Министерство науки и высшего образования
Российской Федерации

**Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)**

Факультет *«Информатика и системы управления»*

ОТЧЕТ
по лабораторной работе №4

по дисциплине

«Информационный поиск и извлечение информации из текстов»

Студент группы ИУ9-21М _____ *С.С. Погосян*
(подпись, дата)

Руководитель _____ *Н.В. Лукашевич*
(подпись, дата)

Москва — 2021

1. Постановка задачи

- Запрос к поисковой системе состоит из двух слов: a b
- В коллекции имеются следующие документы:
-
- a b c d
- a a a
- b b c
- a b b c
-
- Других документов в коллекции нет.
- Примените языковую модель к этой коллекции.
- Сравните лямбда=0.5 и лямбда=0.9
- Как упорядочатся документы при этих значениях лямбда? Какая выдача кажется более правильной?

2. Решение

$$P(Q,d) = \prod_{t \in Q}(1-\lambda)P_1(t) + \lambda P_2(t),$$

где P_1 – вероятность встретить слово во всей коллекции, P_2 – вероятность встретить слово в документе, λ – заданная константа сглаживания, Q – вектор запроса, d – документ.

$$Q = (a, b),$$

1) $d = 1$:
$t = a$:

$$\begin{cases} P_1 = \dfrac{3}{4}, \\ P_2 = \dfrac{1}{4}, \end{cases}$$

$t = b$:

$$\begin{cases} P_1 = \dfrac{3}{4}, \\ P_2 = \dfrac{1}{4} \end{cases}$$

$$P(Q|d_1) = \big((1-\lambda)\frac{3}{4} + \lambda\frac{1}{4}\big)\cdot\big((1-\lambda)\frac{3}{4} + \lambda\frac{1}{4}\big),$$

где $\lambda = \{0.5, 0.9\}$ Аналогично вычисляем вероятности для каждого документа (для каждого λ):
1)$\lambda = 0.5$

$$P(Q|d_1) = 0.25,$$
$$P(Q|d_2) = 0.328,$$
$$P(Q|d_3) = 0.26,$$
$$P(Q|d_4) = 0.3125$$

2)$\lambda = 0.9$

$$P(Q|d_1) = 0.09,$$
$$P(Q|d_2) = 0.073,$$
$$P(Q|d_3) = 0.0506,$$
$$P(Q|d_4) = 0.1575$$

Документы распределятся по вероятностям следующим образом (в порядке убывания вероятности):

$$(d_2, d_4, d_3, d_1), \quad (\lambda = 0.5),$$
$$(d_4, d_1, d_2, d_3), \quad (\lambda = 0.9).$$

В первом случае, самым релевантным документом выступает d_2, во втором – d_4.

Полный ход вычислений доступен по ссылке:
`https://github.com/legion15q/sem2/blob/master/num4/%D0%9E%D1%82%D1%87%D0%B5%D1%82/towlr-mfXMA.jpg`