MCS 第12次作业

李青林*

June 12, 2012

7.21
对于$0-9$这十个数字，共有10^k中排列
$\therefore \forall N > 10^k + k$，对于集合$S_1, S_2$
$|S_1| > N, |S_2| > N \Longrightarrow S_1 \bigcup S_2 = S_1 \bigcap S_2$
$n \to \infty, resemblance \to 1$ □

7.22
令字母i接在字母j后面的概率为p_{ij}
令$p_{min} = min\{p_{ij}\}$
扩充字母表大小为$[1/p_{min}] + 1$，每个字母取的概率相等
则$resemblance$不增加
类比7.21,新模型下$resemblance = 1$
则原模型下$resemblance = 1$ □

7.23

(a) 即计算$10000 - k \approx 10000$长度为$k-1$的子串中有两个相同的概率

$$p = 1 - \frac{100^2 \times (100^2 - 1) \times \cdots \times (100^2 - 10000 + 1)}{(100^2)^{10000}} \approx 1$$

(b) $$p = 1 - \frac{100^4 \times (100^4 - 1) \times \cdots \times (100^4 - 10000 + 1)}{(100^4)^{10000}} \approx 0.39$$

□

*jack951753@gmail.com

7.24
确定一个k
将两遍文章长度为k的子串取出来组成集合A与B
计算$resemblance(A, B)$
如果没有抄袭，$resemblance$很小，反之会很大 □

7.25
确定一个k
将网页W_i长度为k的子串取出来组成集合S_i
由于网页非常多，只能将S_i中的元素哈希之后在存储
计算$resemblance(hash(S_i), hash(S_j))$
如果$resemblance$接近于1，W_i与W_j就是重复的 □

7.27
最长的重复子串为"you'll never walk along"
所以答案是23 □