1. 假设一个布隆过滤器的容量为 8×10^9 位，集合中有 1×10^9 个元素。如果使用 3 个哈希函数，试计算误判率。如果使用 4 个哈希函数呢?

解　由题意，布隆过滤器容量 $m=8\times10^9$，元素数量 $n=1\times10^9$，误判率

$$p=\left(1-\left(1-\frac{1}{m}\right)^{kn}\right)^k\approx(1-e^{-kn/m})^k$$

当 $k=3$ 时，误判率为

$$p_3=0.030579$$

当 $k=4$ 时，误判率为

$$p_4=0.023969$$

2. 假设有 n 位内存容量可用，集合 S 中有 m 个成员。将 n 位内存分为 k 组，使用一个哈希函数，每次哈希到每组中对应的位置，而不是使用 k 个哈希函数，请计算误判率。这种方法和使用 k 个哈希函数有什么区别吗?

解　仅使用一个哈希函数时，对于给定的一个元素，经哈希之后，在每组中的相对位置都相同，因此当一个组中发生碰撞时，在所有的组中都将发生碰撞。此时对于某个组中的某一位，插入一个词经一个哈希函数哈希后，该位为 1 的概率为

$$\frac{1}{\frac{n}{k}}=\frac{k}{n}$$

为 0 的概率为

$$1-\frac{k}{n}$$

插入 m 个元素后，该位仍为 0 的概率为

$$\left(1-\frac{k}{n}\right)^m$$

为 1 的概率为

$$1-\left(1-\frac{k}{n}\right)^m$$

所以对于测试集中的一个元素，与布隆过滤器中元素发生碰撞的概率即误判率为

$$1-\left(1-\frac{k}{n}\right)^m\approx1-e^{-mk/n}$$

但是如果采用 k 个哈希函数，误判率为

$$\left(1-\left(1-\frac{k}{n}\right)^{m}\right)^{k} \approx (1-e^{-mk/n})^{k}$$

所以采用 k 个哈希函数时误判率要低于仅采用 1 个哈希函数。

3. 计算以下三个集合两两之间的 Jaccard 相似度：

$$\{1,2,3,4\},\{2,3,5,7\},\{2,4,6\}$$

解 令 $S_1=\{1,2,3,4\}, S_2=\{2,3,5,7\}, S_3=\{2,4,6\}$

$$sim(S_1,S_2)=\frac{2}{6}=\frac{1}{3}$$

$$sim(S_1,S_3)=\frac{2}{5}$$

$$sim(S_2,S_3)=\frac{1}{6}$$

4. 证明：如果两个集合的 Jaccard 相似度为 0，则 Min-hashing 一定可以给出一个正确的估计。

证明 假设两个文档的 Jaccard 相似度为 0，则在这两个文档的特征矩阵中不存在 $(1,1)$ 这样的组合对，因此对于同一个随机排列 π，有

$$h\pi(C_1) \neq h\pi(C_2)$$

恒成立。因此，这两个文档的 Min-hashing 对中每一对都不相等，所以 Min-hashing 给出的相似度也为 0，即 Min-hashing 一定可以给出一个正确的估计。

5. 根据图 1 中的集合表示，回答以下问题：

a. 计算任意两列之间的 Jaccard 相似度。

b. 使用以下三个哈希函数，计算每一列的 minhash 签名。

$$h_1(x)=7x+1 \bmod 6; h_2(x)=11x+2 \bmod 6; h_3(x)=5x+2 \bmod 6$$

解 a.

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

图 1: 集合表示

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

$$sim(S_2, S_3) = \frac{0}{4} = 0$$

b. 如表 1和表 2所示。

表 1: 特征矩阵

h_3	h_2	h_1	Ele	S_1	S_2	S_3
2	2	1	0	1	1	0
1	1	2	1	0	1	0
0	0	3	2	1	0	0
5	5	4	3	0	0	1
4	4	5	4	1	0	1
3	3	0	5	0	0	0

表 2: min-hashing 签名

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

6. 对于 LSH，假设两个集合被哈希到同一桶中的概率为 $\frac{1}{2}$，计算相似度阈值 t（关于 b 和 r 的函数）。

解 由 **4.4.4 基于 Min-hashing 的 LSH 过程**可得

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

解得

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

7. 设 S_1 和 S_2 为两个布尔向量，$h_1, \cdots, h_k$ 表示 k 个随机的排列，$h_i(S)$ 记录了随机排列之后,第一个行值为 1 的行号。试证明当 $k = O(\frac{\ln(1/\delta)}{JS \cdot \epsilon^2})$ 时，

$$P(|\widehat{JS}(S_1, S_2) - JS(S_1, S_2)| > \epsilon JS(S_1, S_2)) < \delta,$$

其中 $\widehat{JS}(S_1, S_2) = \frac{1}{k}\sum_{i=1}^{k} X_i$，$JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$，并且定义：

$$X_i = \begin{cases} 1, & \text{if } h_i(S_1) = h_i(S_2); \\ 0, & \text{otherwise.} \end{cases}$$

解 根据 Chernoff bound 有

$$\begin{cases} P(X \le (1-\delta)\mu) \le e^{-\delta^2 \mu/2} \\ P(X \ge (1+\delta)\mu) \le e^{-\delta^2 \mu/4} \end{cases}$$

有

$$P\left(\left|\frac{X-\mu}{\mu}\right| \ge \delta\right) \le e^{-\delta^2\mu/4}$$

根据 X_i 的定义，可以求得其期望为

$$E[X_i] = P(h_i(S_1) = h_i(S_2)) = JS(S_1, S_2)$$

令 $X^k = \sum_{i=1}^{k} X_i$, 则 $E[X^k] = k \cdot JS(S_1, S_2)$, 因为 $\widehat{JS}(S_1, S_2) = \frac{1}{k}\sum_{i=1}^{k} X_i$, 所以

$$\begin{aligned} & P(|\widehat{JS}(S_1, S_2) - JS(S_1, S_2)| > \epsilon JS(S_1, S_2)) \\ =& P\left(\left|\frac{\frac{X^k}{k} - JS(S_1, S_2)}{JS(S_1, S_2}\right| \ge \epsilon\right) \\ \le& 2e^{-\epsilon^2 \cdot k \cdot JS(S_1, S_2)/4} \end{aligned}$$

假设当 $\widehat{JS}$ 的误差在 ϵ 以内的概率不超过 $1-\delta$，即大于 ϵ 的概率至多为 δ，因此，令

$$\delta = 2e^{-\epsilon^2 \cdot k \cdot JS(S_1, S_2)/4}$$

得

$$k = \frac{4\ln 2/\delta}{\epsilon^2 \cdot JS(S_1, S_2)}$$

即

$$k = O(\frac{\ln 1/\delta}{JS \cdot \epsilon^2})$$

8. 假定全集 A 有 n 个元素，随机从中抽取出两个子集 A_1 和 A_2 ，且每个子集都有 m 个元素，求 A_1 和 A_2 两个集合的期望相似度。

解　根据集合的性质，有

$$|A_1| + |A_2| - |A_1 \cap A_2| = |A_1 \cup A_2|$$

所以，

$$Jaccard(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} = \frac{|A_1| + |A_2|}{|A_1 \cup A_2|} - 1$$

由于集合 A_1 和 A_2 都是随机抽取的子集，其模为 0 到 n 之间的一个随机整数，所以

$$E[|A_1|] = E[|A_2|] = \frac{1}{n+1}\sum_{i=0}^{n} i = \frac{n}{2}$$

$$E[|A_1| + |A_2|] = n$$

而 $|A_1 \cup A_2|$ 独立于 $|A_1|$ 和 $|A_2|$ ，且 $0 \leq |A_1 \cup A_2| \leq |A_1| + |A_2|$，所以

$$E[|A_1 \cup A_2|] = \frac{1}{n+1}\sum_{i=0}^{n} i = \frac{n}{2}$$

所以，

$$E[Jaccard(A_1, A_2)] = E[\frac{|A_1| + |A_2|}{|A_1 \cup A_2|}] - 1 = \frac{E[|A_1| + |A_2|]}{E[|A_1 \cup A_2|]} - 1 = 1$$

9. 基于例 **??** 的数据

(1) 计算通过如下两个哈希函数随机重排特征矩阵行号的最小哈希签名矩阵。

$$(4x + 1) \mod 5$$

$$(2x + 3) \mod 5$$

(2) 分析通过最小哈希签名矩阵估计出的 Jaccard 相似度与真实值的差异。

解 (1) 如表 3和表 4所示。

(2)

$$Jac(d_1, d_2) = 0.25 \qquad mh(d_1, d_2) = 0$$

$$Jac(d_1, d_3) = 0.75 \qquad mh(d_1, d_3) = 1$$

表 3: 特征矩阵

(4x+1)%5	(2x+3)%5	Elm	d_1	d_2	d_3	d_4
1	3	0	1	1	0	1
0	0	1	1	0	1	0
4	2	2	1	0	1	0
3	4	3	1	0	1	1
2	1	4	0	0	0	1

表 4: min-hashing 签名

	d_1	d_2	d_3	d_4
(4x+1)%5	0	1	0	1
(2x+3)%5	0	3	0	1

$$Jac(d_1, d_4) = 0.40 \qquad mh(d_1, d_4) = 0$$

$$Jac(d_2, d_3) = 0.00 \qquad mh(d_2, d_3) = 0$$

$$Jac(d_2, d_4) = 0.33 \qquad mh(d_2, d_4) = 0.5$$

$$Jac(d_3, d_4) = 0.20 \qquad mh(d_3, d_4) = 0$$

可以看出，Jaccard 相似度真实值较小的，经过 min-hashing 后，其估计值也比较小，反之，真实值较大的，经过 min-hashing 之后得到的真实值也比较大。

10. 对 $s = 0.1, 0.2, ..., 1$, 计算在给定如下 b 和 r 值的情况下， S -曲线的值，并简要分析你的发现。

$$r = 3, b = 20$$

$$r = 5, b = 20$$

$$r = 5, b = 30$$

解 如图 2所示。

可以发现，当 b 保持不变，增大 r 时，S-型曲线右移；当保持 r 不变，增大 b 时，S-型曲线左移。除此之外，相对于 b，r 较小的变化会引起 S-型曲线发生较大的变化。

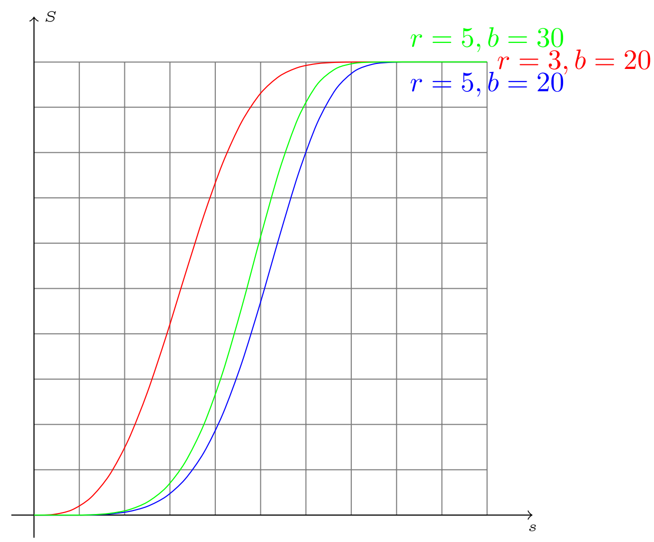

图 2: S-型曲线

11. 请设计合适的哈希函数解决例 **??** 和例 **??** 中的问题。

解 对于**例 4.2**,可以在 A,B 两张表上分别建立布隆过滤器 BF_A 和 BF_B，做自然连接时，遍历 A 中的数据，如果存在于 BF_B 中，则取出来该部分数据，对 B 做同样的操作，将两者共有数据做自然连接即可。对于**例 4.3**，可以采用局部敏感哈希 LSH ，对已经爬取过的网页构建 LSH，对于新爬取的网页，根据 LSH 判断其内容是否已经爬取过，如果已爬取过，就丢弃，否则，将其加入到 LSH 中。

12. 有 10 个大小为 1 G 的文件，每个文件的每一行存放的都是用户的查询字符串，每个文件的查询串都可能重复。请你设置方案，按照查询串的频度排序。

解

方案一：顺序读取 10 个文件，对于每个查询串，采用 $hash(\quad)\%10$ 的方法将其分配到 10 个文件中 (经过处理后对于同一个词只会出现在同一个文件中)。然后在一台内存约为 $2G$ 的机器上，对每个文件采用 $hash-map(\quad,\quad)$ 的方法统计频度，然后根据频数进行排序，这样得到了 10 个排

好序的文件，最后将这 10 个文件进行归排序得到最终的结果。

方案二：在方案一采用 $hash(\quad)\%10$ 方法分配到 10 个文件中后，可以在多台机器上对每个文件中的内容进行 $hash-map(\quad,\ )$ 排序，最后在使用归并排序进行合并。

方案三：查询串中肯呢个有很多都是重复的，因此可能对于所有不同的查询串，一次性就可以加载到内存中，这样可以直接采用 $hash-map$ 直接统计频数然后进行排序。

13. 有一个大小为 1 G 的一个文件，每一行是一个词，词的大小不超过 16 字节，内存限制大小是 1 M。请设计方案，返回频数最高的 100 个词。

解

映射：顺序读取文件，将每个词采用 $hash(word)\%5000$ 方法映射到 5000 个文件中，这样每个文件的平均大小为 $200K$，如果有哪个文件大小超过了 $1M$，可以继续采用这样的方法进行映射 (注意经过映射处理后相同的词一定在同一个文件中)；

统计：对每个小文件，采用 $hash-map$ 的方法统计词频；

排序：对每个小文件取出词频前 100 的词重写到对应的文件中并进行排序，这样得到了 5000 个每个包含 100 个词及词频的文件，然后对这些小文件进行归并排序

14. 有 1000 万字符串，其中有些是重复的，请设计恰当的方法把重复的字符串去掉，保留没有重复的字符串。

解

方案一：采用 $hash-map$ 的方法，每到一个字符串，判断其是否已经存在于 $hash-map$ 中，如果存在，则丢弃，否则将其写到输出文件中并将其添加到 $hash-map$ 中。

方案二：如果能够接受一定的错误率，可以采用布隆过滤器来实现。遍历这 1000 万条字符串，并同时构建布隆过滤器。如果一个字符串不在布隆过滤器中，则将其保存下来并添加到布隆过滤器中，否则将该条字符串丢弃。

15. 计算三个集合 $\{a,b,c,d,e\}$ 、 $\{a,b,c\}$ 、 $\{a,b,c,d,g,h\}$ 的 Jaccard 相似度。

解 令 $C_1 = \{a, b, c, d, e\}$ 、$C_2 = \{a, b, c\}$ 、$C_3 = \{a, b, c, d, g, h\}$ ，则

$$Jac(C_1, C_2) = \frac{3}{5}$$

$$Jac(C_1, C_3) = \frac{4}{7}$$

$$Jac(C_2, C_3) = \frac{3}{6} = \frac{1}{2}$$

16. 布隆过滤器能够判断某一元素是否在集合中，但是不能给出元素在集合中出现次数。请设计一个能够对集合元素进行计数的哈希算法。

解 原布隆过滤器是一个位数组，现在将位数组改为整数数组。
初始化：初始化一个空的整数数组，值为 0；
插入：计算 k 个哈希值，在相应的数组元素中加 1；
查询：计算 k 个哈希值，并取出 k 个哈希值对应的数组元素值，取其中的最小值作为对该该元素频率的估计值。
该方法会出现高估，但是不会低估。

17. 给定 a b 两个文件，各存放 50 亿个 url，每个 url 各占 64 字节，内存限制是 4G，请找出 a、b 文件共同的 url?

解 由题意知，一个文件中的元素有 $n = 5{,}000{,}000{,}000$ 个，布隆过滤器大小可以设置为 $m = 4G = 2^{35} bits$，根据 **4.4.3** 中哈希函数个数的选择和位数组大小的设置可得，哈希函数个数为

$$k = \frac{m}{n} \ln 2$$

此时最大误判率由 $m = -1.44n \log_2 \epsilon$ 可得

$$\epsilon = 0.037$$

对 a 文件构建上述的布隆过滤器，遍历 b 中的 url，判断是否存在于布隆过滤器中，存在则表示是两个文件共有的 url，否则则不是共有 url。