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

Задача на дом

- Запрос: отбор кандидатов
- Пользователь отметил релевантными два документа
  - Кандидат отобрать претендент
  - Отбор выбрать претендент
- Объем коллекции – 1 млн.документов
- Df:
  - отбор 70000, кандидат – 70000,
  - Претендент - 30000, отобрать – 50000, выбрать 70000
- Как изменится запрос, если
  - alpha=0.7 (коэффициент учета запроса),
  - beta=0.3 (коэффициент учета релевантных документов),
  - Запрос представляется как вектор частот
  - Документ представляется как нормализованный вектор tf.idf

Полный ход решения доступен по ссылке: `https://github.com/legion15q/sem2/blob/master/num4_02/%D0%9E%D1%82%D1%87%D0%B5%D1%82/o7Re5PuGPNQ.jpg`

2. Решение

1) Представим документы в виде векторов, элементами которых будут веса, вычисляемые по формуле:

$$w_i = tf \cdot idf$$

где $idf = \log_{10} \frac{N}{df}$. Тогда:

$$d_1 = (0.383, 0.43, 0.506, 0, ..., 0)$$

$$d_2 = (0, 0, 0.506, 0.383, 0.383, 0, ..., 0)$$

Представим запрос в виде вектора:

$$q = 1, 0, 0, 1, 0, 0, ..., 0)$$

Нормируем вектора:

$$d_{1_{norm}} = (0.499, 0.560, 0.66, 0, ..., 0)$$

$$d_{2_{norm}} = (0, 0, 0.682, 0.516, 0.516, 0, ..., 0)$$

$$q_{norm} = (\frac{1}{\sqrt{2}}, 0, 0, \frac{1}{\sqrt{2}}, 0, ..., 0)$$

По формуле Роккио вычисляем модифицированный вектор запроса: ($\alpha = 0.7$, $\beta = 0.3$)

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

$$q_m = (0.6487, 0.168, 0.4026, 0.6548, 0.1548, 0, ..., 0)$$

Полный ход решения доступен по ссылке: `https://github.com/legion15q/sem2/blob/master/num4_02/%D0%9E%D1%82%D1%87%D0%B5%D1%82/o7Re5PuGPNQ.jpg`