МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»

ИФТЭБ

ИНСТИТУТ ФИНАНСОВЫХ ТЕХНОЛОГИЙ И ЭКОНОМИЧЕСКОЙ БЕЗОПАСНОСТИ

КАФЕДРА «ФИНАНСОВЫЙ МОНИТОРИНГ»

КУРСОВАЯ РАБОТА ПО ДИСЦИПЛИНЕ

«Информационные ресурсы в финансовом мониторинге»

на тему

«Интеллектуальная обработка текста Конституции Российской Федерации с целью выявления ключевых слов с помощью закона Ципфа»

Выполнил студент группы С19-712:
А.И. Ахремова

Проверил:
В.Ю. Радыгин

Москва, 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 3
ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ. ЗАКОН ЦИПФА 4
ГЛАВА 2. ПРАКТИЧЕСКАЯ ЧАСТЬ. ВЫПОЛНЕНИЕ РАБОТЫ 6
ЗАКЛЮЧЕНИЕ 11
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 12
ПРИЛОЖЕНИЕ А 13

ВВЕДЕНИЕ

В современном мире в период глобализации и модернизации все больше людей обращаются к сети Интернет. С каждым днем информации в мире становится все больше и больше, ее количество увеличивается в геометрической прогрессии [1]. Для своевременной и качественной обработки такого огромного объема информации разработаны специальные методы и способы работы с текстовыми форматами данных.

Для решения задач обработки текстов применяются различные методы, такие как морфологический анализ, классификация, ранжирование и другие. Их важность в современном информационном обществе заключается в том, что они предоставляют средства для автоматической обработки и анализа текстов, с помощью которых можно своевременно получать необходимые сведения из больших объемов данных.

Целью данной курсовой работы является изучение основ работы с различной текстовой информацией в сети Интернет. Данная работа выполняется на примере текста Конституции Российской Федерации на русском языке.

Для полного и тщательного исследования и анализа текста необходимо выделить ключевые этапы, требующие особенно важного внимания. Для этого были определены и сформулированы основные задачи, представляющие собой план, по которому в дальнейшем была проделана работа:

1. Загрузка Конституции Российской Федерации на русском языке с указанного сайта;
2. Преобразование HTML-страницы в текст;
3. Преобразование всех слов в нормальную форму;
4. Удаление в тексте всех стоп-слов русского языка;
5. Составление таблицы частот встречающихся слов;
6. Построение графика для закона Ципфа и выделение ключевых слов.

ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ. ЗАКОН ЦИПФА

Закон Ципфа является статистическим законом, который описывает распределение частоты употребления слов в тексте на естественном языке. Закон был назван в честь американского лингвиста Джорджа Ципфа.

Суть закона Ципфа заключается в том, что частота появления слова в тексте обратно пропорциональна его порядковому номеру. Другими словами, наиболее часто употребляемые слова в тексте имеют меньший порядковый номер, а редко употребляемые слова имеют больший порядковый номер. Например, если наиболее часто употребляемое слово в тексте имеет порядковый номер 1, то второе по частоте слово будет иметь порядковый номер, равный половине частоты первого слова, третье слово - одной третьей, четвертое слово - одной четвертой и так далее [2].

Закон Ципфа применяется в лингвистике, статистике, информатике и других науках для анализа текстов и выделения наиболее значимых слов. Он также может использоваться для определения наиболее важных ключевых слов в поисковой оптимизации.

Чтобы понимать, как работает данный закон на практике, была выведена специальная формула, по которой можно вычислить закономерность использования определенных слов:

$$C = F * R \quad (1)$$

где:

- C – константа,
- F – частота появления слова в тексте,
- R – ранг слова.

Ранг слова определяется путем упорядочивания списка слов в порядке убывания частоты их употребления, где наиболее часто употребляемое слово занимает первое место, следующее за ним – второе, и так далее.

На рисунке 1 представлена теоретическая зависимость закона Ципфа. Графическое изображение закона напоминает гиперболу.

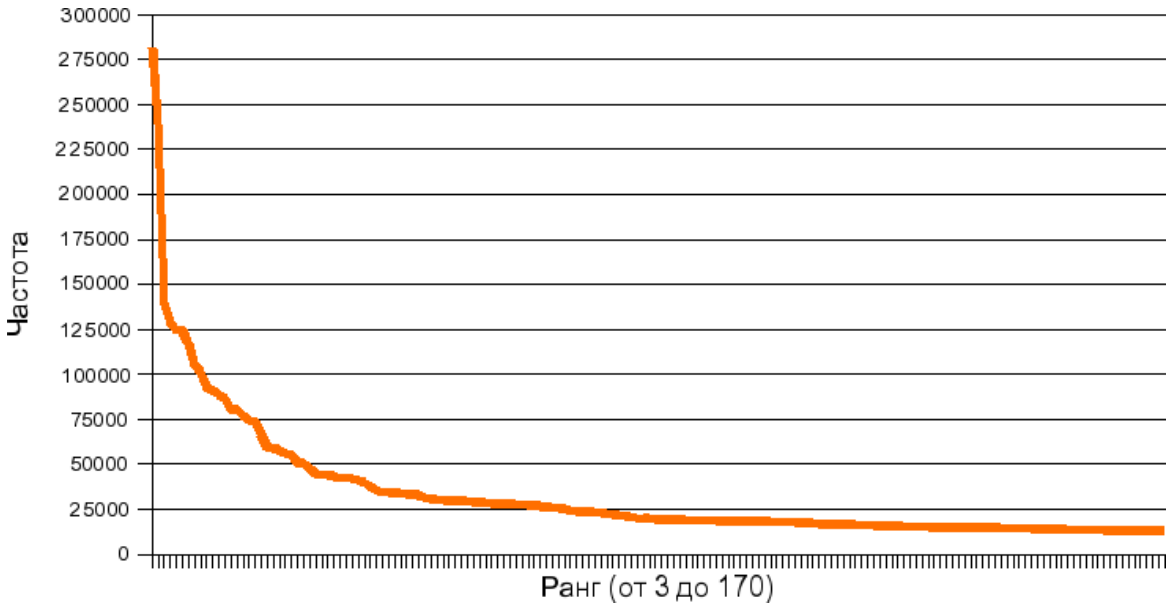

Рисунок 1 – Теоретическая зависимость

ГЛАВА 2. ПРАКТИЧЕСКАЯ ЧАСТЬ. ВЫПОЛНЕНИЕ РАБОТЫ

Для интеллектуальной обработки текста Конституции Российской Федерации была написана программа на языке программирования python, позволяющая получить текст с указанного сайта, провести процесс нормализации слов, удалить из текста стоп-слова русского языка, затем создать словарь частотности слов и на его основе построить график, подтверждающий выполнение закона Ципфа, и выделить ключевые слова в тексте документа.

Для начала необходимо импортировать следующие библиотеки:

- pandas – библиотека для обработки и анализа структурированных данных;
- requests – библиотека для работы с http-запросами;
- nltk – библиотека для символьной и статистической обработки естественного языка;
- re – модуль, предоставляющий операции сопоставления шаблонов регулярных выражений;
- pymorphy2 – морфологический анализатор, выполняющий лемматизацию и морфологический анализ слов;
- bs4 – библиотека для извлечения данных из файлов HTML и XML;
- plotly – библиотека для создания визуализаций (графиков);
- alive_progress – библиотека для визуализации процесса обработки.

На рисунке 2 представлен импорт необходимых библиотек.

```
import pandas as pd
import requests
import nltk
from nltk.probability import FreqDist
from nltk.corpus import stopwords
import re
import pymorphy2
from bs4 import BeautifulSoup, SoupStrainer
import plotly.express as px
from alive_progress import alive_bar
```

Рисунок 2 – Импорт библиотек

С помощью запроса с сайта http://Duma.gov.ru/news/48953 [3] была получена Конституция Российской Федерации на русском языке в формате HTML-документа. Для этого был выделен текст из HTML-элементов с атрибутом class = «article__content». На рисунке 3 приведен фрагмент кода, на рисунке 4 показан результат выполнения.

```
response = requests.get('http://Duma.gov.ru/news/48953')
data = response.text

soup = BeautifulSoup(response.content, 'html.parser', parse_only=SoupStrainer('div', class_='article__content'))
print(soup)
```

Рисунок 3 – Загрузка конституции

```
    3. Особенности осуществления публичной власти на территориях городов федерального значения, административных центров (столиц)
 </strong><strong class="highlight-important">
<br/>
</strong></p><p><b>Статья 132</b></p><p>1. Органы местного самоуправления самостоятельно управляют муниципальной собственностью, ф
    вводят
 </strong> местные налоги и сборы, решают иные вопросы местного значения, <strong class="highlight-important">
    а также в соответствии с федеральным законом обеспечивают в пределах своей компетенции доступность медицинской помощи.
 </strong></p><p>2. Органы местного самоуправления могут наделяться <strong class="highlight-important">
    федеральным
 </strong> законом, <strong class="highlight-important">
    законом субъекта Российской Федерации
 </strong>отдельными государственными полномочиями <strong class="highlight-important">
    при условии передачи
 </strong> необходимых для осуществления <strong class="highlight-important">
    таких полномочий
 </strong> материальных и финансовых средств. Реализация переданных полномочий подконтрольна государству.</p><p><strong class="high
    3. Органы местного самоуправления и органы государственной власти входят в единую систему публичной власти в Российской Федера
 </strong></p><p><b>Статья 133</b></p><p>Местное самоуправление в Российской Федерации гарантируется правом на судебную защиту, на
    в результате выполнения органами местного самоуправления во взаимодействии с органами государственной власти публичных функци
 </strong> запретом на ограничение прав местного самоуправления, установленных Конституцией Российской Федерации и федеральными зак
</div>
115624
```

Рисунок 4 – Результат выполнения загрузки

Далее этот документ был преобразован в простой текст (рис. 5). Результат выполнения показан на рисунке 6.

```
text = soup.get_text()
print(len(text))
```

Рисунок 5 – Преобразование в текст

```
    Статья 75.1

    В Российской Федерации создаются условия для устойчивого экономического роста страны и повышения бла
Статья 761. По предметам ведения Российской Федерации принимаются федеральные конституционные законы и ф
    3. Высшим должностным лицом субъекта Российской Федерации (руководителем высшего исполнительного орг



Статья 781. Федеральные органы исполнительной власти для осуществления своих полномочий могут создавать
    5. Руководителем федерального государственного органа может быть гражданин Российской Федерации, дос
Статья 79Российская Федерация может участвовать в межгосударственных объединениях и передавать им часть
    Российской Федерации
, если это не влечет ограничения прав и свобод человека и гражданина и не противоречит основам конституц
    Решения межгосударственных органов, принятые на основании положений международных договоров Российск
Статья 79.1
    Российская Федерация принимает меры по поддержанию и укреплению международного мира и безопасности,
```

Рисунок 6 – Фрагмент результата

Затем текст был разделен на слова и из него были удалены нетекстовые символы (рис. 7).

```
clear_text = re.sub(r'[^а-я\s\-]', '', text.lower())
clear_text = re.sub(r'\s', ' ', clear_text)
clear_text = re.sub(' +', ' ', clear_text)
```

Рисунок 7 – Удаление нетекстовых символов

После был проведен процесс нормализации (лемматизации) полученных слов (рис. 8).

```
morph = pymorphy2.MorphAnalyzer(lang='ru')
words_norm = []
words = clear_text.split()
bar_max = len(words)

with alive_bar(bar_max, force_tty=True, length=30) as bar:
    for word in words:
        p = morph.parse(word)[0]
        words_norm.append(p.normal_form)
        bar()
```

Рисунок 8 – Нормализация слов

На следующем этапе из полученного набора слов были удалены стоп-слова (рис. 9). Как правило, стоп-словами являются местоимения, частицы и некоторые общеупотребительные глаголы, т. е. слова не несущие смысловую нагрузку.

```
nltk.download('stopwords')
stop_words = set(stopwords.words('russian'))
words_no_stops = [word for word in words_norm if not word in stop_words]
```

Рисунок 9 – Удаление стоп-слов

После удаления из текста стоп-слов, получен «чистый» текст. Теперь можно оценить его мощность и применить закон Ципфа. Далее был получен словарь частотности встречающихся слов (рис. 10).

```
freq_dist = FreqDist(words_no_stops)
freq_dist_df = pd.DataFrame.from_dict(freq_dist, orient='index')

freq_dist_df.sort_values(by=[0], ascending=False, inplace=True)
freq_dist_df.reset_index(inplace=True)
freq_dist_df.columns = ['Слово', 'Частота']
```

Рисунок 10 – Получение словаря частотности слов в тексте

Расположим слова в порядке убывания их частотности, предварительно выделив 10 наиболее популярных в тексте слов (рис. 11). В результате получим список кортежей по убыванию частотности (рис. 12).

```
print('10 самых частых слов')
freqwords = freq_dist_df.head(10)
print(freqwords)
```

Рисунок 11 – Выделение 10 наиболее встречающихся слов

```
10 самых частых слов
             Слово  Частота
0       российский      741
1        федерация      739
2      федеральный      248
3  государственный      227
4            закон      173
5        президент      141
6            орган      122
7            право      120
8             дума      104
9      конституция       99
```

Рисунок 12 – Результат выполнения

Построим график зависимости наиболее встречающихся слов от ранга (рис. 13). Полученный график представлен на рисунке 14.

```
fig = px.line(freq_dist_df, title="Закон Ципфа", x=freq_dist_df.index, y="Частота", width=1400, height=788)
fig.update_xaxes(title_text='Ранг')
fig.update_yaxes(title_text='Частота')
fig.show()
```

Рисунок 13 – Построение графика зависимости частотности слов от их ранга

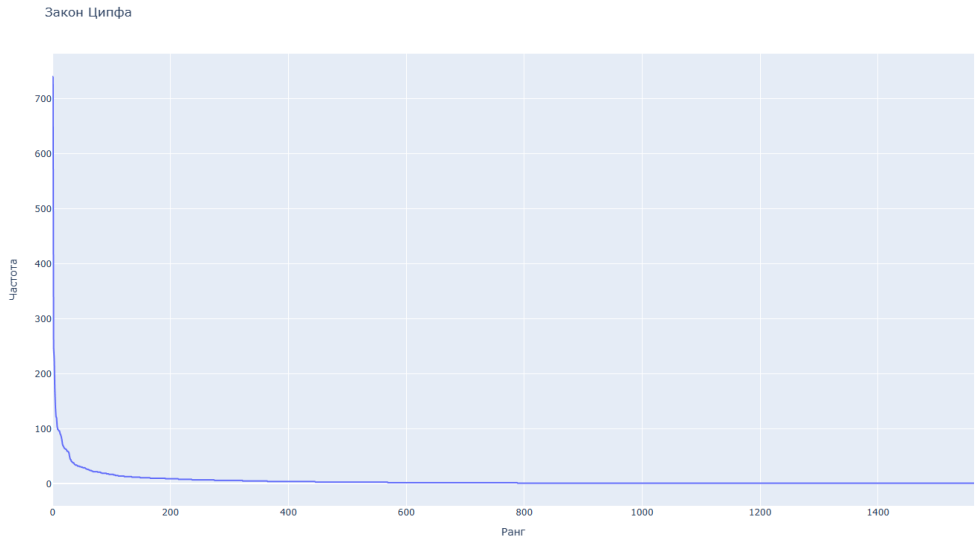

Рисунок 14 – Полученный график зависимости частотности слов от их ранга

Можно увидеть, что характер зависимости, полученной в ходе лабораторной работы, совпадает с характером теоретической зависимости, приведенной в теоретической части отчета.

Затем выделим ключевые слова (рис. 15). Результат представлен на рисунке 16.

```
min_rang = 9
max_rang = 19
print("Ключевые слова")
keywords = freq_dist_df.iloc[min_rang:max_rang]
print(keywords)
```

Рисунок 15 – Выделение ключевых слов

```
Ключевые слова
             Слово  Частота
9       конституция       99
10          власть       98
11  конституционный       96
12    правительство       96
13            мочь       91
14             суд       89
15     председатель       85
16        гражданин       80
17          область       72
18        должность       69
```

Рисунок 16 – Ключевые слова

Полный код программы, написанной на python, представлен в Приложении А.

ЗАКЛЮЧЕНИЕ

В ходе выполнения данной курсовой работы были изучены основы работы с различной текстовой информацией в сети Интернет. На примере Конституции Российской Федерации был проведен анализ текста, выделены ключевые слова, и построен график зависимости частотности слов от ранга – то есть была проведена визуализация закона Ципфа. Был составлен словарь частотности слов. Полученная практическая зависимость примерно совпадает с теоретической. Можно утверждать, что закон Ципфа работает для больших текстовых документов.

Словарь частотности слов, полученный в процессе выполнения работы, подробно характеризует сущность обработанного текста. Двумя наиболее часто встречающимися словами являются «российский» и «федерация» с частотой 741 и 739 соответственно. Наиболее важные ключевые слова в тексте: «власть» с частотой 98, «правительство» с частотой 96, «гражданин» с частотой 80. Полученные результаты свидетельствуют о необходимости подготовки, обработки и структурирования текстового источника информации для повышения качества дальнейшей проводимой работы.

Говоря про работу с текстовыми источниками информации, необходимо заметить, что одним из важных аспектов работы с текстами является их структурирование и организация с помощью различных инструментов. Именно грамотное применение таких инструментов позволяет сократить время на выполнение работы и помогает получить более качественные результаты. Навыки работы с текстовой информацией являются важным фактором успеха во многих профессиональных областях.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Что такое «Big Data»? [Электронный ресурс]. – URL: https://vc.ru/s/productstar/129351-chto-takoe-big-data – (дата обращения: 06.05.2023).

2. Закон Ципфа-Мандельброта [Электронный ресурс]. – URL: https://vc.ru/education/370718-zakon-cipfa-mandelbrota-zakon-rang-chastotnost – (дата обращения: 06.05.2023).

3. Конституция Российской Федерации [Электронный ресурс]. – URL: http://Duma.gov.ru/news/48953 – (дата обращения: 06.05.2023).

ПРИЛОЖЕНИЕ А

```
import pandas as pd
import requests
import nltk
from nltk.probability import FreqDist
from nltk.corpus import stopwords
import re
import pymorphy2
from bs4 import BeautifulSoup, SoupStrainer
import plotly.express as px
from alive_progress import alive_bar

response = requests.get('http://Duma.gov.ru/news/48953')
data = response.text

soup = BeautifulSoup(response.content, 'html.parser',
parse_only=SoupStrainer('div', class_='article__content'))
print(soup)
text = soup.get_text()
print(text)
print(len(text))

clear_text = re.sub(r'[^а-я\s\-]', '', text.lower())
clear_text = re.sub(r'\s', ' ', clear_text)
clear_text = re.sub(' +', ' ', clear_text)

morph = pymorphy2.MorphAnalyzer(lang='ru')
words_norm = []
words = clear_text.split()
bar_max = len(words)

with alive_bar(bar_max, force_tty=True, length=30) as bar:
    for word in words:
        p = morph.parse(word)[0]
        words_norm.append(p.normal_form)
        bar()

nltk.download('stopwords')
stop_words = set(stopwords.words('russian'))
words_no_stops = [word for word in words_norm if not word in stop_words]

freq_dist = FreqDist(words_no_stops)
freq_dist_df = pd.DataFrame.from_dict(freq_dist, orient='index')

freq_dist_df.sort_values(by=[0], ascending=False, inplace=True)
freq_dist_df.reset_index(inplace=True)
freq_dist_df.columns = ['Слово', 'Частота']

print('10 самых частых слов')
freqwords = freq_dist_df.head(10)
print(freqwords)

fig = px.line(freq_dist_df, title="Закон Ципфа", x=freq_dist_df.index,
y="Частота", width=1400, height=788)
fig.update_xaxes(title_text='Ранг')
fig.update_yaxes(title_text='Частота')
fig.show()

min_rang = 9
max_rang = 19
print("Ключевые слова")
keywords = freq_dist_df.iloc[min_rang:max_rang]
print(keywords)
```