Отчёт по домашнему заданию 4

Андрей Лебедев, группа 424

Условие

Запрос к поисковой системе состоит из двух слов: **a** и **b**. В коллекции имеются следующие документы:

- Документ 1: a b c d
- Документ 2: a a a
- Документ 3: b b c
- Документ 4: a b b c

Необходимо применить языковую модель к этой коллекции и сравнить результаты при двух значениях параметра сглаживания λ: 0.5 и 0.9. Определить, как упорядочатся документы при этих значениях λ.

Построение языковых моделей

Пусть $P(w|d)$ — вероятность слова w в документе d.

Вероятность слова w в документе без сглаживания:

$$P(w|d) = \frac{tf(w,d)}{|d|}$$

где $tf(w,d)$ — количество вхождений слова w в документе d, а $|d|$ — общее количество слов в документе d.

Документ 1

- $P(a|d1) = \frac{1}{4}$
- $P(b|d1) = \frac{1}{4}$
- $P(c|d1) = \frac{1}{4}$
- $P(d|d1) = \frac{1}{4}$

Документ 2

- $P(a|d2) = 1$
- $P(b|d2) = 0$
- $P(c|d2) = 0$
- $P(d|d2) = 0$

Документ 3

- $P(a|d3) = 0$
- $P(b|d3) = \frac{2}{3}$
- $P(c|d3) = \frac{1}{3}$
- $P(d|d3) = 0$

Документ 4

- $P(a|d4) = \frac{1}{4}$
- $P(b|d4) = \frac{2}{4} = \frac{1}{2}$
- $P(c|d4) = \frac{1}{4}$
- $P(d|d4) = 0$

Сглаживание

Сглаженная вероятность рассчитывается по формуле:

$$P_\lambda(w|d) = \lambda \cdot P(w|d) + (1 - \lambda) \cdot P(w|C)$$

где $P(w|C)$ — вероятность слова w в коллекции документов.

Частоты слов в коллекции

- $tf(a) = 5$
- $tf(b) = 5$
- $tf(c) = 3$
- $tf(d) = 1$
- $|C| = 14$

Вероятности слов в коллекции

- $P(a|C) = \frac{5}{14} \approx 0.357$
- $P(b|C) = \frac{5}{14} \approx 0.357$
- $P(c|C) = \frac{3}{14} \approx 0.214$
- $P(d|C) = \frac{1}{14} \approx 0.071$

Вычисление вероятности запроса

Вычислим вероятность запроса $Q =$ "ab" для каждого документа при значениях $\lambda = 0.5$ и $\lambda = 0.9$.

Для $\lambda = 0.5$

$$P_{\lambda=0.5}(ab|d1) = \left(0.5 \times \frac{1}{4} + 0.5 \times \frac{5}{14}\right) \times \left(0.5 \times \frac{1}{4} + 0.5 \times \frac{5}{14}\right) \approx 0.092$$

$$P_{\lambda=0.5}(ab|d2) = \left(0.5 \times 1 + 0.5 \times \frac{5}{14}\right) \times \left(0.5 \times 0 + 0.5 \times \frac{5}{14}\right) \approx 0.121$$

$$P_{\lambda=0.5}(ab|d3) = \left(0.5 \times 0 + 0.5 \times \frac{5}{14}\right) \times \left(0.5 \times \frac{2}{3} + 0.5 \times \frac{5}{14}\right) \approx 0.091$$

$$P_{\lambda=0.5}(ab|d4) = \left(0.5 \times \frac{1}{4} + 0.5 \times \frac{5}{14}\right) \times \left(0.5 \times \frac{1}{2} + 0.5 \times \frac{5}{14}\right) \approx 0.130$$

Для $\lambda = 0.9$

$$P_{\lambda=0.9}(ab|d1) = \left(0.9 \times \frac{1}{4} + 0.1 \times \frac{5}{14}\right) \times \left(0.9 \times \frac{1}{4} + 0.1 \times \frac{5}{14}\right) \approx 0.067$$

$$P_{\lambda=0.9}(ab|d2) = \left(0.9 \times 1 + 0.1 \times \frac{5}{14}\right) \times \left(0.9 \times 0 + 0.1 \times \frac{5}{14}\right) \approx 0.034$$

$$P_{\lambda=0.9}(ab|d3) = \left(0.9 \times 0 + 0.1 \times \frac{5}{14}\right) \times \left(0.9 \times \frac{2}{3} + 0.1 \times \frac{5}{14}\right) \approx 0.023$$

$$P_{\lambda=0.9}(ab|d4) = \left(0.9 \times \frac{1}{4} + 0.1 \times \frac{5}{14}\right) \times \left(0.9 \times \frac{1}{2} + 0.1 \times \frac{5}{14}\right) \approx 0.127$$

Ранжирование документов

Для $\lambda = 0.5$

1. Документ 4: 0.130
2. Документ 2: 0.121
3. Документ 1: 0.092
4. Документ 3: 0.091

Для $\lambda = 0.9$

1. Документ 4: 0.127
2. Документ 1: 0.067
3. Документ 2: 0.034
4. Документ 3: 0.023

Вывод

Документ 4 является наиболее релевантным запросу "a b"при обоих значениях λ, занимая первое место с вероятностями 0.130 и 0.127 соответственно. Это указывает на высокую степень соответствия документа запросу. При $\lambda = 0.5$ документ 2 также показывает хорошую релевантность, благодаря высокой частоте терма "a". При $\lambda = 0.9$ увеличивается влияние частот термов внутри документа, что снижает релевантность документа 2, но повышает документ 1, содержащий оба слова запроса.