Московский Авиационный Институт
(Научный Исследовательский Институт)

Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Информационный поиск»

Выполнил: Ефименко Н.А

Преподаватель: Калинин А.Л.

Москва, 2019 г.

ЛР1: Добыча корпуса документов

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. Из чего состоит текст? Есть ли дополнительная метаинформация? Если разметка текста, какая она?
- Разбить на документы.
- Выделить текст.
- Найти существующие поисковики, которые уже можно использовать для поиска по выбранному набору документов (встроенный поиск Википедии, поиск Google с использованием ограничений на URL или на сайт). Если такого поиска найти невозможно, то использовать корпус для выполнения лабораторных работ нельзя!
- Привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.

Метод решения

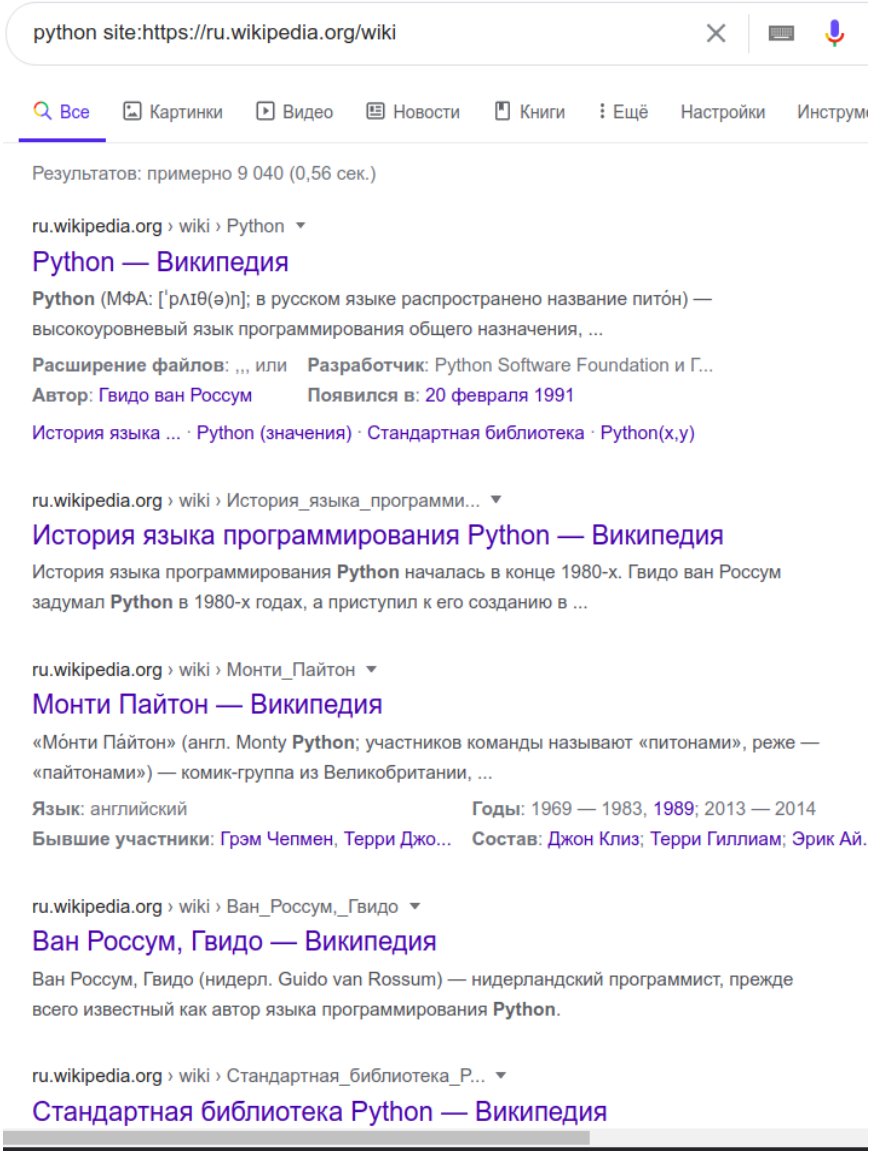
1. Изучение способов скачивания статей из Википедии.
2. Выбор темы для корпуса документов.
3. Экспорт статей при помощи wikiAPi.
4. Выделения из них текста.
5. Написание отчёта с выполнением последних двух пунктов задания.

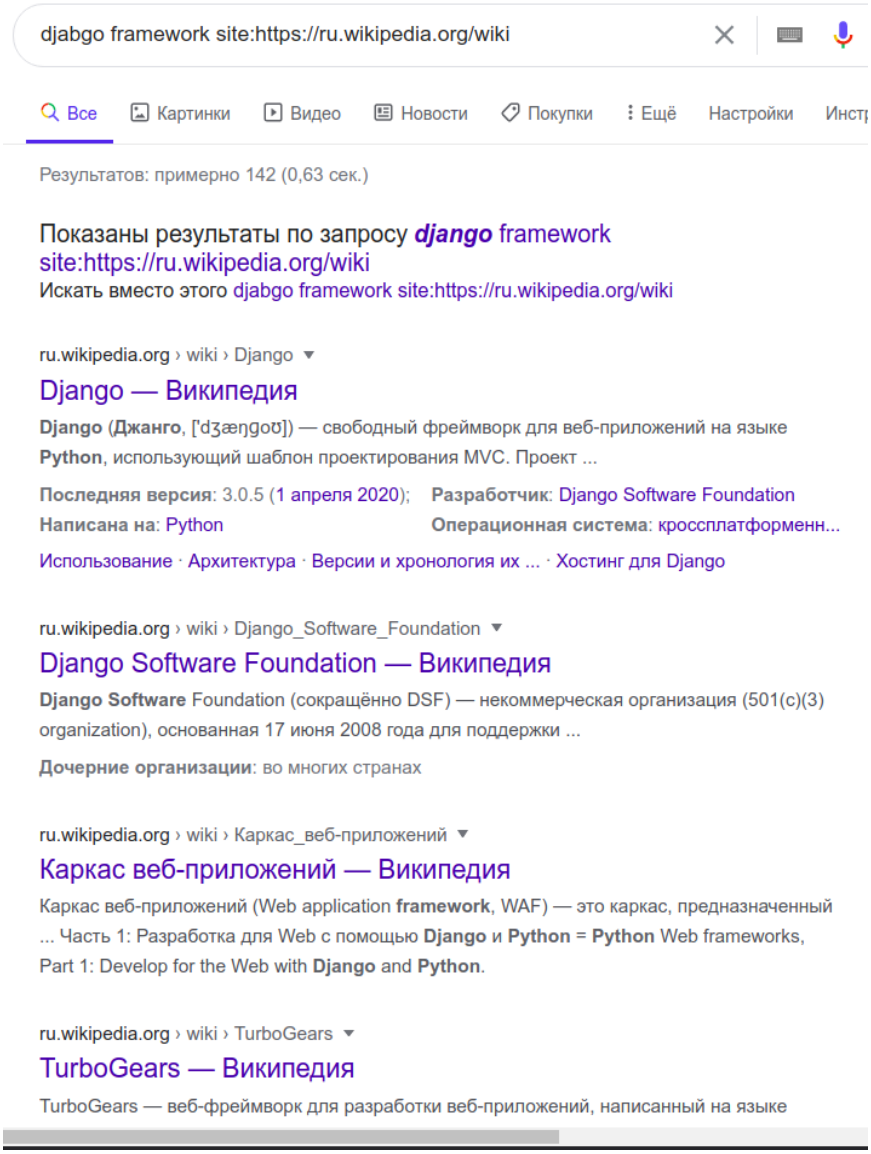
Информация о корпусе

Источник данных	Wikipedia.org
Тема корпуса	Computer programming
Размер «сырых» данных	489 Мб
Количество статей	25000
Размер текста, выделенного из «сырых» данных	123 Мб
Средний размер документа, средний объём текста в документе	20 Кб 17 Кб

Исходные данные статей, скачаны через wikiapi — библотеки python реализующей открытое api wikipedia. Rроме непосредственно текста статей в них имеется html разметка документа и различная метаинформация (например, источник данных, версия медиа-вики).

Примеры запросов





Недостатками в полученной поисковой выдаче является то, что кроме результатов, относящихся непосредственно к теме, так же в выдачу попадают статьи, в которых упоминаются слова из поискового запроса.

Выводы

В процессе выполнения данной лабораторной работы была выбрана тематика документов для последующего информационного поиска в них, скачан соответствующий ей корпус документов в размере 25000 статей. Разбит на документы по одной подкатегории, из которых выделен текст. Изучена информация о размере данных.