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: М80-106М
Выполнил: Забелин М. К.
Преподаватель: Калинин А.Л.

Москва, 2019

Лабораторная работа №1. Добыча корпуса документов

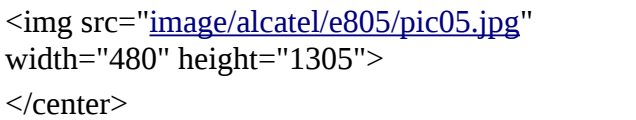
Необходимо подготовить корпус документов, который будет использован при выполнении остальных лабораторных работ

Для получения статей был написан парсер статей с сайта mobile-review.com. Все статьи сайта касаются мобильной техники. В результате был получен набор текстовых документов, разделенных по темам. Все документы находятся по адресу: https://mobile-review.com/review.shtml. При парсинге html страниц сохранял только саму статью без рекламы и ссылок на другие статьи.

Пример «грязного текста» (без очистки от тегов)

На левой боковой стороне находятся три музыкальных клавиши, с их помощью можно включить воспроизведение музыки в любой момент. Наличие FM-радио обусловлено исключительно гарнитурой, и все управление идет с нее же и только.

Разъем на корпусе предназначен для карт microSD, хотя во всех материалах говорится о поддержке miniSD-карт, но это не так. Поддерживается «горячая» замена карт памяти, объем карт до 1 Гб. Заглушка слота для карт из резины и прочно крепится к корпусу.



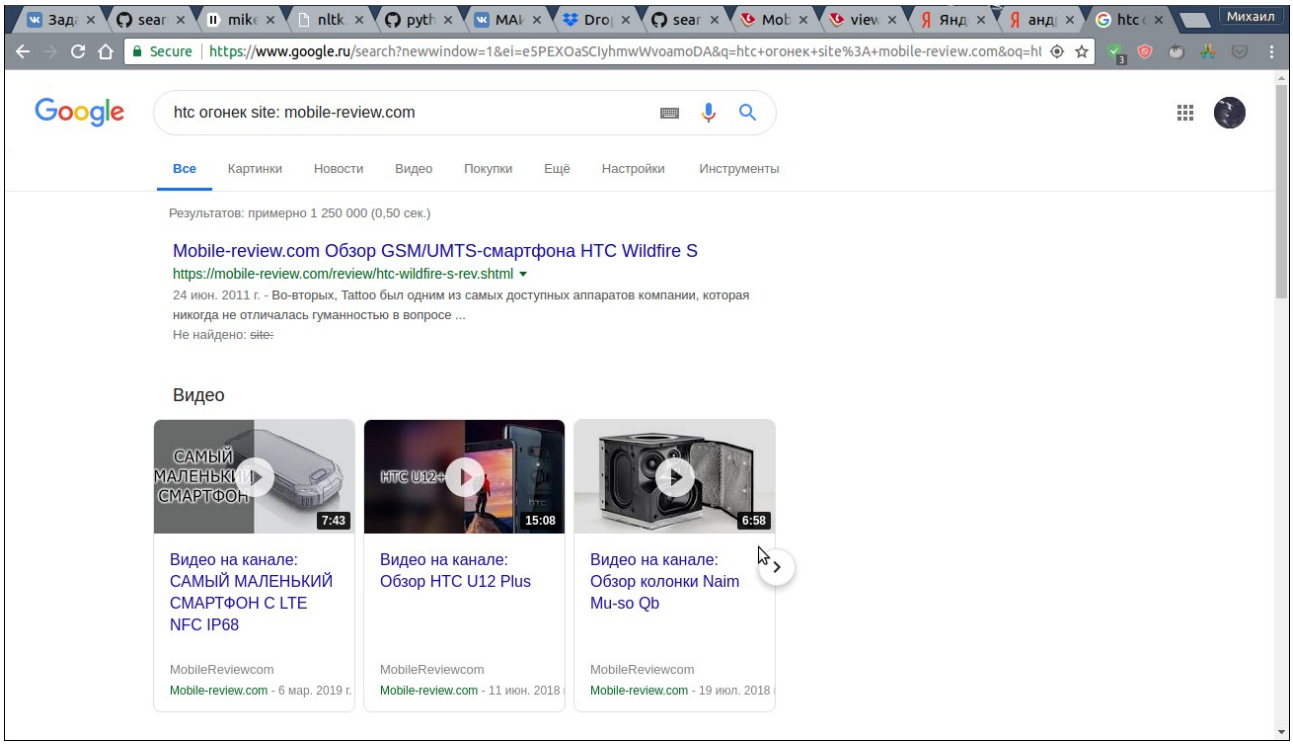
Этот исходный текст был обработан, и в результате была получен итоговый текста, который будет участвовать в поиске:

«На левой боковой стороне находятся три музыкальных клавиши, с их помощью можно включить воспроизведение музыки в любой момент. Наличие FM-радио обусловлено исключительно гарнитурой, и все управление идет с нее же и только.

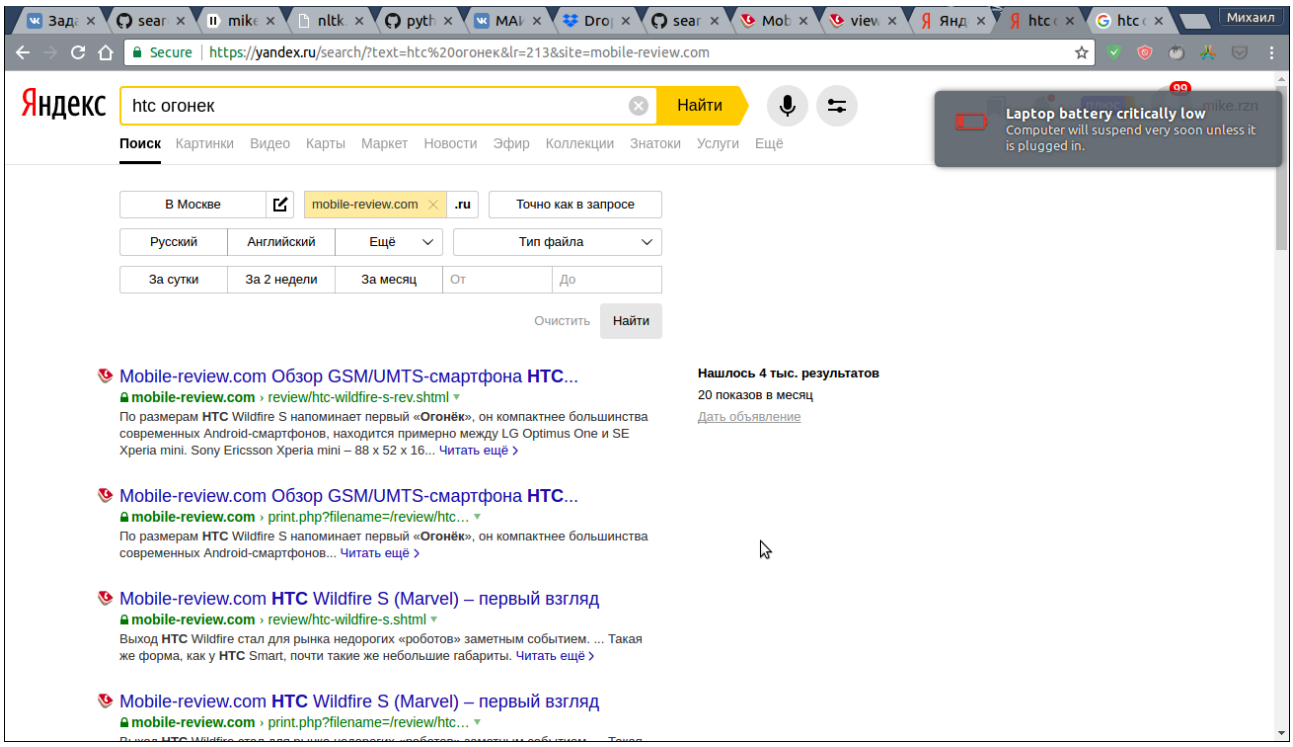
Разъем на корпусе предназначен для карт microSD, хотя во всех материалах говорится о поддержке miniSD-карт, но это не так. Поддерживается «горячая» замена карт памяти, объем карт до 1 Гб. Заглушка слота для карт из резины и прочно крепится к корпусу.»

Кокретно по этому сайту можно искать обычным поисковиком. Например, для яндекса нужно в самом интерфейса поика в поле ввода ввести сайт, по которому будет вестись поиск. Для гугла достаточно в самом запросе ввести после замого запроса site: some-site.com.

Поиск в google



Поиск в Yandex:



Статистическая информация о корпусе

- Размер сырых данных — неизвестно, так как не сохранялись
- Количество документов — 2994
- Размер «чистого» текста — 69,4 МБ
- Средний размер документа — 24.3 Кб
- Средний объем текста в документе — 13000 символов

Вывод

В ходе выполнения лабораторной я получил корпус документов по категории «Мобильная техника». Познакомился с библиотекой Beautiful Soup для Python.

Ссылка на Git репозиторий

https://github.com/Mikhail-Z/MAI/tree/master/sem10/Informational%20Search/is1