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 я воспользовался api википедии через библиотеку wikipediaapi для Python. В качестве категорий выступают «Информация» и «Техника».

Статистическая информация о корпусе

- Размер сырых данных — неизвестно, так как не сохранялись
- Количество документов — 13959
- Размер «чистого» текста — 119 МБ
- Средний размер документа — 8.7 Кб
- Средний объем текста в документе — 8300 символов

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)

<p>Minolta AF 50 F1.4 — нормальный объектив системы Minolta AF фирмы Minolta.

Итоговый документ после очистки от тегов имеет вид:

Minolta AF 50 F1.4 — нормальный объектив системы Minolta AF фирмы Minolta. Оригинальная версия объектива была выпущена в 1985 году. В 1990 году был несколько усовершенствован (версия RS). С 2006 года и по настоящее время усовершенствованная версия RS производится под маркой Sony, код — SAL-50F14. Один из самых светосильных объективов системы.

По википедии можно искать несколькими способами:

1. Поиск самой википедии
2. Обычными поисковиками с указанием сайта для поисковиков

Вывод

В ходе выполнения лабораторной я получил корпус документов по категории «Иформация» и «Техника». Познакомился api википедии.

Ссылка на Git репозиторий

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