor: a Counter: 899790
Charactor: e Counter: 0
Charactor: x Counter: 1
Charactor: u Counter: 0
Charactor: c Counter: 607161
Charactor: b Counter: 406116
Charactor: o Counter: 0
Charactor: g Counter: 1
```

a = 899790
b = 406116
c = 607161

A (40 points): In each stream, use just the counters to report how many objects might occur more than 20% of the time, and which must occur more than 20% of the time.

$S1$
$20\% = 600,000$
$fq - 300,000 \leq 355,715 \leq fq$
$355,715 \leq fq \leq 655,715$
a Might Occur

$20\% = 600,000$
$fq - 300,000 \leq 625,715 \leq fq$
$625,715 \leq fq \leq 925,715$
b Must Occur

$20\% = 600,000$
$fq - 300,000 \leq 475,715 \leq fq$
$475,715 \leq fq \leq 775,715$
c Might Occur

u, m, i, z, g, and r are for sure not going to occur more than 20%

$S2$
$20\% = 800,000$
$fq - 400,000 \leq 899,790 \leq fq$
$899,790 \leq fq \leq 1,299,790$
a Must Occur

$20\% = 800,000$
$fq - 400,000 \leq 406,116 \leq fq$
$406,116 \leq fq \leq 860,116$
b Might Occur

$20\% = 800,000$
$fq - 400,000 \leq 607,161 \leq fq$
$607,161 \leq fq \leq 1,007,161$
c Might Occur

p, e, x, u, o, and g are for sure not going to occur more than 20%

B (40 points): Build a Count-Min Sketch (see L12.1.1) with k = 10 counters using t = 5 hash functions. Run it on streams S1 and S2. For both streams, report the estimated counts for objects a, b, and c. Just from the output of the sketch, which of these objects, with probably 1 - δ = 31/32 (that is assuming the randomness in the algorithm does not do something bad), might occur more than 20% of the time?

Count-Min Sketch S1

```
CountMinSketch(0)(0) = 836154
CountMinSketch(0)(1) = 47423
CountMinSketch(0)(2) = 285302
CountMinSketch(0)(3) = 237380
CountMinSketch(0)(4) = 515707
CountMinSketch(0)(5) = 142276
CountMinSketch(0)(6) = 47599
CountMinSketch(0)(7) = 874380
CountMinSketch(0)(8) = 47475
CountMinSketch(0)(9) = 438
CountMinSketch(1)(0) = 237943
CountMinSketch(1)(1) = 731892
CountMinSketch(1)(2) = 94785
CountMinSketch(1)(3) = 94841
CountMinSketch(1)(4) = 94977
CountMinSketch(1)(5) = 47456
CountMinSketch(1)(6) = 658044
CountMinSketch(1)(7) = 930927
CountMinSketch(1)(8) = 95605
CountMinSketch(1)(9) = 47664
CountMinSketch(2)(0) = 47460
CountMinSketch(2)(1) = 95295
CountMinSketch(2)(2) = 189965
CountMinSketch(2)(3) = 827190
CountMinSketch(2)(4) = 563804
CountMinSketch(2)(5) = 47512
CountMinSketch(2)(6) = 94724
CountMinSketch(2)(7) = 237133
CountMinSketch(2)(8) = 836199
CountMinSketch(2)(9) = 94852
CountMinSketch(3)(0) = 94834
CountMinSketch(3)(1) = 47376
CountMinSketch(3)(2) = 142548
CountMinSketch(3)(3) = 47301
CountMinSketch(3)(4) = 47482
CountMinSketch(3)(5) = 94972
CountMinSketch(3)(6) = 189610
CountMinSketch(3)(7) = 143043
CountMinSketch(3)(8) = 875133
CountMinSketch(3)(9) = 1351835
CountMinSketch(4)(0) = 142483
CountMinSketch(4)(1) = 142311
CountMinSketch(4)(2) = 142540
CountMinSketch(4)(3) = 94967
CountMinSketch(4)(4) = 95009
CountMinSketch(4)(5) = 95017
CountMinSketch(4)(6) = 95706
CountMinSketch(4)(7) = 610607
CountMinSketch(4)(8) = 883745
CountMinSketch(4)(9) = 731749

a =515707
b = 836154
```

c = 731749

20% = 600,000
$-84,293 \leq fq \leq 515,707$
a Will Not Occcur more than 20%

20% = 600,000
$236,154 \leq fq \leq 836,154$
b Might Occur more than 20%

20% = 600,000
$164,686.25 \leq fq \leq 731,749$
c Might Occur more than 20%

Count-Min Sketch S2
CountMinSketch(0)(0) = 686437
CountMinSketch(0)(1) = 64534
CountMinSketch(0)(2) = 385538
CountMinSketch(0)(3) = 320877
CountMinSketch(0)(4) = 1121011
CountMinSketch(0)(5) = 192826
CountMinSketch(0)(6) = 64032
CountMinSketch(0)(7) = 1145163
CountMinSketch(0)(8) = 64510
CountMinSketch(0)(9) = 584
CountMinSketch(1)(0) = 321578
CountMinSketch(1)(1) = 953840
CountMinSketch(1)(2) = 128177
CountMinSketch(1)(3) = 128707
CountMinSketch(1)(4) = 128194
CountMinSketch(1)(5) = 64399
CountMinSketch(1)(6) = 1313585
CountMinSketch(1)(7) = 814015
CountMinSketch(1)(8) = 129040
CountMinSketch(1)(9) = 63977
CountMinSketch(2)(0) = 64288
CountMinSketch(2)(1) = 128782
CountMinSketch(2)(2) = 256891
CountMinSketch(2)(3) = 1082481
CountMinSketch(2)(4) = 1185570
CountMinSketch(2)(5) = 63948
CountMinSketch(2)(6) = 128266
CountMinSketch(2)(7) = 320958
CountMinSketch(2)(8) = 685750
CountMinSketch(2)(9) = 128578
CountMinSketch(3)(0) = 128849

CountMinSketch(3)(1) = 64065
CountMinSketch(3)(2) = 192439
CountMinSketch(3)(3) = 63921
CountMinSketch(3)(4) = 64747
CountMinSketch(3)(5) = 128236
CountMinSketch(3)(6) = 256743
CountMinSketch(3)(7) = 193369
CountMinSketch(3)(8) = 1146043
CountMinSketch(3)(9) = 1807100
CountMinSketch(4)(0) = 192196
CountMinSketch(4)(1) = 192863
CountMinSketch(4)(2) = 193887
CountMinSketch(4)(3) = 128592
CountMinSketch(4)(4) = 128807
CountMinSketch(4)(5) = 128464
CountMinSketch(4)(6) = 128568
CountMinSketch(4)(7) = 1248964
CountMinSketch(4)(8) = 749779
CountMinSketch(4)(9) = 953392

a = 1121011
b = 685750
c = 953392

20% = 600,000
$521,011 \leq fq \leq 1,121,011$
a Might Occcur more than 20%

20% = 600,000
$85,750 \leq fq \leq 685,750$
b will not occur more than 20%

20% = 600,000
$353,392 \leq fq \leq 953,392$
c Might Occur more than 20%

C (10 points): 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?
So the algorithm will have to handle words and not character. You have to make decisions whether or not upper case and lower case letters are counted as the same or not. You also have to change where new words need to be broken down by, in just white spaces or other things like commas. You will also have to decide wether or not counting punctuation will count or not. Based on the

way you coded it you had to handle characters, but I turned it to Strings to do the hashing, so my algorithm will not change much, but it you did not turn the char to string then you would also have to make this change. Because I turned my char to string I won't have to do much change to my code, but others will.

D (10 points): Describe one advantage of the Count-Min Sketch over the Misra-Gries Algorithm.
Count−Min Sketch over counts for each element so you are guaranteed to account for everything, while Misra-Gries under counts so you might mis things. In order to make sure you account for everything Count−Min Sketch is the best to use.