Задание 9 от 19.10.20 (Вероятностные модели)

Валентин Александров, 620 группа

Постановка задачи

- Запрос к поисковой системе q: "а b"
- В коллекции имеются следующие документы (d_1, d_2, d_3, d_4) :
 - o abcd
 - o aaa
 - o bbc
 - o abbc
- Применить языковую модель к этой коллеции
- Сравнить $\lambda_1=0.5$ и $\lambda_2=0$.
- Как упорядочатся документы при этих значениях лямбда? Какая выдача кажется более правильной?

Решение

$$P(w|d) = \lambda P_{mle}(w|M_d) + (1-\lambda)P_{mle}(w|M_c)$$

$$\lambda_1 = 0.5$$

$$P(q|d_1) = (0.5\frac{1}{4} + 0.5\frac{5}{14}) \cdot (0.5\frac{1}{4} + 0.5\frac{5}{14}) = 0.092155$$
 $P(q|d_2) = (0.5 + 0.5\frac{5}{14}) \cdot 0.5\frac{5}{14} = 0.121173$
 $P(q|d_3) = 0.5\frac{5}{14} \cdot (0.5\frac{2}{3} + 0.5\frac{5}{14}) = 0.091411$
 $P(q|d_4) = (0.5\frac{1}{4} + 0.5\frac{5}{14}) \cdot (0.5\frac{2}{4} + 0.5\frac{5}{14}) = 0.130102$

$$\lambda_2=0.9$$

$$P(q|d_1) = 0.067971$$

$$P(q|d_2) = 0.033418$$

$$P(q|d_3) = 0.022704$$

$$P(q|d_4) = 0.126632$$

	$\lambda_1=0.5$	$\lambda_2=0.9$
$P(q\mid d_1)$	0.092155	0.067971
$P(q\mid d_2)$	0.121173	0.033418
$P(q\mid d_3)$	0.091411	0.022704
$P(q\mid d_4)$	0.130102	0.126632

$$\lambda_1: d_4 > d_2 > d_1 > d_3 \ \lambda_2: d_4 > d_1 > d_2 > d_3$$

Ранжирование второй лябмды выглядит разумней ("a b c d" > "a a a"). Более того, d2 и d3 имеют низкие вероятности, что тоже выглядит разумно, так как эти документы не имеют оба слова а и b одновременно.