Московский Авиационный Институт
(Научный Исследовательский Институт)

Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу

«Информационный поиск»

Студент: Мхитарян С.А.

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 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 в xml формате.
4. Выделения из них текста.
5. Написание отчёта с выполнением последних двух пунктов задания.

Журнал выполнения

№	Действие	Проблема	Решение
1	Установка и использование библиотеки Wikiextractor (https://github.com/attardi/wikiextractor) для обработки статей Википедии	Ярко выраженных проблем не обнаружено. Возможны проблемы со скоростью обработки англоязычных статей википедии из-за их количества.	Использование библиотеки.

Информация о корпусе

Источник данных	https://dumps.wikimedia.org/ruwiki/latest/ruwiki-latest-pages-articles.xml.bz2
Размер «сырых» данных	Со сжатием – 4.47ГБ , без сжатия – 25.8ГБ
Количество документов	19265
Количество статей	3938068
Размер текста, выделенного из «сырых» данных	18ГБ
Средний размер документа	1МБ
Средний объем текста	0.5МБ

Исходный текст в xml

```
<page>
    <title>Литва</title>
    <ns>0</ns>
    <id>7</id>
    <revision>
      <id>117955298</id>
      <parentid>117942923</parentid>
      <timestamp>2021-11-18T12:43:36Z</timestamp>
      <contributor>
        <ip>46.56.246.35</ip>
      </contributor>
      <comment>/* Язык */Исправлена ошибка.</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="124694" xml:space="preserve">{{другие значения}}
{{Государство
| Русское название           = Литовская Республика
| Оригинальное название      = {{lang-lt|Lietuvos Respublika}}
| Родительный падеж          = Литвы
| Герб                       = Coat of Arms of Lithuania.svg
| Название гимна             = Tautiška giesmė
| Флаг                       = Flag of Lithuania.svg
| Аудио                      = Tautiška_giesme_instumental.ogg
| Этап1                      = [[Битва при Сауле]]
| Дата1                      = [[1236 год]]
| Этап2                      = Коронация [[Миндовг]]а
| Дата2                      = [[1253 год]]
```

Обработанный текст в json


```
{"id": "7", "revid": "117955298", "url":
"https://ru.wikipedia.org/wiki?curid=7",
"title": "Литва", "text": "Литва́ ( ),
официальное название — Литóвская
Респу́блика () — государство,
расположенное в северной части Европы.
Площадь — км². Протяжённость с севера на
юг — 280 км, а с запада на восток — 370
км. Население составляет человек
(январь, 2021). Занимает 137-е место в
мире по численности населения и 121-е по
территории. Имеет выход к Балтийскому
морю, расположена на его восточном
побережье. Береговая линия составляет
всего 99 км (наименьший показатель среди
государств Балтии). На севере граничит с
Латвией, на юго-востоке — с Белоруссией,
на юго-западе — с Польшей и
Калининградской областью России. По
площади и населению является самым
крупным прибалтийским государством.
Столица — Вильнюс. Официальный язык — ли
товский. Денежная единица — евро.
Восстановление независимос
```

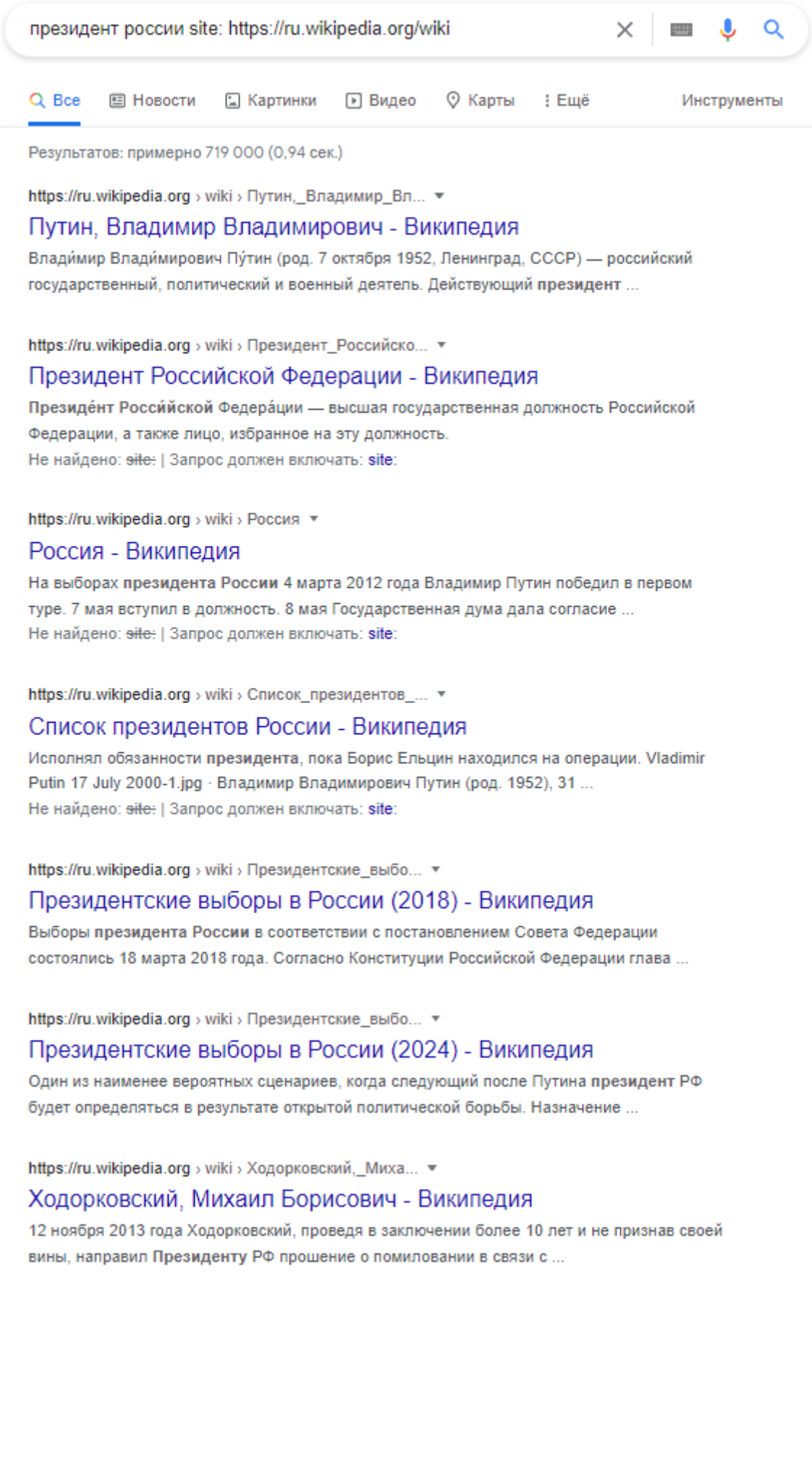

Лог преобразования из xml в json

```
 ~/projects/MAI-IR  on main ?2 ..................... INT  MAI-IR py  at
20:20:20
> python -m wikiextractor.WikiExtractor --json ~/Downloads/ruwiki-latest-
pages-articles.xml.bz2
INFO: Preprocessing '/Users/samvel/Downloads/ruwiki-latest-pages-
articles.xml.bz2' to collect template definitions: this may take some
time.
INFO: Preprocessed 100000 pages
INFO: Preprocessed 200000 pages
INFO: Preprocessed 300000 pages
INFO: Preprocessed 400000 pages
INFO: Preprocessed 500000 pages
INFO: Preprocessed 600000 pages
INFO: Preprocessed 700000 pages
INFO: Preprocessed 800000 pages
INFO: Preprocessed 900000 pages
INFO: Preprocessed 1000000 pages
INFO: Preprocessed 1100000 pages
INFO: Preprocessed 1200000 pages
INFO: Preprocessed 1300000 pages
INFO: Preprocessed 1400000 pages
INFO: Preprocessed 1500000 pages
INFO: Preprocessed 1600000 pages
INFO: Preprocessed 1700000 pages
INFO: Preprocessed 1800000 pages
INFO: Preprocessed 1900000 pages
```

```
INFO: Preprocessed 2000000 pages
INFO: Preprocessed 2100000 pages
INFO: Preprocessed 2200000 pages
INFO: Preprocessed 2300000 pages
INFO: Preprocessed 2400000 pages
INFO: Preprocessed 2500000 pages
INFO: Preprocessed 2600000 pages
INFO: Preprocessed 2700000 pages
INFO: Preprocessed 2800000 pages
INFO: Preprocessed 2900000 pages
INFO: Preprocessed 3000000 pages
INFO: Preprocessed 3100000 pages
INFO: Preprocessed 3200000 pages
INFO: Preprocessed 3300000 pages
INFO: Preprocessed 3400000 pages
INFO: Preprocessed 3500000 pages
INFO: Preprocessed 3600000 pages
INFO: Preprocessed 3700000 pages
INFO: Preprocessed 3800000 pages
INFO: Preprocessed 3900000 pages
INFO: Preprocessed 4000000 pages
INFO: Preprocessed 4100000 pages
INFO: Preprocessed 4200000 pages
INFO: Preprocessed 4300000 pages
INFO: Preprocessed 4400000 pages
INFO: Preprocessed 4500000 pages
INFO: Preprocessed 4600000 pages
INFO: Preprocessed 4700000 pages
INFO: Preprocessed 4800000 pages
INFO: Preprocessed 4900000 pages
INFO: Loaded 184647 templates in 896.7s
INFO: Starting page extraction from /Users/samvel/Downloads/ruwiki-latest-
pages-articles.xml.bz2.
INFO: Using 11 extract processes.
INFO: Extracted 100000 articles (1224.9 art/s)
INFO: Extracted 200000 articles (1976.9 art/s)
INFO: Extracted 300000 articles (3209.9 art/s)
INFO: Extracted 400000 articles (2378.5 art/s)
INFO: Extracted 500000 articles (2592.4 art/s)
INFO: Extracted 600000 articles (2594.1 art/s)
INFO: Extracted 700000 articles (2402.6 art/s)
INFO: Extracted 800000 articles (2264.8 art/s)
INFO: Extracted 900000 articles (2176.6 art/s)
INFO: Extracted 1000000 articles (2388.3 art/s)
INFO: Extracted 1100000 articles (2312.5 art/s)
INFO: Extracted 1200000 articles (2327.7 art/s)
INFO: Extracted 1300000 articles (2556.7 art/s)
INFO: Extracted 1400000 articles (2513.4 art/s)
INFO: Extracted 1500000 articles (2372.6 art/s)
INFO: Extracted 1600000 articles (2216.4 art/s)
INFO: Extracted 1700000 articles (2495.4 art/s)
INFO: Extracted 1800000 articles (2349.6 art/s)
INFO: Extracted 1900000 articles (2325.4 art/s)
INFO: Extracted 2000000 articles (2337.8 art/s)
INFO: Extracted 2100000 articles (2279.3 art/s)
INFO: Extracted 2200000 articles (2296.8 art/s)
INFO: Extracted 2300000 articles (2380.9 art/s)
INFO: Extracted 2400000 articles (2616.6 art/s)
INFO: Extracted 2500000 articles (2918.7 art/s)
INFO: Extracted 2600000 articles (3155.8 art/s)
INFO: Extracted 2700000 articles (2697.5 art/s)
```

```
INFO: Extracted 2800000 articles (2498.5 art/s)
INFO: Extracted 2900000 articles (2701.1 art/s)
INFO: Extracted 3000000 articles (2616.3 art/s)
INFO: Extracted 3100000 articles (2537.1 art/s)
INFO: Extracted 3200000 articles (2693.8 art/s)
INFO: Extracted 3300000 articles (2215.4 art/s)
INFO: Extracted 3400000 articles (2186.5 art/s)
INFO: Extracted 3500000 articles (2183.1 art/s)
INFO: Extracted 3600000 articles (2129.2 art/s)
INFO: Extracted 3700000 articles (2207.7 art/s)
INFO: Extracted 3800000 articles (2175.6 art/s)
INFO: Extracted 3900000 articles (2184.7 art/s)
INFO: Finished 11-process extraction of 3938068 articles in 1676.1s
(2349.6 art/s)
```

Примеры запросов



Можно заметить, что попадаются не совсем правильные варианты для поискового запроса.

Выводы

При выполнении лабораторной работы был получен корпус документов в формате json для выполнения следующих лабораторных работ. Найденная библиотека Wikiextractor оказалась очень даже эффективной, преобразовав 25ГБ русской Википедии из формата xml в json за 30 минут.