МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №7
по курсу «Информационный поиск»

«Языковая модель. Задача»

Выполнил:
студент группы ИУ9-21М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2019

1 Цель работы

Необходимо оценить, как упорядочатся документы при использовании языковой модели для запроса a b, где документы представлены следующим списком:

- a b c d
- a a a
- b b c
- a b b c

При этом коэффициент $\lambda_1 = 0.5, \lambda_2 = 0.9$.

2 Ход работы

Воспользуемся формулировкой языковой модели:

$$P(Query \| doc) = \prod_{term \in Query} ((1 - \lambda) * p(term) + \lambda * p(term \| M_{doc}))$$

Здесь $p(term)$ – вероятность встречи терма в коллекции. $p(term \| M_{doc}) = \frac{TF(term,doc)}{len(doc)}$.
Веорятность встречи термов запроса: $p(a) = \frac{5}{14} = p(b)$.
Тогда формула для документов принимает следующий вид:

- $d_1 : ((1 - \lambda) * \frac{5}{14} + \lambda * \frac{1}{4}) * ((1 - \lambda) * \frac{5}{14} + \lambda * \frac{1}{4})$. Подставим lambda: $d_1(\lambda_1) = 0.09, d_1(\lambda_2) = 0.07$
- $d_2 : ((1-\lambda)*\frac{5}{14}+\lambda*\frac{3}{3})*((1-\lambda)*\frac{5}{14})$. Подставим lambda: $d_2(\lambda_1) = 0.12, d_2(\lambda_2) = 0.03$
- $d_3 : ((1-\lambda)*\frac{5}{14})*((1-\lambda)*\frac{5}{14}+\lambda*\frac{2}{3})$. Подставим lambda: $d_3(\lambda_1) = 0.09, d_3(\lambda_2) = 0.02$
- $d_4 : ((1 - \lambda) * \frac{5}{14} + \lambda * \frac{1}{4}) * ((1 - \lambda) * \frac{5}{14} + \lambda * \frac{2}{4})$. Пдоставим lambda: $d_4(\lambda_1) = 0.13, d_4(\lambda_2) = 0.127$

При коэффициенте 0.5 результат ранжирования следующий: D_4, D_2, D_1, D_3.
При коэффициенте 0.9 результат ранжирования следующий: D_4, D_1, D_2, D_3.
Такое ранжирование совпадает с ожиданием. Стоит отеметить, что разные коэффициенты lambda дают разные результаты ранжирования, что связано с бОльшим упором на наличие слова в докмуенте (документная модель), нежели в коллекции.

3 Выводы

В лабораторной работе было вычислено ранжирование документов по запросу с помощью языковой модели. В зависимости от объема коллекции и техник «компенсации» нехватки данных, результаты получаются более или менее похожими на ожидаемые.