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

Отчет по лабораторной работе №1
«Добыча корпуса документов»
по курсу
«Информационный поиск»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

Необходимо подготовить корпус документов, который будет использован при выполнении остальных лабораторных работ:

- Скачать его к себе на компьютер. В отчёте нужно указать источник данных.
- Ознакомиться с ним, изучить его характеристики.
- Разбить на документы.
- Выделить текст.
- Найти существующие поисковики, которые уже можно использовать для поиска по выбранному набору документов и привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.

Метод решения

Дамп английской википедии брался с ресурса https://meta.wikimedia.org/wiki/Data_dump_torrents, конкретно взят архив enwiki-20170301-pages-articles.xml.bz2.

В архиве находится xml-файл размера около 60Гб.

Для хранения извлечения текстов и их хранения, был написал скрипт на python3. Он выполняет следующие функции:

- построчный разбор xml файла,
- выделение текста, заголовка timestamp и id,
- очистка текста статьи от «форумных» тегов,
- сохранение полученных данных в sqlite, с использованием sqlalchemy, для дальнейшей обработки.

Журнал выполнения задания

В процессе обработки файла ядро питона «умерло», обработав порядка 310 000 статей, не смотря на то, что память полностью очищалась после обработки каждой статьи. Этот набор данных я и буду использовать, как начальный корпус документов. Так же взяты 10 000 случайных статей из этих 310 000 и дублированы в отдельную, тестовую базу данных.

Так же возникло порядка 15 ошибок следующего вида: удвоенный текст статьи не поместился в память.

Результаты

Выделено порядка 310 000 тысяч статей английской википедии.

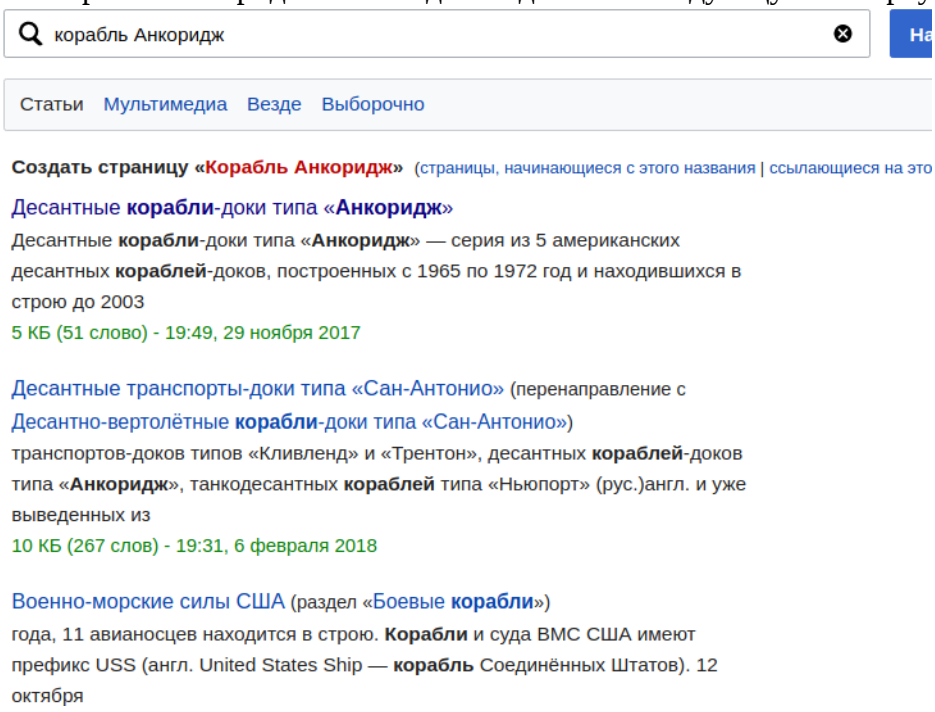
Вес базы данных с документами: 2.8Гб.

Для полного разбора необходимо переписать скрипт парсера из ipython notebook в обычный скрипт, а так же сменить БД на postgress.

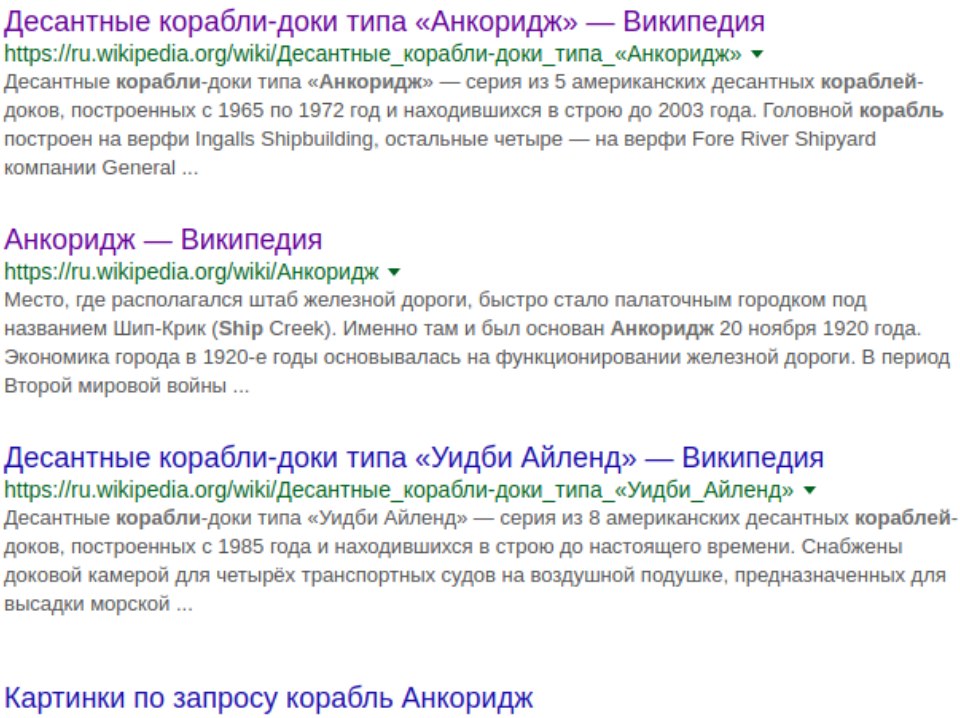
Примеры запросов к поисковикам

Возьмем случайную статью с википедии: Десантные корабли-доки типа «Анкоридж». Попробуем поискать ее, поисковым запросом будет Анкоридж.

Поиск по википедии отправляет на статью о городе с таким названием. Попробуем изменить запрос: корабль Анкоридж. Википедия выдаст нам следующую выборку:



Поиск в Google по запросу «Анкоридж» предлагает нам ту же статью об американском городе Анкоридж. Поменяем запрос на «корабль Анкоридж».



Видим, что первый результат в выдаче это искомая статья, второй же результат отличается от выдачи википедии, это статья про одноименный город, а не про другие корабли.

Яндекс на запрос выдает «Анкоридж» знакому статью о городе Анкоридж, при этом отличия в выдаче заключаются в наличии видео. При поиске «корабль Анкоридж» находим статью, а так же видео с различными кораблями.

Выводы

- Полученный корпус документов имеет относительно небольшой размер, это будет удобно для отладки «движка».
- Обработка большого объема документов требует большого количества ресурсов. Поэтому желательно использование только асимптотически оптимальных или близких к этому алгоритмов. Так же, в таких задачах узким местом чаще всего является чтение/запись данных с диска, а не их обработка.