Algorithmic Foundations of Data Science: Assignment #6

Khaled AlHosani (kah579), Myunggun Seo (ms9144)

New York University Abu Dhabi

March 26, 2018

Problem 1

Please refer to `Assignment #6 (ms9144, kah579) 1.a` for the code. The median and the harmonic mean seems to work better than the mean. Estimate of the distinct number of words using FM algorithm:

1. mean of the estimates = 107642.88
2. median of the estimates = 32768.0
3. harmonic mean of the estimates = 26886.56

Please refer to `Assignment #6 (ms9144, kah579) 1.b` for the code. The 9-shingles were normalized by removing punctuation and spaces; they were changed to lowercase too. Estimate of distinct number of 9-shingles using HyperLogLog algorithm:
HyperLogLog harmonic mean = 107642.88

Problem 2

Please refer to `Assignment #6 (ms9144, kah579) 2.a` for the code. Using the Misra-Gries Algorithm, the pages that were viewed by more than 10% of users are: 1, 3, 6, 7, 11, 218. The counters produced were as follows (item, count): (1, 4742), (3, 12205), (6, 16458), (7, 1428), (11, 9508), (218, 1420)

Please refer to `Assignment #6 (ms9144, kah579) 2.b` for the code. Using the Count-Min Sketch Algorithm, the pages that were viewed by more than 10% of users are: 1, 3, 6, 7, 11, 218. The counters produced were as follows (item, count): (218, 2510), (7, 2519), (1, 5833), (11, 10599), (3, 13296), (6, 17549)

Problem 3

In the version of AMS where n is known in advance, we choose k random positions in the stream and those position are used to define the starting positions of the corresponding k variables. In order to modify the AMS algorithm for a stream of an unknown length n, we need to choose the variable positions in the stream in such a way that every location in the stream has an equal chance of being the starting variable position.

As you progress through the stream you keep a counter c that is the number of elements seen so far. At each location, return a true or false (whether or not the position has been picked) with a $\frac{1}{c}$ chance. Whenever true is returned, pick one of the existing k variables with equal probability and discard it. Replace the discarded variable by one in position c, holding the element at position c and initialized with a value of 1. No discarding is needed if there are $< k$ variables so far.