Data Mining Assignment 5

Author: Lukas Gust

Due: Feb. 27th

1 Streaming Algorithms

Description:

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

In each stream, use 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.

B: Build a Count-Min Sketch with $k = 10$ counters using $t = 5$ hash functions. Run it on stream `S1` and `S2`.

For both streams, report the estimated counts for objects `a, b, c`. Just from the output of the sketch, which of these objects, with probability $1 - \delta = 31/32$, *might* occur more than 20% of the time?

C: How would you implementation of these algorithms need to change if each object of the stream was a "word" seen on Twitter, and the stream contained all tweets concatenated together?

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

Solution:

A: After running Misra-Gries on `S1` we get the following counts:

```
Counter({'a': 330000, 'b': 600000, 'c': 450000})
```

and the output of the interpretation is:

```
'a' Maybe
'c' Maybe
'b' No.
```

For `S2` the counts are:

```
Counter({'a': 865008,
         'b': 371334,
         'c': 572379,
         'g': 1,
         'p': 1,
         'q': 1,
         'x': 1})
```

and the output of the interpretation is:

```
'a' must occur more than 20% of the time
'c' Maybe
'b' No.
'x' No.
'q' No.
'p' No.
'g' No.
```

B: After building the Count-Min Sketch for `S1` the estimated counts for `a,b,c` are:

```
a : 557077
b : 826791
c : 677106
```

and the output of the interpretation is:

```
'a' No
'b' Maybe
'c' Maybe
```

For `S2` the counts for `a,b,c` are:

```
a : 1235102
b : 868829
c : 943539
```

and the output of the interpretation:

```
'a' Maybe
'b' Maybe
'c' Maybe
```

C: If the words are separated by spaces and the sub-sequent tweets were also separated by a space then we would just treat each object as a word instead of a character. The interpretation would be the same and the algorithms would not need to change. The way we read the stream would be slightly different. We would need to make sure that we are treating each word as an object in the algorithms and we must ignore white space, punctuation, stop words, etc. Because these items would have a large amount of frequency.

D: The advantage of the Count-Min Sketch is that we can updated it at any time. Meaning that if we want to subtract a frequency from it we can. This requires additional implementation, but is not difficult.

End