Data Mining - Asmt3 - Frequent Items

Joyce Chen

March 14, 2018

1 Streaming Algorithms

A: Running Majority Count on Data 1 gives me: $a : 194715$ $b : 147715$ $c : 104715$ $e : 1$ $v : 1$ $j : 1$ $n : 1$ $o : 1$

nothing must occur more than 20 percent of the time a, b, c might occur more than 20 percent of the time

Running Majority Count on Data 2 gives me: $a : 231429$ $b : 121429$ $c : 161430$ $e : 1$ $g : 1$
a must occur more than 20 percent of the time b, c might occur more than 20 percent of the time

B: Running Count-min Sketch on Data 1 gives me: $a : 283890$ $b : 219588$ $c : 176790$
a, b must occur more than 20 percent of the time
c might occur more than 20 percent of the time
Running Count-min Sketch on Data 2 gives me: $a : 309583$ $b : 184742$ $c : 224755$
a, c must occur more than 20 percent of the time
b might occur more than 20 percent of the time

C: I would hash the word first into an integer value

D: It is easier to implement and the number of hash functions and buckets are independent of the input size. Most importantly it only requires one pass.