Отчёт по домашнему заданию 3

Андрей Лебедев, группа 424

Условие

Данные:

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

Запрос: **car insurance**.

Построение векторов и вычисление косинусовой близости

1 способ: count

Запрос (Q)

- car: $1 \times 1.65 = 1.65$
- insurance: $1 \times 1.62 = 1.62$

Вектор запроса $Q = [1.65, 0, 1.62, 0]$

Документ d1

- car: $27 \times 1.65 = 44.55$
- auto: $3 \times 2.08 = 6.24$
- insurance: $0 \times 1.62 = 0$
- best: $14 \times 1.5 = 21$

Вектор $d1 = [44.55, 6.24, 0, 21]$

Документ d2

- car: $4 \times 1.65 = 6.6$
- auto: $33 \times 2.08 = 68.64$
- insurance: $33 \times 1.62 = 53.46$
- best: $0 \times 1.5 = 0$

Вектор $d2 = [6.6, 68.64, 53.46, 0]$

Документ d3

- car: $24 \times 1.65 = 39.6$
- auto: $0 \times 2.08 = 0$
- insurance: $29 \times 1.62 = 46.98$
- best: $17 \times 1.5 = 25.5$

Вектор $d3 = [39.6, 0, 46.98, 25.5]$

Косинусовое сходство

$$\cos(Q, D) = \frac{Q \cdot D}{|Q| \cdot |D|} \tag{1}$$

Запрос (Q)

$$|Q| = \sqrt{1.65^2 + 1.62^2} = 2.31$$

Документ d1

$$|d1| = \sqrt{44.55^2 + 6.24^2 + 0 + 21^2} = 49.65$$

$$\cos(Q, d1) = \frac{73.5075}{2.31 \times 49.65} = 0.64$$

Документ d2

$$|d2| = \sqrt{6.6^2 + 68.64^2 + 53.46^2 + 0} = 87.22$$

$$\cos(Q, d2) = \frac{97.4952}{2.31 \times 87.22} = 0.48$$

Документ d3

$$|d3| = \sqrt{39.6^2 + 0 + 46.98^2 + 25.5^2} = 66.54$$

$$\cos(Q, d3) = \frac{141.4596}{2.31 \times 66.54} = 0.92$$

Ранжирование документов после обычного способа вычисления сходства:

1. **Документ d3**: **0.92**
2. **Документ d1**: **0.64**
3. **Документ d2**: **0.48**

2 способ: log(1+count)

Документ d1

- car: $\log(1 + 27) \times 1.65 = 4.03 \times 1.65 = 6.65$
- auto: $\log(1 + 3) \times 2.08 = 1.39 \times 2.08 = 2.89$
- insurance: $\log(1 + 0) \times 1.62 = 0 \times 1.62 = 0$
- best: $\log(1 + 14) \times 1.5 = 2.71 \times 1.5 = 4.07$

Вектор $d1 = [6.65, 2.89, 0, 4.07]$

Документ d2

- car: $\log(1 + 4) \times 1.65 = 1.61 \times 1.65 = 2.66$
- auto: $\log(1 + 33) \times 2.08 = 3.53 \times 2.08 = 7.35$
- insurance: $\log(1 + 33) \times 1.62 = 3.53 \times 1.62 = 5.72$
- best: $\log(1 + 0) \times 1.5 = 0 \times 1.5 = 0$

Вектор $d2 = [2.66, 7.35, 5.72, 0]$

Документ d3

- car: $\log(1+24) \times 1.65 = 3.22 \times 1.65 = 5.31$
- auto: $\log(1+0) \times 2.08 = 0 \times 2.08 = 0$
- insurance: $\log(1+29) \times 1.62 = 3.40 \times 1.62 = 5.51$
- best: $\log(1+17) \times 1.5 = 2.89 \times 1.5 = 4.34$

Вектор $d3 = [5.31, 0, 5.51, 4.34]$

Косинусовое сходство

Документ d1

$$|d1| = \sqrt{6.65^2 + 2.89^2 + 4.07^2} = 8.30$$

$$\cos(Q, d1) = \frac{1.65 \times 6.65 + 1.62 \times 0}{2.31 \times 8.30} = 0.47$$

Документ d2

$$|d2| = \sqrt{2.66^2 + 7.35^2 + 5.72^2} = 9.55$$

$$\cos(Q, d2) = \frac{1.65 \times 2.66 + 1.62 \times 5.72}{2.31 \times 9.55} = 0.38$$

Документ d3

$$|d3| = \sqrt{5.31^2 + 5.51^2 + 4.34^2} = 8.40$$

$$\cos(Q, d3) = \frac{1.65 \times 5.31 + 1.62 \times 5.51}{2.31 \times 8.40} = 0.62$$

Ранжирование документов после логарифмического способа вычисления сходства:

1. **Документ d3**: **0.62**
2. **Документ d1**: **0.47**
3. **Документ d2**: **0.38**

Вывод

В обоих вариантах вычислений документ d3 наиболее релевантен запросу «car insurance», хотя логарифмическая частота снижает абсолютные значения косинусного сходства.