

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

Отчет по лабораторной работе №4
«Булев Поиск»
по курсу
«Информационный поиск»

Группа: 80-106М

Выполнил: Демин И.А.

Преподаватель: Калинин А.Л.

Москва, 2019

Задание

Нужно реализовать ввод поисковых запросов и их выполнение над индексом, получение поисковой выдачи.

Выполнение лабораторной работы

Для выполнения работы была немного изменена структура запросов, так как использовались логические операции над `set()`. По этому для поисковых запросов синтаксис следующий:

1. Пробел или один амперсанд, «&», соответствуют логической операции «И»
2. Одна вертикальных «палочки», «|» – логическая операция «ИЛИ»
3. «!» – логическая операция «НЕ»

Условие устойчивости к количеству пробелов осталось неизменным.

Примеры поисковых запросов из задания выглядят следующим образом:

1. [московский авиационный институт]
2. [(красный | желтый) автомобиль]
3. [руки !ноги]

Работа поиска происходит в консольной утилите.

Поиск работает следующим образом:

0. При запуске утилиты загружается инвертированный индекс.
1. Пользователь вводит поисковый запрос.
2. Запрос проходит предпроцессинг запроса:
 - удаляются лишние пробелы
 - происходит замена символа пробела на « & »
 - происходит замена символа «!» на « - »
 - запрос разбивается на отдельные слова
3. Для каждого слова получается hash, происходит поиск в инвертированном индексе и затем из бинарных файлов получаем статьи и формируем выдачу. (Ограничения по выдаче было решено не делать, для проверки результатов. В случае необходимости добавляется в одну строчку кода).

Описание работы получения статей. Для этого была использована встроенная функция `python eval()`, которая берет строку, переданную ей в качестве аргумента, и превращает ее в исполняемый код. В качестве такой строки передается запрос, в котором на место слова ставится набор `doc_id` в виде `set()`. После этого производятся логические операции над `set`-ами и формируется результат. Учтены случаи, когда одни слова являются частью другими слов, например «известный» и «в». Так же не производится повторный поиск по уже обработанным словам.

Проверка работы осуществлялась при помощи сравнения руками. Использовался встроенный поиск в PyCharm.

```
kompekomp-65:~/stud/search_MAI/l6_boolsearch$ python l6.py
Запрос: заслуженный мастер спорта
Заголовок: Заслуженный мастер спорта. Url: https://ru.wikipedia.org/wiki/Заслуженный\_мастер\_спорта
Заголовок: Гомельский, Владимир Александрович. Url: https://ru.wikipedia.org/wiki/Гомельский,\_Владимир\_Александрович
Заголовок: Калмыков, Виталий Николаевич. Url: https://ru.wikipedia.org/wiki/Калмыков,\_Виталий\_Николаевич
Заголовок: Степаненков, Виктор Александрович. Url: https://ru.wikipedia.org/wiki/Степаненков,\_Виктор\_Александрович
Заголовок: 1981 год в спорте. Url: https://ru.wikipedia.org/wiki/1981\_год\_в\_спорте
Заголовок: 2012 год в спорте. Url: https://ru.wikipedia.org/wiki/2012\_год\_в\_спорте
Заголовок: Белоглазов, Сергей Алексеевич. Url: https://ru.wikipedia.org/wiki/Белоглазов,\_Сергей\_Алексеевич
Заголовок: Ильенко, Наталья Никитична. Url: https://ru.wikipedia.org/wiki/Ильенко,\_Наталья\_Никитична
Заголовок: Костылева, Наталья Геннадьевна. Url: https://ru.wikipedia.org/wiki/Костылева,\_Наталья\_Геннадьевна
Заголовок: 1975 год в спорте. Url: https://ru.wikipedia.org/wiki/1975\_год\_в\_спорте
Заголовок: Челышев, Анатолий Васильевич. Url: https://ru.wikipedia.org/wiki/Челышев,\_Анатолий\_Васильевич
Заголовок: Апанасенко, Марина Геннадьевна. Url: https://ru.wikipedia.org/wiki/Апанасенко,\_Марина\_Геннадьевна
Заголовок: Фрай, Юрий Владимирович. Url: https://ru.wikipedia.org/wiki/Фрай,\_Юрий\_Владимирович
Заголовок: Хоменков, Леонид Сергеевич. Url: https://ru.wikipedia.org/wiki/Хоменков,\_Леонид\_Сергеевич
Заголовок: Кузьмин, Кирилл Константинович. Url: https://ru.wikipedia.org/wiki/Кузьмин,\_Кирилл\_Константинович
Заголовок: Голоухов, Валерий Георгиевич. Url: https://ru.wikipedia.org/wiki/Голоухов,\_Валерий\_Георгиевич
Заголовок: Билялетдинов, Динияр Ринатович. Url: https://ru.wikipedia.org/wiki/Билялетдинов,\_Динияр\_Ринатович
Заголовок: Фархутдинов, Виктор Борисович. Url: https://ru.wikipedia.org/wiki/Фархутдинов,\_Виктор\_Борисович
Заголовок: Вараев, Башир Магомедович. Url: https://ru.wikipedia.org/wiki/Вараев,\_Башир\_Магомедович
Заголовок: Аухадов, Хамзат Ахмаевич. Url: https://ru.wikipedia.org/wiki/Аухадов,\_Хамзат\_Ахмаевич
Заголовок: Руденко, Станислав Николаевич. Url: https://ru.wikipedia.org/wiki/Руденко,\_Станислав\_Николаевич
Заголовок: Михайлин, Вячеслав Вячеславович. Url: https://ru.wikipedia.org/wiki/Михайлин,\_Вячеслав\_Вячеславович
Заголовок: Вольховский, Сергей Андреевич. Url: https://ru.wikipedia.org/wiki/Вольховский,\_Сергей\_Андреевич
Заголовок: Пасхаев, Умар Алуевич. Url: https://ru.wikipedia.org/wiki/Пасхаев,\_Умар\_Алуевич
Заголовок: Список заслуженных мастеров спорта России по хоккею с мячом. Url: https://ru.wikipedia.org/wiki/Список\_заслуженных\_мастеров\_спорта\_России\_по\_хоккею\_с\_мячом
Заголовок: Шкурлатова, Галина Александровна. Url: https://ru.wikipedia.org/wiki/Шкурлатова,\_Галина\_Александровна
Заголовок: Кистяковский, Андрей Юльевич. Url: https://ru.wikipedia.org/wiki/Кистяковский,\_Андрей\_Юльевич
Заголовок: Виноградов, Эдуард Юрьевич. Url: https://ru.wikipedia.org/wiki/Виноградов,\_Эдуард\_Юрьевич
Заголовок: Белоглазов, Анатолий Алексеевич. Url: https://ru.wikipedia.org/wiki/Белоглазов,\_Анатолий\_Алексеевич
Заголовок: Парамонов, Алексей Александрович. Url: https://ru.wikipedia.org/wiki/Парамонов,\_Алексей\_Александрович
Заголовок: Михайлов, Владимир Николаевич. Url: https://ru.wikipedia.org/wiki/Михайлов,\_Владимир\_Николаевич
Заголовок: 2016 год в спорте. Url: https://ru.wikipedia.org/wiki/2016\_год\_в\_спорте
Заголовок: Стивенсон, Теофило. Url: https://ru.wikipedia.org/wiki/Стивенсон,\_Теофило
Заголовок: Цыплаков, Александр Викторович. Url: https://ru.wikipedia.org/wiki/Цыплаков,\_Александр\_Викторович
Заголовок: Митянин, Юрий Павлович. Url: https://ru.wikipedia.org/wiki/Митянин,\_Юрий\_Павлович
Заголовок: Катинасов, Сайгид Абдусаламович. Url: https://ru.wikipedia.org/wiki/Катинасов,\_Сайгид\_Абдусаламович
Заголовок: Никифоров, Григорий Исаевич. Url: https://ru.wikipedia.org/wiki/Никифоров,\_Григорий\_Исаевич
Заголовок: Жох, Олег Сергеевич. Url: https://ru.wikipedia.org/wiki/Жох,\_Олег\_Сергеевич

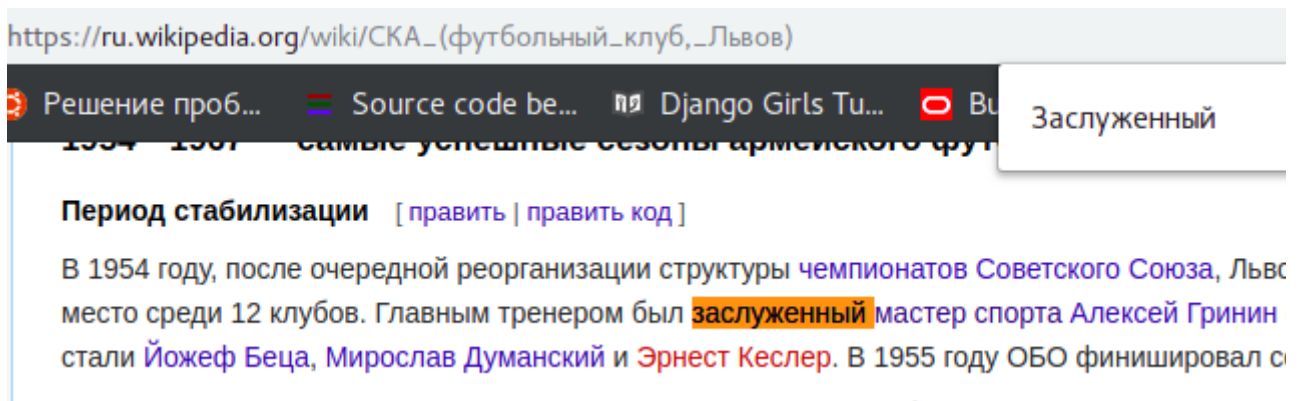
Заголовок: Березуцкий, Алексей Владимирович. Url: https://ru.wikipedia.org/wiki/Березуцкий,\_Алексей\_Владимирович
Заголовок: Веселов, Виктор Анатольевич. Url: https://ru.wikipedia.org/wiki/Веселов,\_Виктор\_Анатольевич
Заголовок: Емельянова, Ирина Витальевна. Url: https://ru.wikipedia.org/wiki/Емельянова,\_Ирина\_Витальевна
Заголовок: Шурыгин, Валерий Иванович. Url: https://ru.wikipedia.org/wiki/Шурыгин,\_Валерий\_Иванович
Заголовок: Каштанова, Наталья Анатольевна. Url: https://ru.wikipedia.org/wiki/Каштанова,\_Наталья\_Анатольевна
Заголовок: Дворников, Борис Геннадьевич. Url: https://ru.wikipedia.org/wiki/Дворников,\_Борис\_Геннадьевич
Заголовок: Гулиев, Зафар Сафар оглы. Url: https://ru.wikipedia.org/wiki/Гулиев,\_Зафар\_Сафар\_оглы
Заголовок: Жунисбаев, Несип Жунисбаевич. Url: https://ru.wikipedia.org/wiki/Жунисбаев,\_Несип\_Жунисбаевич
get_search_res works 0:00:00.203900

Запрос:
```

Результат для поискового запроса «заслуженный мастер спорта» - время работы 2 мск — это время поиска + вывода в консоль данных. Время работы поиска - 0.002 с. Запрос является не сложным.

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l6_boolsearch$ python l6.py
Запрос: заслуженный мастер спорта
Результат получен за 0:00:00.002327
```

Для наглядности не будем использовать поиск PyCharm, а воспользуемся поиском на странице wiki. Первый запрос слишком очевиден, возьмем статью про [https://ru.wikipedia.org/wiki/СКА_\(футбольный_клуб,_Львов\)](https://ru.wikipedia.org/wiki/СКА_(футбольный_клуб,_Львов)).



На странице все 3 этих слова встречаются.

Попробуем запрос «(заслуженный мастер) !спорта». Количество таких статей должно быть очень малым.

```
komp@komp-G5:~/stud/search_MAI/l6_boolsearch$
Запрос: (заслуженный мастер) !спорта
get_search_res works 0:00:00.007720
```

Как видно - ожидания оправдались. Поисковая выдача равна 0. PyCharm — так же не нашел файлов с таким сочетанием (использовался поиск по регулярным выражениям).

Заменяем слово «спорта» на какое-то высокочастотное слово — например «В».

```
Запрос: (заслуженный мастер) !v
Заголовок: Дикиев, Хасмагомед Магомедович. Url: https://ru.wikipedia.org/wiki/Дикиев, Хасмагомед Магомедович
get_search_res works 0:00:00.011679
```

```
Дикийев, Хасмагомед Магомедович.txt x
Q v
1 Хасмагомед Магомедович Дикиев (Аргун, Чечено-Ингушская АССР, РСФСР, СССР) — советский и российский чеченский самбист
2
3 Биография
4 Первый чеченец ставший чемпионом мира по самбо.
5
6 Спортивные достижения
7 Чемпионат СССР по самбо 1989 года — ;
8 Чемпионат мира по самбо 1989 года (Вест-Оранж, США) — ;
9 Кубок мира 1993 по самбо года — ;
10
11 Литература
12 С-А. М. Аслаханов, Х. Х. Хизриев. Физическая культура и спорт Чечни: истоки и современность. — Москва: «Перо», 201
13
14 Ссылки
15 Мовлади Абдулаев. Хасмагомед Дикиев (неопр.). Энциклопедия чеченского спорта (16 сентября 2013). Проверено 28 фев
16 Хасмагомед Дикиев — первый и единственный... (неопр.). Информационное агентство «Чечен-Инфо» (6 января 2015). Провер
17 Ваха Автуринский. Династия богатырей (неопр.). Вести республики (29 октября 2012). Проверено 1 марта 2015.
```

Была найдена одна статья. На этот раз я приведу пример поиска используя PyCharm, так как на web-странице не в тексте статьи есть предлог «в».

```
Дикийев, Хасмагомед Магомедович.txt x
Q заслуженный мастер
1 (Аргун, Чечено-Ингушская АССР, РСФСР, СССР) — советский и российский чеченский самбист, чемпион СССР и мира по самбо, победитель Кубка мира по самбо, заслуженный мастер
2
3 м мира по самбо.
4
5
6
7 да — ;
8 да (Вест-Оранж, США) — ;
9 - ;
10
11
12 в. Физическая культура и спорт Чечни: истоки и современность. — Москва: «Перо», 2015. — С. 136. — 240 с. — ISBN 978-500086-997-0.
13
14
15 киев (неопр.). Энциклопедия чеченского спорта (16 сентября 2013). Проверено 28 февраля 2015.
16 иственный... (неопр.). Информационное агентство «Чечен-Инфо» (6 января 2015). Проверено 1 марта 2015.
17 тьрей (неопр.). Вести республики (29 октября 2012). Проверено 1 марта 2015.
```

Теперь приведем время поиска статей с высокочастотными словами. Используем запрос «оно | в | на | он | я». В результат вошла почти вся тестовая выборка.

```
Заголовок: Летние Олимпийские игры 1940. Url: https://ru.wikipedia.org/wiki/Летние Олимпийские игры 1940
Заголовок: Панамериканские игры 1971. Url: https://ru.wikipedia.org/wiki/Панамериканские игры 1971
Заголовок: Чемпионат мира по крикету 2011. Url: https://ru.wikipedia.org/wiki/Чемпионат мира по крикету 2011
Заголовок: Виера, Юсниер. Url: https://ru.wikipedia.org/wiki/Виера, Юсниер
Заголовок: Григорьев, Владимир Яковлевич. Url: https://ru.wikipedia.org/wiki/Григорьев, Владимир Яковлевич
Заголовок: Зимние Паралимпийские игры 2010. Url: https://ru.wikipedia.org/wiki/Зимние Паралимпийские игры 2010
Заголовок: Летние Всемирные военные игры 2015. Url: https://ru.wikipedia.org/wiki/Летние Всемирные военные игры 2015
Заголовок: Список главных тренеров, выигравших Кубок УЕФА и Лигу Европы УЕФА. Url: https://ru.wikipedia.org/wiki/Список главных тренеров, выигравших Кубок УЕФА и Лигу Европы УЕФА
Заголовок: Чемпионат СССР по тяжёлой атлетике 1986. Url: https://ru.wikipedia.org/wiki/Чемпионат СССР по тяжёлой атлетике 1986
Заголовок: Бадминтон на Европейских играх 2019. Url: https://ru.wikipedia.org/wiki/Бадминтон на Европейских играх 2019
Заголовок: Хансен, Алан. Url: https://ru.wikipedia.org/wiki/Хансен, Алан
Заголовок: Гран-при Марсельезы. Url: https://ru.wikipedia.org/wiki/Гран-при Марсельезы
Articles count: 2705
get_search_res works 0:00:02.290107
```

Как видно — время работы почти 2 секунды. Однако стоит учитывать, что это полное время работы консольной утилиты, то есть львиная часть этого времени — это вывод запросов. Так сделано для более красивого вывода

данных. Если оставить только получение поисковой выдачи без ее печати, то время выполнения — 0.01 секунды.

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l6_boolsearch$ python l6.py
Запрос: оно | в | на | он | я
Результат получен за 0:00:00.013194
```

Оценка качества поиска

Воспользуемся тестовыми запросами, которые использовались в ЛР2 для оценки качества работы поисковых систем. Оценки были следующими:

Для поиска по википедии:

```
"search_sys": "wiki",

"P@1": 0.9655172413793104,
"P@3": 0.9195402298850575,
"P@5": 0.9195402298850575,
"DCG@1": 4.551724137931035,
"DCG@3": 9.223221721921302,
"DCG@5": 12.284175920616793,
"NDCG@1": 0.9103448275862069,
"NDCG@3": 0.6148814481280869,
"NDCG@5": 0.4913670368246717,
"ERR@1": 0.9103448275862069,
"ERR@3": 0.6148814481280869,
"ERR@5": 0.4913670368246717
```

Для google:

```
"search_sys": "google",

"P@1": 1.0,
"P@3": 1.0,
"P@5": 1.0,
"DCG@1": 4.896551724137931,
"DCG@3": 9.975266464532151,
"DCG@5": 13.657926517972932,
"NDCG@1": 0.9793103448275862,
"NDCG@3": 0.6650177643021434,
"NDCG@5": 0.5463170607189172,
"ERR@1": 0.9793103448275862,
"ERR@3": 0.6650177643021434,
"ERR@5": 0.5463170607189172
```

Оценки для созданной поисковой системы:

```
"search_sys": "bool",

"P@1": 0.8275862068965517,
```

```
"P@3": 0.8275862068965518,  
"P@5": 0.8275862068965518,  
"DCG@1": 3.2413793103448274,  
"DCG@3": 7.036870908374482,  
"DCG@5": 9.691327170149046,  
"NDCG@1": 0.6482758620689654,  
"NDCG@3": 0.4691247272249654,  
"NDCG@5": 0.38765308680596183,  
"ERR@1": 0.6482758620689654,  
"ERR@3": 0.4691247272249654,  
"ERR@5": 0.38765308680596183
```

Как видно, созданная поисковая система значительно уступает. Это обуславливается тем, что большая часть тестовых запросов, все-таки, больше рассчитана на цитатный поиск. К тому же выдача никак не ранжируется, что тоже ухудшает результат. Однако, если искать отдельные слова и сравнить количество найденных статей с тем, что выдают утилита того же PyCharm, то можно судить о том, что поиск работает правильно и количество найденных статей совпадает.

Вывод

В ходе лабораторной работы был реализован булев поиск. Изучены инструменты Python для работы с парсингом выражений.