Домашнее задание 8

Андрей Лебедев, группа 424

Условие задачи

Запрос: отбор кандидатов

Релевантные документы:

- Кандидат отобрать претендент
- Отбор выбрать претендент

Объём коллекции: 1 млн. документов

Частоты термов (df):

- отбор — 70000
- кандидат — 70000
- претендент — 30000
- отобрать — 50000
- выбрать — 70000

Параметры:

- $\alpha = 0.7$ — вес исходного запроса
- $\beta = 0.3$ — вес релевантных документов

Запрос представляется как вектор частот (count), а документы — как нормализованные векторы TF-IDF.

Решение

1. Вычисление TF-IDF для термов

- Для терма «отбор»:

$$\text{tf-idf}(\text{отбор}) = \log\left(\frac{1000000}{70000}\right) \approx 1.156$$

- Для терма «кандидат»:

$$\text{tf-idf}(\text{кандидат}) = \log\left(\frac{1000000}{70000}\right) \approx 1.156$$

- Для терма «претендент»:

$$\text{tf-idf}(\text{претендент}) = \log\left(\frac{1000000}{30000}\right) \approx 1.522$$

- Для терма «отобрать»:

$$\text{tf-idf}(\text{отобрать}) = \log\left(\frac{1000000}{50000}\right) \approx 1.301$$

- Для терма «выбрать»:

$$\text{tf-idf}(\text{выбрать}) = \log\left(\frac{1000000}{70000}\right) \approx 1.156$$

2. Формирование векторов

На основе частоты термов

Исходный запрос: *отбор, кандидат*

$$q_0 = [1,\ 1,\ 0,\ 0,\ 0]$$

Векторы релевантных документов:

- Документ 1: *Кандидат отобрать претендент*

$$d_{\text{rel1}} = [0,\ 1,\ 1,\ 1,\ 0]$$

- Документ 2: *Отбор выбрать претендент*

$$d_{\text{rel2}} = [1,\ 0,\ 1,\ 0,\ 1]$$

На основе TF-IDF

TF-IDF исходного запроса:

$$q_0^{\text{tf-idf}} = [1.155,\ 1.155,\ 0,\ 0,\ 0]$$

TF-IDF векторы релевантных документов:

- Документ 1:
$$d_{\text{rel1}}^{\text{tf-idf}} = [0,\ 1.155,\ 1.523,\ 1.301,\ 0]$$
- Документ 2:
$$d_{\text{rel2}}^{\text{tf-idf}} = [1.155,\ 0,\ 1.523,\ 0,\ 1.155]$$

Вычисление норм векторов

Норма вектора $d_{\text{rel1}}^{\text{tf-idf}}$:

$$\|d_{\text{rel1}}\| = \sqrt{0^2 + 1.155^2 + 1.523^2 + 1.301^2 + 0^2} \approx 2.311$$

Норма вектора $d_{\text{rel2}}^{\text{tf-idf}}$:

$$\|d_{\text{rel2}}\| = \sqrt{1.155^2 + 0^2 + 1.523^2 + 0^2 + 1.155^2} \approx 2.233$$

Нормированные векторы документов:

- $d_{\text{rel1}}^{\text{norm}} = \frac{1}{2.311} \times [0,\ 1.155,\ 1.523,\ 1.301,\ 0] \approx [0,\ 0.5,\ 0.659,\ 0.563,\ 0]$
- $d_{\text{rel2}}^{\text{norm}} = \frac{1}{2.233} \times [1.155,\ 0,\ 1.523,\ 0,\ 1.155] \approx [0.517,\ 0,\ 0.682,\ 0,\ 0.517]$

3. Применение формулы Rocchio

Формула для модификации запроса по методу Rocchio:

$$q_m = \alpha \cdot q_0^{\text{norm}} + \frac{\beta}{2} \cdot (d_{\text{rel1}}^{\text{norm}} + d_{\text{rel2}}^{\text{norm}})$$

Сумма нормированных векторов релевантных документов:

$$d_{\text{rel1}}^{\text{norm}} + d_{\text{rel2}}^{\text{norm}} = [0.517,\ 0.5,\ 1.341,\ 0.563,\ 0.517]$$

Вычислим компоненты модифицированного запроса:

$$q_m = 0.7 \times [1,\ 1,\ 0,\ 0,\ 0] + \frac{0.3}{2} \times [0.517,\ 0.5,\ 1.341,\ 0.563,\ 0.517]$$

$$q_m \approx [0.778,\ 0.775,\ 0.201,\ 0.085,\ 0.078]$$

Вывод

После применения метода Rocchio исходный запрос был дополнен новыми термами, извлечёнными из релевантных документов. Наибольшие веса остались у исходных термов «отбор» и «кандидат», что свидетельствует о том, что они по-прежнему являются ключевыми для запроса. Однако были добавлены новые термы — «претендент», «отобрать» и «выбрать» — с меньшими весами, что позволяет улучшить полноту поиска за счёт включения дополнительных релевантных результатов.

Таким образом, модифицированный запрос лучше учитывает релевантные документы и может обеспечить более точную и полную поисковую выдачу.