

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 метainформация? 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 — библиотеки python реализующей открытое api wikipedia. Кроме непосредственно текста статей в них имеется html разметка документа и различная метainформация (например, источник данных, версия медиа-вики).

Примеры запросов

python site:https://ru.wikipedia.org/wiki

✕

🔍 Все

🖼️ Картинки

📺 Видео

📰 Новости

📖 Книги

⋮ Ещё

⚙️ Настройки

🛠️ Инструм

Результатов: примерно 9 040 (0,56 сек.)

ru.wikipedia.org › wiki › Python ▾

Python — Википедия

Python (МФА: [ˈpɫɪθ(ə)n]; в русском языке распространено название **пито́н**) — высокоуровневый язык программирования общего назначения, ...

Расширение файлов: [...](#) или [Разработчик: Python Software Foundation и Г...](#)

Автор: [Гвидо ван Россум](#) [Появился в: 20 февраля 1991](#)

[История языка ...](#) · [Python \(значения\)](#) · [Стандартная библиотека](#) · [Python\(x,y\)](#)

ru.wikipedia.org › wiki › История_языка_программи... ▾

История языка программирования Python — Википедия

История языка программирования **Python** началась в конце 1980-х. Гвидо ван Россум задумал **Python** в 1980-х годах, а приступил к его созданию в ...

ru.wikipedia.org › wiki › Монти_Пайтон ▾

Монти Пайтон — Википедия

«Мóнти Па́йтон» (англ. **Monty Python**; участников команды называют «питонами», реже — «пайтонами») — комик-группа из Великобритании, ...

Язык: английский [Годы: 1969 — 1983, 1989; 2013 — 2014](#)

Бывшие участники: [Грэм Чепмен](#), [Терри Джо...](#) [Состав: Джон Клиз; Терри Гиллиам; Эрик Ай.](#)

ru.wikipedia.org › wiki › Ван_Россум,_Гвидо ▾

Ван Россум, Гвидо — Википедия

Ван Россум, Гвидо (нидерл. Guido van Rossum) — нидерландский программист, прежде всего известный как автор языка программирования **Python**.

ru.wikipedia.org › wiki › Стандартная_библиотека_P... ▾

Стандартная библиотека Python — Википедия

djabgo framework site:https://ru.wikipedia.org/wiki

Результатов: примерно 142 (0,63 сек.)

Показаны результаты по запросу **djabgo framework**
site:https://ru.wikipedia.org/wiki
Искать вместо этого djabgo framework site:https://ru.wikipedia.org/wiki

ru.wikipedia.org › wiki › Django ▼

Django — Википедия

Django (Джанго, [ˈdʒæŋɡoʊ]) — свободный фреймворк для веб-приложений на языке Python, использующий шаблон проектирования MVC. Проект ...

Последняя версия: 3.0.5 (1 апреля 2020); Разработчик: Django Software Foundation
Написана на: Python Операционная система: кроссплатформенн...

Использование · Архитектура · Версии и хронология их ... · Хостинг для Django

ru.wikipedia.org › wiki › Django_Software_Foundation ▼

Django Software Foundation — Википедия

Django Software Foundation (сокращённо DSF) — некоммерческая организация (501(c)(3) organization), основанная 17 июня 2008 года для поддержки ...

Дочерние организации: во многих странах

ru.wikipedia.org › wiki › Каркас_веб-приложений ▼

Каркас веб-приложений — Википедия

Каркас веб-приложений (Web application framework, WAF) — это каркас, предназначенный ... Часть 1: Разработка для Web с помощью Django и Python = Python Web frameworks, Part 1: Develop for the Web with Django and Python.

ru.wikipedia.org › wiki › TurboGears ▼

TurboGears — Википедия

TurboGears — веб-фреймворк для разработки веб-приложений, написанный на языке

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.