

## 衡量文档相关性

概率：已知事件的观察  $\rightarrow$  未知事件的概率

## 信息检索的语言概率模型

语言产生器

语句的产生是一个随机的过程，语言产生器每次按照一定的概率输出一个单词

单词之间的相关性

$$P(w_1, w_2, w_3, w_4 | M) = P(w_1 | M)P(w_2 | w_1 M)P(w_3 | w_1 w_2 M)P(w_4 | w_1 w_2 w_3 M)$$

单词不相关

$$P(w_1, w_2, w_3, w_4 | M) = P(w_1 | M)P(w_2 | M)P(w_3 | M)P(w_4 | M)$$

相邻单词相关

$$P(w_1, w_2, w_3, w_4 | M) = P(w_1 | M)P(w_2 | w_1 M)P(w_3 | w_2 M)P(w_4 | w_3 M)$$

## 文档匹配的概率

已知：查询语句 $q$ ，一个文档 $d$ 和该文档对应的语言产生器 $M_d$

求： $P(M_d | q)$

$$P(M_d | q) = \frac{P(q | M_d)P(M_d)}{P(q)}$$

当我们进行比较时， $P(M_d)$ [假设相同]， $P(q)$ 不用求

简化为

$$P(q | M_d) = \prod_{t \in q} P_{ml}(t | M_d)$$

$$P_{ml}(t | M_d) = \frac{tf(t, d)}{dl_d}$$

以上公式有个缺陷，比如查询语句中有一个词不在文档生成器中，那结果就是0

做平滑处理,混合  $tf$  和  $df$  得到

$$P_{ml}(t|M_d) = \lambda \frac{tf(t,d)}{dl_d} + (1-\lambda) \frac{df_t}{cs}$$