CS 261 - Data Structure **Spring 2017**

4

HomeWork 2

May 5, 2017 *Liangjian Chen*

Problem 1 **Solution:**

(a) for any new incoming x_i, $p \leftarrow p + x_i, q \leftarrow q + x_i^2$

(b) average: $\frac{p}{n}$

standard deviation:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2x_i * \bar{x_i} + \bar{x}^2}$$

$$= \sqrt{\frac{1}{n}(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2)}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

By the previous soultion, we can maintain the $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i$. Thus by using the formula shown above, standard deviation can be maintained as well.

Problem 2 **Solution:**

assume $|A| = |B| = t$

$$\frac{|A \cap B|}{\sqrt{|A| * |B|}}$$
$$= \frac{|A \cap B|}{t}$$
$$= x$$
$$\rightarrow |A \cap B| = tx$$

$$\frac{|A \cap B|}{|A \cup B|}$$
$$= \frac{tx}{2t - tx}$$
$$= \frac{x}{2 - x}$$

Problem 3 **Solution:**

AAABBBBCC

Algorithm would return C which appear the least time in the sequence.

Problem 4 **Solution:**

I would choose the minHash. My algorithm would consider all left key as one set$L$, and all right key as one set$R$. Then our target is to estimate the size of the union $|L \cap R|$. Then we can applied minHash algorithm.

estimated Bound:

$J(L,R) = \dfrac{|L \cap R|}{2n - |L \cap R|}$

assume p is $J(L,R)$ and $\tilde{p}$ is the estimation of p

$|L \cap R| = \dfrac{2np}{1+p}$

which indicate the following:

$(1+\epsilon)\dfrac{2np}{1+p} > \dfrac{2n\tilde{p}}{1+\tilde{p}}$

$\tilde{p} \leq \dfrac{p(1+\epsilon)}{1-p\epsilon} \leq p(1+\dfrac{2\epsilon}{1-\epsilon})$

assume $\theta = \dfrac{2\epsilon}{1-\epsilon}$, and according to **Multiplicative Chernoff Bound**, we know that:

$Pr(\tilde{p} \geq p(1+\theta)) \leq e^{-\frac{\theta^2 p}{3}}$

which indicate that

$Pr(\tilde{p} \leq p(1+\theta)) \geq 1 - e^{-\frac{\theta^2 p}{3}} = 1 - \delta(\delta = e^{-\frac{\theta^2 p}{3}})$

Thus, we would have at least $1 - e^{-\frac{4\epsilon^2 p}{3(1-\epsilon)^2}}$ probability to get $1+\epsilon$ factor of accurate number.

Multiplicative Chernoff Bound from **Wikipedia**

Suppose X_1, ..., X_n are independent random variables taking values in $\{0,1\}$. Let X denote their sum and let $\mu = E[X]$ denote the sum's expected value. Then for any $\sigma > 0$,

$$Pr(X \geq (1+\sigma)\mu) \leq e^{-\frac{\theta^2 \mu}{3}} (0 < \sigma < 1)$$