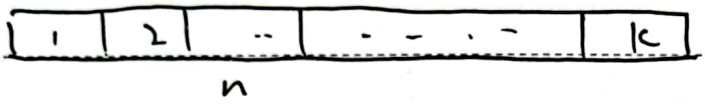
2- n 位内存容量，集合S中有 m 个成员，如果使用 k 个哈希函数，则误判率为

$$(1-e^{-\frac{km}{n}})^k$$

如果将内存分为 k 组，每组 $\frac{n}{k}$ 位，记 $t=\frac{n}{k}$



则首先假设哈希函数的选择是完全随机的。那么，任意一次哈希选中这一位的概率为 $\frac{1}{n}=\frac{1}{t}$，因此没有选中这一位的概率为 $1-\frac{1}{t}$。

插入一个元素需要哈希 k 次，所以经过 k 次哈希之后，某个特定元素未被置为1的概率为 $(1-\frac{1}{t})^k$

在 m 个元素都被插入后，该位仍为0的概率为

$$P(X_i=0)=(1-\frac{1}{t})^{km}\approx e^{-\frac{km}{t}}$$

∴ 在 m 个元素都被插入后，该位为1的概率是

$$P(X_i=1)=1-P(X_i=0)\approx 1-e^{-\frac{km}{t}}$$

又有当某个元素经过 k 个组哈希之后，对应七组中的七个位置都恰好被置为1才会发生误判，则误判率为

$$(1-e^{-\frac{km}{t}})^k=(1-e^{-\frac{k^2m}{n}})^k$$

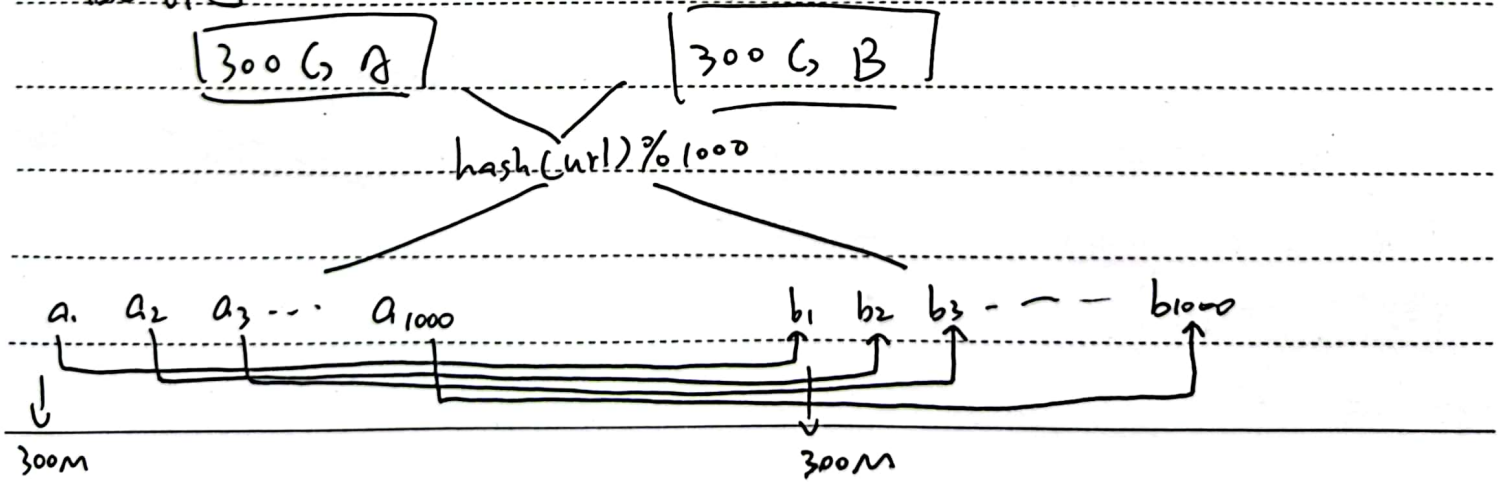
3、 50亿 x 64B ÷ 1024 ÷ 1024 ÷ 1024 = 298G ≈ 300G

而内存限制为4GB，所以无法一次读入内存，所以需要采用大文件切割的方法

假设将a，b两个大文件都分割成1000个小文件，所以每个小文件大小为300G ÷ 1000 x 1024 = 307M ≈ 300M，所以同时加载两个文件需要300M x 2 = 600M

对于a，b两个文件的集合运用相同的哈希函数，这个每个小文件集合中的小文件都有独立的哈希值

如果两个url相同，则这两个url所在的文件序号也一定相同，所以比较的时候只需要在序号相同的两个文件中进行url匹配即可



5. A: {1, 2, 3, 4} B: {2, 3, 5, 7} C: {2, 4, 6}

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Jaccard}(A, B) = \frac{2}{6} = \frac{1}{3}$$

$$\text{Jaccard}(B, C) = \frac{1}{6}$$

$$\text{Jaccard}(A, C) = \frac{2}{5}$$

7、证明

$\because P(mh(S_1)=mh(S_2))=Jaccard(S_1,S_2)$

如果 $Jaccard(S_1,S_2)=0$

则 $P(mh(S_1)=mh(S_2))=0$

即 $mh(S_1)$ 与 $mh(S_2)$ 有100%的概率不相等

即用 Min-Hashing 方法可以知道

S_1 与 S_2 两个集合必不相似

$\therefore$ 此时 Min-Hashing 一定可以给出正确的估计

6. 题目中的 X_i 是独立的伯努利随机向量序列

$$P(X_i=1)=P(h_i(S_1)=h_i(S_2))=J_S(S_1,S_2)$$

令 $X=\sum_{i=1}^{k} X_i$ 和 $\mu=\sum_{i=1}^{k} P_i$，对任意 $\epsilon\in(0,1)$，则 chernoff 不等式

$$P(X>(1+\epsilon)\mu)<\exp\left(-\frac{\mu\epsilon^2}{3}\right)$$

$$\hat{J}S(S_1,S_2)=\frac{1}{k}\sum_{i=1}^{k} X_i=\frac{X}{k}$$

$$E(\hat{J}S(S_1,S_2))=\frac{1}{k}\times k\times E(X_i)=P(h_i(S_1)=h_i(S_2))=J_S(S_1,S_2)=\frac{\mu}{k}$$

$$P(|\hat{J}S(S_1,S_2)-JS(S_1,S_2)|>\epsilon JS(S_1,S_2))$$

$$=P(\hat{J}S(S_1,S_2)>JS(S_1,S_2)+\epsilon JS(S_1,S_2))+P(\hat{J}S(S_1,S_2)<JS(S_1,S_2)-\epsilon JS(S_1,S_2))<P(\hat{J}S(S_1,S_2)>(1+\epsilon)JS(S_1,S_2))$$

$$P(X>(1+\epsilon)\mu)<\exp\left(-\frac{\mu\epsilon^2}{3}\right)$$

$$\Downarrow\quad P\left(\frac{X}{k}<\frac{(1+\epsilon)\mu}{k}\right)<\exp\left(-\frac{\mu\epsilon^2}{3}\right)$$

$$\therefore P(\hat{J}S(S_1,S_2)>(1+\epsilon)JS(S_1,S_2))<\exp\left(-\frac{kJS(S_1,S_2)\epsilon^2}{3}\right)$$

需比较其与 δ 的大小

$k=O\left(\frac{-\ln\delta}{JS\epsilon^2}\right)$ $\therefore \frac{-\ln\delta}{JS\epsilon^2}\le k$ $-\ln\delta\le JS\cdot\epsilon^2\cdot k$

$$\ln\delta\ge -JS\cdot\epsilon^2\cdot k$$

$$\delta\ge\exp(-JS\cdot\epsilon^2\cdot k)\ge\exp\left(\frac{-kJS(S_1,S_2)\epsilon^2}{3}\right)$$

$$\therefore P(\hat{J}S(S_1,S_2)>(1+\epsilon)JS(S_1,S_2))<\delta$$

$$\therefore P(|\hat{J}S(S_1,S_2)-JS(S_1,S_2)|>\epsilon JS(S_1,S_2))<\delta$$

得证

9.

(1) Jaccard $(S_1, S_2) = \frac{1}{4}$

Jaccard $(S_2, S_3) = 0$

Jaccard $(S_1, S_3) = \frac{1}{4}$

(2)

	S_1	S_2	S_3	$(7X+1) \bmod 6$ $h_1(X)$	$(11X+2) \bmod 6$ $h_2(X)$	$(5X+2) \bmod 6$ $h_3(X)$
0	1	1	0	1	2	2
1	0	1	0	2	1	1
2	1	0	0	3	0	0
3	0	0	1	4	5	5
4	1	0	1	5	4	4
5	0	0	0	0	3	3

$mh_1(S_1) = 1$ $mh_1(S_2) = 1$ $mh_1(S_3) = 4$

$mh_2(S_1) = 0$ $mh_2(S_2) = 1$ $mh_2(S_3) = 4$

$mh_3(S_1) = 0$ $mh_3(S_2) = 1$ $mh_3(S_3) = 4$

最小哈希签名矩阵

	S_1	S_2	S_3
h_1	1	1	4
h_2	0	1	4
h_3	0	1	4

11. 每个 signature 向量被分成 b 段，每段 r 行

两个 signature 向量的任意一个段所有行都相同 t^r

至少有一行不相同 $1-t^r$

所有 b 段都不同 $(1-t^r)^b$

至少有一个段相同 $1-(1-t^r)^b$

由题目可知两个集合成为 candidate 用户（即被哈希到同一个桶中的概率为 $\frac{1}{2}$）

$$\therefore \quad 1-(1-t^r)^b = \frac{1}{2}$$

$$(1-t^r)^b = \frac{1}{2}$$

$$1-t^r = 由(2^{-\frac{1}{b}})$$

$$t^r = 1-(2^{-\frac{1}{b}})$$

$$t = [1-(2^{-\frac{1}{b}})]^{\frac{1}{r}}$$

$$t = \sqrt[r]{1-\sqrt[b]{\frac{1}{2}}}$$