# Московский авиационный институт (национальный исследовательский университет)

## Факультет информационных технологий и прикладной математики

## Кафедра вычислительной математики и программирования


**Лабораторная работа №1 по курсу «Информационный поиск»**


Студент: И. А. Мариничев
Преподаватель: А. А. Кухтичев
Группа: М8О-408Б-19

Дата:
Оценка:
Подпись:


**Москва, 2023**

# Лабораторная работа №1 «Добыча корпуса документов»

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 мета-информация? 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.

В результатах работы должна быть указаны статистическая информация о корпусе:

- Размер «сырых» данных.
- Количество документов.
- Размер текста, выделенного из «сырых» данных.
- Средний размер документа, средний объём текста в документе.

# 1 Описание

В качестве корпуса докуметов был выбран дамп английской википедии. Из этого дампа было выбрано 50000 документов, которые будут ипользоваться в дальнейшем. Каждый документ имеет следующее строение:

<doc id=«<Id>» url=«https://en.wikipedia.org/wiki?curid=<Id>» title=«<Title>»>

<Title>

<Contents>

</doc>

Где <Id> - это идентификатор документа в википедии, <Title> - это название документа, <Contents>- это содержимое документа.

# 2 Исходный код

Для того, чтобы получить корпус нужного размера, был использован следующий скрипт на языке Python:

```
FILE_NAME ='corpus/wikipedia'
PREVIEW_SIZE =1557199
BUFFER =list()
TOTAL =0

with open(FILE_NAME) as f:
    for _ in range(PREVIEW_SIZE):
        BUFFER.append(f.readline())

with open(f'{FILE_NAME}_preview', 'w') as f:
    for i in range(PREVIEW_SIZE):
        if (BUFFER[i] =="</doc>\n"):
            TOTAL +=1
        f.write(BUFFER[i])

print(TOTAL)
```

# 3 Выводы

Выполнив первую лабораторную работу по курсу «Информационный поиск», я осознал понятие документа и корпуса документов, лучше понял строение википедии и узнал о том, какие метаданные содержаться на ее страницах.

# Список литературы

[1] Маннинг, Рагхаван, Шютце *Введение в информационный поиск* — Издательский дом «Вильямс», 2011. Перевод с английского: доктор физ.-мат. наук Д. А. Клюшина — 528 с. (ISBN 978-5-8459-1623-4 (рус.))