Ранжирование документов по текстовой релевантности

(ник на kaggle NastyBoget)

Автор: Богатенкова Анастасия

Tf-idf #1

- Из документов извлекались слова, состоящие из русских и латинских букв, цифр
- Lowercase + stemming
- Удаление стоп-слов
- Создание индекса по всем документам
- Ранжирование каждого документа в соответствии с tf-idf мерой
- Итог 0.41 score

- term-frequency

$$tf = \frac{n_t}{\sum n_k}$$

n_t - число вхождений слова t в документ

$\sum n_k$ - общее число слов в данном документе

- inverse document frequency

$$idf = log \frac{|D|}{|d_i : t \in d_i|}$$

$|D|$ - число документов в коллекции

$|d_i : t \in d_i|$ - число документов из коллекции D, в которых встречается t

Tf-idf #2

- Вместо всего текста документа извлечение только title
- Изменение запросов: исправление опечаток, раскладки (вручную)
- Расширение запросов: прибавление синонимов к набору слов в запросах, расшифровка некоторых аббревиатур, исправление транслитерации
- Использование помимо слов их символьных 3-грамм
- Итог 0.42 score

```
"соц": ["социальный"],
"вк": ["vk", "vkontakte", "вконтакте"],
"кс": ["cs", "counter", "strike"],
"дискорд": ["discord"],
"киви": ["kiwi"],
"трейнз": ["trainz"],
"мерседес": ["mercedes"],
"симс": ["sims"],
"биос": ["bios"],
"псп": ["playstation" "portable", "psp"],
"мод": ["mode"],
"одн": ["общедомовые", "нужды"],
```

Tf-idf #3

- Использование семплирования: создание индекса только для документов, соответствующих конкретному запросу
- Стоп-слова, поисковые расширения
- Использование всего содержимого документа
- Итог: score почти не изменился ~ 0.418
- Вывод: попробовать другую модель

BM25

- Идея 1: использование новой модели BM25
- Идея 2: выбирать из документов текст разных уровней (заголовок, хедеры, содержимое (body)) и считать score для каждого уровня по отдельности с разным весом (например, у заголовка вес больше)

BM-25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

$f(q_i, D)$ — частота слова (term frequency, TF) q_i в документе D

$|D|$ — длина документа (количество слов в нём)

$avgdl$ — средняя длина документа в коллекции

k_1 и b — свободные коэффициенты, обычно их выбирают как $k_1 = 2.0$ и $b = 0.75$

N — общее количество документов в коллекции

$n(q_i)$ — количество документов, содержащих q_i

BM25

- Выбор из документов текста различных уровней
- Стемминг, lowercase
- Исправленные запросы
- Модель BM25 с функцией ранжирования, учитывающей «тип» текста документа
- Итог: public score 0.74673, private score 0.72111

Спасибо на внимание!