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 №2

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

Term	df	idf	d1	d2	d3
car	18165	1.65	27	4	24
auto	6723	2.08	3	33	0
Insu-rance	19241	1.62	0	33	29
best	25235	1.5	14	0	17

Рис. 1. Условие задачи

Дан запрос Car insurance
Необходимо вычислить вес каждого документа

1. Представить запрос как вектор
2. Представить документ как вектор
3. Вычислить сходство запроса и документа
   - Tf – 1) число вхождений и 2) log от числа вхождений
   - Idf –
   - Вектор документа нормализуется, вектор запроса без idf, т.е. просто вектор числа вхождений
4. Показать, какие веса у документов по отношению к запросу и как упорядочатся документы

2. Реализация

Имеем пространство состояний (набор слов) V следующего вида:

$$V = (\text{car}, \text{auto}, \text{insurance}, \text{best})$$

В $|V|$ векторы документов будут иметь следующий вид:

$$d_1 = (27, 3, 0, 14),$$
$$d_2 = (4, 33, 33, 0),$$
$$d_3 = (24, 0, 29, 17),$$

где компоненты векторов это частоты термов в соответствующих документах в пространстве состояний. Вектор запроса имеет вид:

$$Q = (1, 0, 1, 0).$$

Зная idf для каждого слова в запросе можно вычислить tf-idf веса для документов по формуле

$$W_{t,d} = tf \cdot idf.$$

$$W_{t_1,d_1} = 27 \cdot 1.65 = 44.55,$$

$$W_{t_2,d_1} = 3 \cdot 2.08 = 6.24,$$

$$W_{t_3,d_1} = 0 \cdot 1.62 = 0,$$

$$W_{t_4,d_1} = 14 \cdot 1.5 = 21,$$

и т.д. В итоге получим матрицу весов документа в пространстве состояний следующего вида:

$$W = \begin{pmatrix} 44.55 & 6.6 & 39.6 \\ 6.24 & 68.64 & 0 \\ 0 & 53.46 & 46.98 \\ 21 & 0 & 25.5 \end{pmatrix}.$$

По условию задачи td-idf веса для запроса находить не нужно (и бессмысленно для данной задачи). Далее нормализуем векторы весов документов для вычисления сходства запроса и документа.

$$d^n_{i_k} = \frac{d_{i_k}}{||d_i||},$$

где $d^n_{i_k}$ нормализованные компоненты взвешенных векторов документов.

$$d_1 = (0.897369, 0.125692, 0, 0.423002),$$

$$d_2 = (0.0756426, 0.786683, 0.612705, 0),$$

$$d_3 = (0.595268, 0, 0.706204, 0.383317).$$

Нормализуем вектор запроса:

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

Поскольку векторы d_i и q нормализованы, то косинусная мера будет вычислена по формуле:

$$\cos(\vec{q}, \vec{d}) = \sum_i q_i d_i,$$

$$\cos(\vec{q}, \vec{d_1}) = 0.634536,$$

$$\cos(\vec{q}, \vec{d_2}) = 0.486735,$$

$$\cos(\vec{q}, \vec{d_3}) = 0.920280.$$

Далее используем модифицированную формулу для вычисления весов документа:

$$W_{t,d} = \log(1 + tf) \cdot idf.$$

$$W = \begin{pmatrix} 2.3878 & 1.1533 & 2.3066 \\ 1.2523 & 3.1855 & 0 \\ 0 & 2.4810 & 2.3929 \\ 1.7641 & 0 & 1.8829 \end{pmatrix}.$$

Нормализуем взвешенные векторы документов:

$$d_1 = (0.741071, 0.38866, 0, 0.547501),$$

$$d_2 = (0.274651, 0.758606, 0.590834, 0),$$

$$d_3 = (0.603837, 0, 0.626429, 0.492918).$$

Тогда косинусные меры будут иметь вид:

$$\cos(\vec{q}, \vec{d_1}) = 0.524016,$$

$$\cos(\vec{q}, \vec{d_2}) = 0.611990,$$

$$\cos(\vec{q}, \vec{d_3}) = 0.869929.$$

Видно, что в первом случае документы упорядочатся следующим образом:

$$(d_3, d_1, d_2),$$

а во втором

$$(d_3, d_2, d_1),$$

где d_3 самый релевантный документ (далее по убыванию весов).