Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №3

по курсу:
«Информационный поиск»
по теме:
«Булев индекс»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

Требуется построить поисковый индекс, пригодный для булева поиска, по подготовленному в ЛР1 корпусу документов. Требования к индексу:

• Самостоятельно разработанный, бинарный формат представления данных. Формат необходимо описать в отчёте, в побайтовом представлении.

• Формат должен предполагать расширение, т.к. в следующих работах он будет меняться под требования новых лабораторных работ.

• Использование текстового представления или готовых баз данных не допускается.

• Кроме обратного индекса, должен быть создан «прямой» индекс, содержащий в себе как минимум заголовки документов и ссылки на них (понадобятся для выполнения ЛР4, при генерации страницы поисковой выдачи).

• Для термов должна быть как минимум понижена капитализация. В отчёте должно быть отмечено как минимум:

• Выбранное внутренне представление документов после токенизации.

• Выбранный метод сортировки, его достоинства и недостатки для задачи индексации.

Оборудование:

Компьютер HP Omen 15 под управлением операционной системы Windows 10, Intel Core i5-7300HQ 2.50 GHz, 12 Gb RAM

Программное обеспечение

Язык программирования	Python 3.6
Среда программирования	Anaconda, Jupyter Notebook

Структура

В данной лабораторной работе я решил рассмотреть следующий формат для булева индекса: Использования словаря для хранения токенов и использования смещения для координатных блоков (два файла: словарь + файл координатных блоков со смещениями)

По битам:

Словарь:

					…									
4 байта- длина токена				N байт - токен			4 байт – указатель на начало координатных блоков							

Координатный блок

					…	
4 байта - длина блока doc-ids(Len)				Len*2(short) байт – блок doc-ids		

Время выполнения индексации

```
0.0
0.0
0.000997304916381836
0.0
0.0009624958038330078
0.001996278762817383
0.0009984970092773438
0.0009949207305908203
0.0
0.0
0.0009963512420654297
0.0
0.0009980201721191406
0.0
0.0
0.0010173320770263672
0.0
0.0009789466857910156

Wall time: 1min 35s
```

Прогон осуществлялся при размере выборки 40.000 статей.

На изображении отображено сколько времени требуется для работы с одним файлом.

Стоит заметить, что в моем случае алгоритм создания булева индекса от скорости считывания записи с диска/ оперативки и для записи в файл.

Как выглядят токены:

командование 1908 5504004

генерал 4150 5507820

люциан 24 5516120

Тестировалось все следующим образом: взял выборку из статей (три статьи) и проверил вхождение терминов из этих статей в координатный блок, если какого-то термина не было в той статье в которой была, то значит вся проверка выпадает по эксепшену. Из возможных ошибок, коллизия, местами я не позаботился о обработке тестирующего текста