

**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: Синдюков В.Р.

Преподаватель: Кухтичев А.А.

Группа: М8О-208М

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 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	Использование библиотеки <code>wikiextractor</code> для извлечения текста из сырых данных	1) текст извлекается в формате <code>xml</code> 2) Некоторые теги, необходимые для дальнейшего считывания <code>xml</code> файла отсутствуют	Был написан скрипт, который добавлял в файлы необходимый тег

Информация о корпусе

- статистика

Источник данных	enwikinews-20211201-pages-articles-multi-stream.xml.bz2
Размер «сырых» данных	46.8 мб
Количество документов	40
Средний размер документа	1 мб

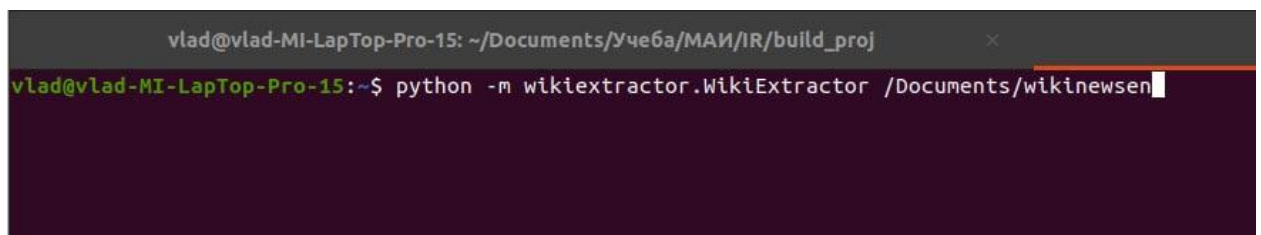
- характеристики

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 извлечения данных команда извлечения была вызвана прямо из терминала

A screenshot of a terminal window with a dark background. The title bar at the top reads "vlad@vlad-MI-LapTop-Pro-15: ~/Documents/Учеба/МАН/IR/build_proj" with a close button on the right. The terminal shows a command prompt "vlad@vlad-MI-LapTop-Pro-15:~\$" followed by the command "python -m wikiextractor.WikiExtractor /Documents/wikinews.xml". The cursor is at the end of the command.

```
vlad@vlad-MI-LapTop-Pro-15: ~/Documents/Учеба/МАН/IR/build_proj
vlad@vlad-MI-LapTop-Pro-15:~$ python -m wikiextractor.WikiExtractor /Documents/wikinews.xml
```

- Фрагмент сырых данных

```

1 <mediawiki xmlns="http://www.mediawiki.org/xml/export-0.10/" xmlns:xsi="http://www.w3.org
2 <siteinfo>
3 <sitename>Wikinews</sitename>
4 <dbname>enwikinews</dbname>
5 <base>https://en.wikinews.org/wiki/Main_Page</base>
6 <generator>MediaWiki 1.38.0-wmf.7</generator>
7 <case>first-letter</case>
8 <namespaces>
9 <namespace key="-2" case="first-letter">Media</namespace>
10 <namespace key="-1" case="first-letter">Special</namespace>
11 <namespace key="0" case="first-letter" />
12 <namespace key="1" case="first-letter">Talk</namespace>
13 <namespace key="2" case="first-letter">User</namespace>
14 <namespace key="3" case="first-letter">User talk</namespace>
15 <namespace key="4" case="first-letter">Wikinews</namespace>
16 <namespace key="5" case="first-letter">Wikinews talk</namespace>
17 <namespace key="6" case="first-letter">File</namespace>
18 <namespace key="7" case="first-letter">File talk</namespace>
19 <namespace key="8" case="first-letter">MediaWiki</namespace>
20 <namespace key="9" case="first-letter">MediaWiki talk</namespace>
21 <namespace key="10" case="first-letter">Template</namespace>
22 <namespace key="11" case="first-letter">Template talk</namespace>
23 <namespace key="12" case="first-letter">Help</namespace>
24 <namespace key="13" case="first-letter">Help talk</namespace>
25 <namespace key="14" case="first-letter">Category</namespace>
26 <namespace key="15" case="first-letter">Category talk</namespace>
27 <namespace key="90" case="first-letter">Thread</namespace>
28 <namespace key="91" case="first-letter">Thread talk</namespace>
29 <namespace key="92" case="first-letter">Summary</namespace>
30 <namespace key="93" case="first-letter">Summary talk</namespace>
31 <namespace key="100" case="first-letter">Portal</namespace>
32 <namespace key="101" case="first-letter">Portal talk</namespace>
33 <namespace key="102" case="first-letter">Comments</namespace>
34 <namespace key="103" case="first-letter">Comments talk</namespace>
35 <namespace key="828" case="first-letter">Module</namespace>
36 <namespace key="829" case="first-letter">Module talk</namespace>
37 <namespace key="2300" case="first-letter">Gadget</namespace>
38 <namespace key="2301" case="first-letter">Gadget talk</namespace>
39 <namespace key="2302" case="case-sensitive">Gadget definition</namespace>
40 <namespace key="2303" case="case-sensitive">Gadget definition talk</namespace>
41 </namespaces>
42 </siteinfo>

```

- Фрагмент извлеченного текста

```

1 <documents>
2 <doc id="15956" url="https://en.wikinews.org/wiki?curid=15956" title="Four small explosions strike London's transport system">
3 Four small explosions strike London's transport system
4
5 London Metropolitan Police commissioner Sir Ian Blair has confirmed that there have been three small explosions on tube trains at W
6 The London Ambulance Service has not found any injured people, but one person has reported themselves to a local hospital. It is no
7 All the devices were "conventional" but possibly faulty, and contained no chemical or biological agents. Not all the bombs exploded
8 Incidents.
9 A spokesman for Stagecoach said the driver of the number 26 bus travelling through Shoreditch had heard a bang on upper deck, gone
10 One injury sustained at Warren Street tube station has been confirmed by authorities. There have been no other reports of injuries,
11 At around 15:25, a man was arrested by armed police in Whitehall, which is cordoned off. A second man was arrested in the Whitehall
12 Armed police were deployed at UCL Hospital, near Warren Street tube station, after reports of a suspect entering the hospital. Ther
13 Responses.
14 Prime Minister Tony Blair has cancelled a visit to an east London school and a photocall with visiting Prime Minister John Howard c
15 In a public press conference, Blair said that "there appear to have been no casualties", and that he wanted people to "React calmly
16 Police initially advised against unnecessary travel in London, asking Londoners to keep travel to a minimum and avoid the public tr
17 A release from Scotland Yard stated that there was no chemical agents found after checking the Oval. Investigations at Shepherd's B
18 Closures.
19 The Northern line, the Hammersmith and City Line, the Piccadilly Line, and the Bakerloo Line have been suspended. Victoria Line and
20 Unconfirmed reports.
21 Various news sites are reporting a minor explosion in a passenger's backpack. A BBC correspondent, claiming to have sources working
22 A spokesman for London Underground has stated the nature of the incidents is unknown.
23 Eyewitness reports from Warren Street say that something happened towards the front of the train. The passengers all headed towards
24 25% of Shepherd's Bush / Uxbridge Road and all of Shepherd's Bush Green is sealed off.
25 Eyewitness report of "bang" in a carriage at Oval station. No injuries. After being spotted the suspect fled the station, leaving t
26 London Police are not regarding this as a major incident yet. (BBC News 24)
27 British Transport Police report there has been one injury at Warren Street Station. No details as to the cause and nature of this i
28 There are reports of problems sending text messages or making calls from mobiles phones on some networks. (02 confirmed, Orange is
29 Emergency Numbers.
30 "Note: Please don't call these numbers just because you can't get through - some of the mobile networks are temporarily down or dis
31 Sources.
32
33
34 </doc>
35 <doc id="15958" url="https://en.wikinews.org/wiki?curid=15958" title="'Incidents' spark Tube evacuation">
36 'Incidents' spark Tube evacuation
37
38

```

Примеры запросов

- Израиль

Израиль site: <https://www.wikinews.org/>



Все

Новости

Картинки

Видео

Карты

Ещё

Инструменты

Результатов: примерно 154 000 (0,51 сек.)

[https://ru.wikinews.org/wiki/Файл:IMRI_\(Israel\).P...](https://ru.wikinews.org/wiki/Файл:IMRI_(Israel).P...) ▼

Файл:IMRI (Israel). Photo 343.jpg - Викиновостей

4928 × 3264 (8,25 Мб), Alina Voznaya, User created **page** with UploadWizard ...

Использование в [he.wikipedia.org](#) ... Использование в [pt.wikinews.org](#).

https://ru.wikinews.org/wiki/В_Израиле_разработа... ▼

В Израиле разработан метод лечения ВИЧ - Викиновости

6 сент. 2010 г. — По сообщению **израильской** газеты Haaretz, к сентябрю 2010 года в Еврейском университете в Иерусалиме разработан новый метод лечения ...

<https://www.wikinews.org> ▼ [Перевести эту страницу](#)

Wikinews

English · Writing an article 21 000+ articles. 中文 · 新聞投稿 15 000+ 篇. Français · Écrire un article 23 000+ articles. Deutsch · Schreibe einen Artikel

Не найдено: **Израиль** | Запрос должен включать: [Израиль](#)

[https://ru.wikinews.org/wiki/Файл:IMRI_\(Israel\).P...](https://ru.wikinews.org/wiki/Файл:IMRI_(Israel).P...) ▼

Файл:IMRI (Israel). Photo 345.jpg - Викиновости

4928 × 3264 (8,27 Мб), Alina Voznaya, User created **page** with UploadWizard ...

Использование в [nl.wikipedia.org](#) ... Использование в [pt.wikinews.org](#).

https://ru.wikinews.org/wiki/Комментарии:Итоги_... ▼

Комментарии:Итоги израильской агрессии - Викиновости

This is to remind that neither **Wikinews** nor Wikimedia Foundation bear any ... Civil discussion and polite dispute turn comments **page** into a friendly space.

https://ru.wikinews.org/wiki/Файл:Flickr_-_Israel_... ▼

Файл:Flickr - Israel Defense Forces - Climbing Up the Infantry ...

The Israel Defense Forces Facebook **page** The Israel Defense Forces blog ... Эта фотография была сделана сотрудниками Армии обороны **Израиля** и передана в ...

- Facebook

facebook site: <https://www.wikinews.org/>

[Все](#) [Новости](#) [Картинки](#) [Видео](#) [Покупки](#) [Ещё](#) [Инструменты](#)

Результатов: примерно 640 000 (0,52 сек.)

<https://www.facebook.com> > ... [Перевести эту страницу](#)

Wikinews - Home | Facebook

<http://www.wikinews.org/> ... Please help us keep the **Wikinews Facebook page** as open as possible, if an issue is dear-to-your heart — write an article on ...

<https://www.wikinews.org> [Перевести эту страницу](#)

Wikinews

English · Writing an article 21 000+ articles. 中文 · 新聞投稿 15 000+ 篇. Français · Écrire un article 23 000+ articles. Deutsch · Schreibe einen Artikel

Не найдено: facebook | Запрос должен включать: **facebook**

<https://en.wikinews.org> > wiki [Перевести эту страницу](#)

Wikinews:Social media - Wikinews, the free news source

30 авг. 2021 г. — Follow **Wikinews** on your favourite social networking **sites!** ... **Facebook** is a social networking **website** that is operated and privately owned ...

<https://ru.wikinews.org> > wiki > Файл:Screen_of_Faceb... [▼](#)

Файл:Screen of Facebook.PNG - Викиновости

1583 × 852 (88 Кб), EEIM, User created **page** with UploadWizard ... **Facebook**. Использование в bg.wikipedia.org ... Использование в fr.wikinews.org.

<https://en.wikinews.org> > wiki [Перевести эту страницу](#)

Wikinews

5 июн. 2020 г. — In a talk made to officials at a meeting of the Political Bureau of the Workers' Party of Korea held on Wednesday, North Korean leader Kim Jong ...

[Social media](#) · [Wikinews:Audio Wikinews](#) · [Wikinews:Introduction](#) · [Writing an article](#)

<https://ru.wikinews.org> > wiki > Файл:Facebook_Logo... [▼](#)

Файл:Facebook Logo (2019).png - Викиновости

File:**Facebook** Logo (2019).png → File:**Facebook** f logo (2019).svg ... 2143 × 2143 (55 Кб), Smithr32, User created **page** with UploadWizard ...

<https://ru.wikinews.org> > wiki > Russian_Wikinews_ove... [▼](#)

Russian Wikinews overtake other Wikinews - Викиновости

14 сент. 2020 г. — Main "hub" **page** of **Wikinews** at www.wikinews.org at September 14, 2020 ... 2020, Russian **Wikinews** took the 1st string of the title **page** of ...

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

Выводы

В ходе выполнения лабораторной работы был получен корпус документов для выполнения следующих лабораторных работ. Сырые и подготовленные данные хранятся в формате xml, для генерации корпуса документов использовалась сторонняя библиотека, дабы оптимизировать процесс.