Домашнее задание №2

Асланов А.Б., ИУ9-21М

Задание 1

Имеется запрос `car insurance`. Необходимо вычислить вес каждого документа.

- Представить запрос как вектор;
- Представить документ как вектор;
- Вычислить сходство запроса и документа (`tf` - число вхождений, `idf` - обратная подокументная частота, вектор документа нормализуется);
- Показать, какие веса у документов по отношению к запросу и как упорядочатся документы.

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

Смысл векторной модели: найти ближайшие документы к запросу.

Запрос: `car insurence`. Векторизация запроса: $q = [1, 0, 1, 0]$, потому что <слово_есть>, <слова_нет>, <слово_есть>, <слова_нет> среди термов.

TF-IDF = $TF * IDF$ = сколько раз в конкретном док-те было необходимое слово * насколько слово тематически значимое.

$w_{car,d_1} = 27 * 1,65 = 44,55$
$w_{auto,d_1} = 3 * 2,08 = 6,24$
$w_{insurance,d_1} = 0 * 1,62 = 0$
$w_{best,d_1} = 14 * 1,5 = 21$

$w_{car,d_2} = 4 * 1,65 = 6,6$
$w_{auto,d_2} = 33 * 2,08 = 68,64$
$w_{insurance,d_2} = 33 * 1,62 = 53,46$
$w_{best,d_2} = 0 * 1,5 = 0$

$w_{car,d_3} = 24 * 1,65 = 39,6$
$w_{auto,d_3} = 0 * 2,08 = 0$
$w_{insurance,d_3} = 29 * 1,62 = 46,98$
$w_{best,d_3} = 17 * 1,5 = 25,5$

На выходе получаем векторы документов, **но сравнивать между собой их пока нельзя**. Потому что в таком случае мы будем сопоставлять векторы разных длин. Для того чтобы привести все векторы к одной длине, и сравнение было верным, нужно провести нормализацию:

$$||w_{car,d_1}|| = \frac{44,55}{\sqrt{(27*1,65)^2+(3*2,08)^2+(0*1,62)^2+(14*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,634$$
$$||w_{auto,d_1}|| = \frac{6,24}{\sqrt{(27*1,65)^2+(3*2,08)^2+(0*1,62)^2+(14*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,088$$
$$||w_{insurance,d_1}|| = \frac{0}{\sqrt{(27*1,65)^2+(3*2,08)^2+(0*1,62)^2+(14*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} = 0$$
$$||w_{best,d_1}|| = \frac{21}{\sqrt{(27*1,65)^2+(3*2,08)^2+(0*1,62)^2+(14*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,299$$

$$||w_{car,d_2}|| = \frac{6,6}{\sqrt{(4*1,65)^2+(33*2,08)^2+(33*1,62)^2+(0*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,053$$
$$||w_{auto,d_2}|| = \frac{68,64}{\sqrt{(4*1,65)^2+(33*2,08)^2+(33*1,62)^2+(0*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,555$$
$$||w_{insurance,d_2}|| = \frac{53,46}{\sqrt{(4*1,65)^2+(33*2,08)^2+(33*1,62)^2+(0*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,432$$
$$||w_{best,d_2}|| = \frac{0}{\sqrt{(4*1,65)^2+(33*2,08)^2+(33*1,62)^2+(0*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} = 0$$

$$||w_{car,d_3}|| = \frac{39,6}{\sqrt{(24*1,65)^2+(0*2,08)^2+(29*1,62)^2+(17*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,42$$
$$||w_{auto,d_3}|| = \frac{0}{\sqrt{(24*1,65)^2+(0*2,08)^2+(29*1,62)^2+(17*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} = 0$$
$$||w_{insurance,d_3}|| = \frac{46,98}{\sqrt{(24*1,65)^2+(0*2,08)^2+(29*1,62)^2+(17*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,499$$
$$||w_{best,d_3}|| = \frac{25,5}{\sqrt{(24*1,65)^2+(0*2,08)^2+(29*1,62)^2+(17*1,5)^2}\sqrt{(1^2+0^2+1^2+0^2)}} \approx 0,27$$

Здесь первый множитель в знаменателе - нормировка по d, а второй - нормировка по q.

Term	d_1	d_2	d_3
car	0,634	0,053	0,42
auto	0,088	0,555	0
insurance	0	0,432	0,499
best	0,299	0	0,27

Вычислим сходство между запросом и документами.

Векторы уже нормализованы по длине, следовательно, можно вычислить косинусное расстояние, ничего не нормируя:

$d_1 = [0.634, 0.088, 0, 0.299], q = [1, 0, 1, 0]$
$Cosine(d_1, q) = 0,634 * 1 + 0,088 * 0 + 0 * 1 + 0,299 * 0 = 0,634$

$d_2 = [0.053, 0.555, 0.432, 0], q = [1, 0, 1, 0]$
$Cosine(d_2, q) = 0,053 * 1 + 0,555 * 0 + 0,432 * 1 + 0 * 0 = 0,485$

$d_3 = [0.42, 0, 0.499, 0.27], q = [1, 0, 1, 0]$
$Cosine(d_3, q) = 0,42 * 1 + 0 * 0 + 0,499 * 1 + 0,27 * 0 = 0,919$

Документы будут выданы в порядке d_3 --> d_1 --> d_2.