

Московский Авиационный Институт
(Научный Исследовательский Институт)

Факультет информационных технологий и прикладной
математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Информационный поиск»

Студент: Бобылев Д.М.

Преподаватель: Кухтичев А.А.

Группа: М8О-206М

Дата:

Оценка:

Подпись:

Москва, 2021

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 метаданная? 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 получения статей Википедии.
2. Экспорт статей Википедии
3. Изучение характеристик дампа Википедии
4. Изучение способов парсинга
5. Выделение текста из сырых данных
6. Изучение существующих поисковиков, применимых для поиска по выбранному набору документов
7. Формирование статистической информации о корпусе
8. Написание отчета

Журнал выполнения

№	Действие	Проблема	Решение
1	Попытка использовать wikipedia api для формирования корпуса документов по категории	1) требуется время на изучение api 2) Выполнение рекурсивного поиска статей Википедии требует много времени 3) Необходимо	Получение корпуса документов скачиванием полного дампа статей Википедии в формате xml

		придумывать решение по выходу из циклов (если статьи ссылаются друг на друга) и по дедупликации статей	
2	Использование библиотеки wikiextractor для извлечения текста из сырых данных	1) текст извлекается в формате xml 2) Библиотека не работает на python 3.9	1) поиск версии библиотеки, где поддерживается выгрузка в json формате 2) Использование docker образа с python версии 3.7

Информация о корпусе

- статистика

Источник данных	ruwiki-20211001-pages-articles.xml.bz2 - дамп статей российской википедии от 01/10/2021
Размер «сырых» данных	25,56 гб
Количество статей	1757893
Размер текста, выделенного из «сырых» данных	6,5 гб
Количество документов	6711
Средний размер документа	1 мб

- характеристики

Дамп статей википедии представляет из себя один xml файл размером 25,56 гб, содержащий статьи с состоянием, актуальным на момент создания дампа. Xml файл заархивирован в bz2.

- процесс извлечения текста

Удалось извлечь текст статей из сырых данных, используя библиотеку wikiextractor <https://github.com/attardi/wikiextractor>

Для упрощения использования утилиты был написан Dockerfile:

```
FROM python:3.7
RUN pip install wikiextractor==0.1
CMD python -m wikiextractor.WikiExtractor
```

```
docker build --tag wikiextracor:1.0 .
```

Используя утилиту в docker, не нужно задумываться о версии python и о pip зависимостях.

```
docker run -v <host_mount_path>:<docker_mount_path> wikiextracor:1.0
```

Процесс извлечения занял 45 минут:

INFO: Finished 5-process extraction of 1757893 articles in 2654.7s (662.2 art/s)

INFO: total of page: 2756384, total of article page: 1757893; total of used article page: 1757893

- Фрагмент сырых данных

```
<page>
  <title>Киевская Русь</title>
  <ns>0</ns>
  <id>27</id>
  <revision>
    <id>116908122</id>
    <parentid>116609009</parentid>
    <timestamp>2021-09-28T13:20:52Z</timestamp>
    <contributor>
      <username>Furyone648</username>
      <id>2871042</id>
    </contributor>
    <minor />
    <comment>оформление</comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text bytes="277625">
xml:space="preserve">{{перенаправление|Древнерусское государство|Русское
государство|о государстве XV–XVIII веков}}
{{Другие значения}}
{{Историческое государство
|название = Киевская Русь<br>Древнерусское государство<br>Древняя
Русь
|самоназвание = {{lang-ory2|роусьская земля}}
|статус = Историческое государство
|карта = Rus-1015-1113.png
|герб = -
|флаг = -
|размер = 300px
|столица = [[Великий Новгород|Новгород]] &lt;!-- Д. А. Гутнов Популярный
обзор русской истории. VI–XVI вв; Введенский А. М.
[http://slovene.ru/2014_1_Vvedenskiy.pdf Стольный город в древнерусских и
фольклорных источниках] // Slověne = Словѣне.- № 1. — 2014. — С. 207.--
&gt;(862–882),&lt;br>[[Киев]] (882–1240)
|города = [[Киев]], [[Великий Новгород|Новгород]], [[Чернигов]],
[[Переяславль Киевский|Переяславль]], [[Полоцк]], [[Туров]], [[Владимир-
Волынский]], [[Смоленск]], [[Псков]], [[Новгород-Северский]],
[[Пшемысль|Перемышль]], [[Теребовля|Теребовль]], [[Галич (Ивано-Франковская
область)|Галич]], [[Ростов]], [[Суздаль]], [[Владимир (город)|Владимир-на-
Клязьме]], [[Муром]], [[Старая Рязань|Рязань]]
|образовано = [[862]]/[[882]]
|ликвидировано = [[1132]]/[[1240]]
```


```
|династия = [[Рюриковичи]]
|язык = [[Древнерусский язык|древнерусский]] &lt;small&gt;(разговорный,
деловой, язык государственного управления,
права)&lt;/small&gt;;&lt;br&gt;[[Древнерусский извод церковнославянского
языка|древнерусский извод церковнославянского]] &lt;small&gt;(язык книжной
культуры)&lt;/small&gt;&lt;ref
наме=&quot;Крысько&quot;&gt;{{БРЭ|автор=[[Крысько, Вадим Борисович|Крысько В.
Б.]]|заглавие=Древнерусский язык|том=9|страницы=339–
340|год=2007|id=2631460|ref=БРЭ}}&lt;/ref&gt;&lt;ref
наме=&quot;Калугин&quot;&gt;{{БРЭ|автор=Калугин В. В.|заглавие=Древнерусская
литература|том=Россия|страницы=703–
712|год=2004|id=5061392|ref=БРЭ}}&lt;/ref&gt;
|площадь = ок. {{num|1330000|км²}} (к 1000 году){{sfn|Урланис|1941|с=85}}
|население = 5,4 млн чел. (1000){{sfn|Урланис|1941|с=86}}
|валюта = [[Куна (денежная единица Древней Руси)|куна]], [[Гривна (денежная и
весовая единица Древней Руси)|гривна]], [[Ногата (монета)|ногата]]
|p1 = Племенные союзы восточных славян
&lt;!-- |p2 = Новгородская Русь
|p3 =
--&gt;
|s1 = Русские княжества
|форма_правления=[[раннефеодальная монархия]]}}
```

- Фрагмент извлеченного текста

```
{
  "id": "4749434",
  "url": "https://ru.wikipedia.org/wiki?curid=4749434",
  "title": "Чемпионат СССР по международным шашкам среди женщин 1975",
  "text": "Чемпионат СССР по международным шашкам среди женщин 1975\n\nПервый Чемпионат СССР по международным шашкам среди женщин 1975 года прошёл 1–16 сентября в городе Бендеры, Молдавская ССР по круговой системе. В нём приняли участие 14 спортсменов, среди которых было несколько 15–16-летних шашисток из Семипалатинска, Нижнего Тагила и Москвы. За победу давалось 1 очко, за ничью ½ очка и 0 за поражение.\n\nПервой чемпионкой страны стала 17-летняя Любовь Травина из Вильнюса. Второе место у 16-летней киевлянки Ольги Беляевой, на третьем месте финишировала чемпионка мира Елена Михайловская. \n\nЛюбовь Травина на старте набрала 5,5 из 6 очков, и , несмотря на два поражения, одержав больше всех побед — 9 удержала первое место.\n\n"}
}
```

Примеры запросов

- Хирурги 20 века



хирурги 20 века site: <https://ru.wikipedia.org/wiki/>

Всё

Картинки

Новости

Видео

Карты

Ещё

Инструменты

На всех языках ▾ За всё время ▾ Все результаты ▾

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/Хирургия) ▾ wiki ▾ Хирургия ▾

Хирургия - — Википедия

Достижения некоторых **хирургов** средневековья были весьма существенными. Итальянский **хирург** Лукка ещё в XIII **веке** для обезболивания использовал специальные...
Не найдено: site: | Запрос должен включать: [site:](#)

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/История_хирургии) ▾ wiki ▾ История_хирургии ▾

История хирургии - — Википедия

История медицины и **хирургии** — область изучения одной из самых древних наук в истории ... 3) первой половине XX в.; 4) второй половине XX — начале XXI **века**.
Не найдено: site: | Запрос должен включать: [site:](#)

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/Воронов,_Сергей_Аб...) ▾ wiki ▾ Воронов,_Сергей_Аб... ▾

Воронов, Сергей Абрамович - — Википедия

Серге́й (Самуи́л) Абра́мович Во́ронов (фр. Samuel (Serge) Voronoff — Серж Воронов; 10 июля 1866, Шехмань, Тамбовская губерния, Российская империя — 3 ...

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/Бокерия,_Лео_Антон...) ▾ wiki ▾ Бокерия,_Лео_Антон... ▾

Бокерия, Лео Антонович - — Википедия

Сеченова Министерства здравоохранения СССР и аспирантуру того же института в 1968 году. С 1968 года работает в Институте сердечно-сосудистой **хирургии** имени А. Н ...

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/Россия) ▾ wiki ▾ Россия ▾

Россия - — Википедия

Площадь лесных массивов сократилась на 20,3 млн га (1-е место в мире). Распространена также нелегальная вырубка (особенно на Северо-Западе и на Дальнем Востоке) ...
Не найдено: site: | Запрос должен включать: [site:](#)

[https://ru.wikipedia.org](https://ru.wikipedia.org/wiki/Углов,_Фёдор_Григо...) ▾ wiki ▾ Углов,_Фёдор_Григо... ▾

Углов, Фёдор Григорьевич - — Википедия

Заблудовский, услышав о хирургических отчётах Углова, в 30-е годы 20-го **века** сказал: «Это всё выдумки барона Мюнхгаузена». Позднее, изучив операции, проведённые ...
Не найдено: site: | Запрос должен включать: [site:](#)

- Хирург Пирогов

Google

хирург Пирогов site: https://ru.wikipedia.org/wiki/

Всё Картинки Новости Карты Видео Ещё Инструменты

Результатов: примерно 12 700 (0,46 сек.)

https://ru.wikipedia.org › wiki › Пирогов,_Николай_И... ▾
Пирогов, Николай Иванович - — Википедия
Вишня (ныне — в черте Винницы), Подольская губерния) — русский **хирург** и учёный-анатом, естествоиспытатель и педагог, профессор, создатель первого атласа ...
Не найдено: site: | Запрос должен включать: [site:](#)
[Операция Пирогова](#) · [Топографическая анатомия](#) · [Склифосовский, Николай...](#)

https://ru.wikipedia.org › wiki › Файл:Пирогов_Никол... ▾
Файл:Пирогов Николай Иванович.jpg - — Википедия
Русский: Никола́й Ива́нович **Пирого́в** (1810— 1881) — русский **хирург** и анатом, естествоиспытатель и педагог, ... Источник, <http://letopis.msu.ru/peoples/1883>.

https://ru.wikipedia.org › wiki › Пирогов,_Николай ▾
Пирогов, Николай - — Википедия
Николай **Пирогов**: **Пирогов**, Николай Васильевич (1872—1913) — русский художник.
Пирогов, Николай Иванович (1810—1881) — русский **хирург** и анатом, ...
Не найдено: site: | Запрос должен включать: [site:](#)

https://ru.wikipedia.org › wiki › Память_о_Николае_П... ▾
Память о Николае Пирогове - — Википедия
Николай Иванович **Пирогов** (1810—1881) — выдающийся русский **хирург**. В Виннице, где он умер, регулярно проводятся Пироговские чтения.
Не найдено: site: | Запрос должен включать: [site:](#)

https://ru.wikipedia.org › wiki › Памятник_Пирогову_... ▾
Памятник Пирогову (Москва) - — Википедия
Памятник Никола́ю Ива́новичу Пирого́ву — первый памятник врачу в Москве, посвящён **хирургу** Николаю **Пирогову**. Изготовлен скульптором Владимиром Шервудом.
Не найдено: site: | Запрос должен включать: [site:](#)

Среди недостатков полученной поисковой выдачи можно выделить то, что в результат попадают статьи, несоответствующие целевому запросу.

Выводы

В процессе выполнения данной лабораторной работы был получен корпус документов для выполнения последующих лабораторных работ по курсу информационного поиска - статьи Википедии. Статистика по корпусу состоит в следующем: размер «сырых» данных - 25,56 гб ; количество статей - 1757893; размер текста, выделенного из «сырых» данных - 6,5 гб. В ходе выполнения ЛР был неприятно удивлен отсутствием возможности скачать дампы конкретной категории Википедии, форматом дампа (xml). Понравилось то, что мне удалось сэкономить время на извлечение текста из сырых данных, используя утилиту `wikiextractor`.