温兆和. 10205501432. 数据科学算法作业3

2. 解：当假设哈希函数均匀，当一个集合元素被插进滤器组中某位未置1的概率为 $1-\frac{1}{n}$（每组占位）
由于将元素哈希到各组中是分别、独立进行的，故整个n位内存中的任何一位未被置1的概率为 $1-\frac{1}{n}$.
此处与布隆过滤器不同。布隆过滤器中是k个哈希函数将元素哈希进同一集合，此时某特定位未被置1的概率为 $(1-\frac{1}{n})^k$.
将所有m个元素插入此过滤器后，整个内存中任一位未被置1概率为 $(1-\frac{1}{n})^m$.
故某一位被置1的概率为 $1-(1-\frac{1}{n})^m$.
当某元素不在集合中，误判概率为 $(1-(1-\frac{1}{n})^m)^k$.

3. 解：可将a文件中的50亿个url插入大小为 $4\times2^{30}\times8$ 个比特位的布隆过滤器中，再分别判断b中的各个url是否在布隆过滤器中.
为使误判概率更低，需要 $k=\ln2\cdot\frac{4\times2^{30}\times8}{50\times10^8}\approx4.76$ 个哈希函数，当k=5时，误判率为0.0369，比k=4时的误判率0.0379更低.
所以这个布隆过滤器需要5个哈希函数

5. 解：$J(\{1,2,3,4\},\{2,3,5,7\})=\frac{2}{6}=\frac{1}{3}$.
$J(\{2,3,5,7\},\{2,4,6\})=\frac{1}{6}$.
$J(\{1,2,3,4\},\{2,4,6\})=\frac{2}{5}$

7. 证：由于 $P(h_\pi(C_1)=h_\pi(C_2)) = Jaccard(C_1, C_2) = 0$.
故对任意随机排列 π，$h_\pi(C_1) \neq h_\pi(C_2)$ 必定成立.
签名矩阵分组后，这两个集合必定被映射到不同的桶中，我们认为不映射到同一桶中的集合不相似.
故：我们一定可以给出一个正确的估计，即认为这两个集合不相似

8. 证：诸 $X_i \sim b(1, JS)$. $E(X_i) = JS$, $Var(X_i) = JS(1-JS)$
$\hat{JS} = \frac{1}{k}\sum_{i=1}^{k} X_i$ 故 $E(\hat{JS}) = JS$. $Var(\hat{JS}) = \frac{JS(1-JS)}{k}$
由切比雪夫不等式：$P(|\hat{JS} - JS| > \varepsilon JS(S_1, S_2))$
$< \frac{Var(\hat{JS})}{\varepsilon^2 JS^2} = \frac{1-JS}{\varepsilon^2 JS \cdot k}$? $\frac{1-JS}{\varepsilon^2 k}$ 至 $\ln(1/\delta)$.
$< \frac{1}{\varepsilon^2 \ln(1/\delta)} < \delta$. ?

9. 解：(1). $J(S_1, S_2) = \frac{1}{4}$ $J(S_1, S_3) = \frac{1}{4}$. $J(S_2, S_3) = 0$.
(2). 经哈希后，行号变为：

$h_1(x)$	$h_2(x)$	$h_3(x)$
1	2.	2
2.	1	1
3	0.	0.
4.	5	5
5	4	4
0.	3	3.

故，每一列的最小哈希签名为：

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

11、结：当 b, r 较大，在相似度阈值右侧，集合有很大可能被哈希到同一个桶中；在左侧，则不太可能被哈希到同一个桶中。故，我们只要求 S 曲线最陡的地方即可。

$$f(s) = 1-(1-s^r)^b$$

$$f''(s) = -b(b-1)(1-s^r)^{b-2}(r s^{r-1})^2 + b(1-s^r)^{b-1}(r-1)s^{r-2}$$

令 $f''(s)=0$，得：$(1-s^r)(r-1) = (b-1)r s^r$

$$s^r = \frac{r-1}{br-1}$$

$$s = \left(\frac{r-1}{br-1}\right)^{\frac{1}{r}} \approx \left(\frac{r}{br}\right)^{\frac{1}{r}} = \left(\frac{1}{b}\right)^{\frac{1}{r}}$$