МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №8
«Ускорение. Прыжки по индексу»
по курсу
«Информационный поиск»

Группа: 80-106М
Выполнил: Демин И.А.
Преподаватель: Калинин А.Л.

Москва, 2019

Задание

Необходимо сделать ранжированный поиск на основании схемы ранжирования TF-IDF.

Выполнение

TF-IDF - статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Данная метрика будет добавляться в поисковую систему, основанную на булевом поиске. Подсчет метрики будет выполняться на этапе построения индекса по следующим причинам:

1. Для IDF используются общие данные, такие как общее количество статей и количество документов, в которых встречается слово. Если с получением статей для конкретного слова проблем нет, то вот общее количество статей можно получить только храня его в данных.

2. TF так же легче получить в момент индексации, так как для этого нужно получить общее количество слов в документе — здесь, опять же, это возможно только в виде лишнего числа в данных.

Построение индекса и запись такая же, как и в предыдущих ЛР, кроме того, что рядом с каждым doc_id записывается idf, а в обратный индекс добавляется if для каждого слова.

В вывод дополнительно добавлен показатель TFIDF для того, что бы было видно, правильно или нет работает ранжирование.

Результаты

Запросы будут такие же, как из ЛР по булеву поиску. В ЛР4 были следующие результаты:

"спорт экспресс" — 0.4

"виды спорта" — 0.4

"активный отдых" — 0.4

"профессиональный бокс" — 0.6

"боевые искусства" — 0.8

Средняя точность — 0.52

Теперь результаты с ранжированием:

"спорт экспресс" — 0.8

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l10_tfidf$ python l10_tfidf.py
Id: 783. Заголовок: Чемпионат_СССР_по_футболу_1941. Url: https://ru.wikipedia.org/wiki/Чемпионат_СССР_по_футболу_1941. TfIdf: 0.05032547366304026
Id: 228. Заголовок: Спорт-Экспресс. Url: https://ru.wikipedia.org/wiki/Спорт-Экспресс. TfIdf: 0.03585674148956454
Id: 865. Заголовок: Назаркевич,_Юрий_Михайлович. Url: https://ru.wikipedia.org/wiki/Назаркевич,_Юрий_Михайлович. TfIdf: 0.034708404342099264
Id: 2324. Заголовок: Советский_спорт. Url: https://ru.wikipedia.org/wiki/Советский_спорт. TfIdf: 0.028278431566364642
Id: 1591. Заголовок: Дымченко,_Дмитрий_Валерьевич. Url: https://ru.wikipedia.org/wiki/Дымченко,_Дмитрий_Валерьевич. TfIdf: 0.02360170470962037
Articles count: 5
```

"виды спорта" — 1

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l10_tfidf$ python l10_tfidf.py
Id: 822. Заголовок: Национальные_виды_спорта. Url: https://ru.wikipedia.org/wiki/Национальные_виды_спорта. TfIdf: 0.14877845755558833
Id: 514. Заголовок: Всемирные_игры_2009. Url: https://ru.wikipedia.org/wiki/Всемирные_игры_2009. TfIdf: 0.05600141072617597
Id: 2534. Заголовок: Вид_спорта. Url: https://ru.wikipedia.org/wiki/Вид_спорта. TfIdf: 0.05227721457539637
Id: 2553. Заголовок: Всемирные_игры_1997. Url: https://ru.wikipedia.org/wiki/Всемирные_игры_1997. TfIdf: 0.04929489356925961
Id: 455. Заголовок: Списки_призёров_Олимпийских_игр_по_видам_спорта. Url: https://ru.wikipedia.org/wiki/Списки_призёров_Олимпийских_игр_по_видам_спорта. TfIdf: 0.04910121228014161
Articles count: 5
```

"активный отдых" — 1

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l10_tfidf$ python l10_tfidf.py
Id: 615. Заголовок: Офф-роуд. Url: https://ru.wikipedia.org/wiki/Офф-роуд. TfIdf: 0.025525057202518416
Id: 289. Заголовок: Физическая_культура. Url: https://ru.wikipedia.org/wiki/Физическая_культура. TfIdf: 0.004624004434483855
Id: 2390. Заголовок: World_Table_Hockey_Association. Url: https://ru.wikipedia.org/wiki/World_Table_Hockey_Association. TfIdf: 0.004191057825501885
Id: 2640. Заголовок: Виндсёрфинг. Url: https://ru.wikipedia.org/wiki/Виндсёрфинг. TfIdf: 0.001711292309385242
Articles count: 4
```

"профессиональный бокс" — 0.6

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l10_tfidf$ python l10_tfidf.py
Id: 1492. Заголовок: Спорт_на_Филиппинах. Url: https://ru.wikipedia.org/wiki/Спорт_на_Филиппинах. TfIdf: 0.0057968310097486295
Id: 30. Заголовок: Полусредний_вес. Url: https://ru.wikipedia.org/wiki/Полусредний_вес. TfIdf: 0.005521902988741375
Id: 2465. Заголовок: 1991_год_в_спорте. Url: https://ru.wikipedia.org/wiki/1991_год_в_спорте. TfIdf: 0.005455376463023915
Id: 1545. Заголовок: Организации_кикбоксинга. Url: https://ru.wikipedia.org/wiki/Организации_кикбоксинга. TfIdf: 0.004744541969467088
Id: 510. Заголовок: Спорт_в_Дании. Url: https://ru.wikipedia.org/wiki/Спорт_в_Дании. TfIdf: 0.0038439788485998827
Articles count: 5
```

"боевые искусства" — 1

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l10_tfidf$ python l10_tfidf.py
Id: 2086. Заголовок: Окинавские_боевые_искусства. Url: https://ru.wikipedia.org/wiki/Окинавские_боевые_искусства. TfIdf: 0.04070380920945686
Id: 2350. Заголовок: Бодзюцу. Url: https://ru.wikipedia.org/wiki/Бодзюцу. TfIdf: 0.03934163581575073
Id: 2396. Заголовок: Единоборство. Url: https://ru.wikipedia.org/wiki/Единоборство. TfIdf: 0.028156803442606515
Id: 1215. Заголовок: Тэцубодзюцу. Url: https://ru.wikipedia.org/wiki/Тэцубодзюцу. TfIdf: 0.024796630930650893
Id: 52. Заголовок: Боевые_искусства. Url: https://ru.wikipedia.org/wiki/Боевые_искусства. TfIdf: 0.02465723213930325
Articles count: 5
```

Средняя точность — 0.88

Как видно из результатов — поисковая выдача была заметно улучшена. Средняя точность показана такая же, как при использовании координатного поиска.

Вывод

В ходе лабораторной работы я познакомился метрикой TF-IDF для ранжированной выдачи и значительно улучшено качество поиска.