МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет информационных технологий и управления

Кафедра информационных технологий автоматизированных систем

**ОТЧЕТ**

По лабораторной работе №2

«Технология анализа текста и извлечения ключевых слов»

|  |  |  |
| --- | --- | --- |
| Выполнил |  | Проверил: |
| ст. гр. 328501 |  | Ярмолик В. И. |
| Суворов В. В. |  |  |

Минск 2023

**1 ЦЕЛЬ РАБОТЫ**

Практическое освоение технологии анализа текста, извлечения ключевых слов и профессионального поиска информации.

**2 ОСНОВНОЙ ПРИНЦИП ВЫБОРА КЛЮЧЕВЫХ СЛОВ**

Основной алгоритм поиска ключевых слов основан на Законе Ципфа:

Джордж Ципфа установил, что все тексты подчиняются общим закономерностям, и сформулировал в 1946—49 гг. несколько законов, которые нашли применение в технологии поиска информации.

Для ознакомления с положениями первого закона Ципфа введем, следуя Ципфу, необходимые терминологические определения. Рассмотрим некоторый произвольный текст. Выпишем все различающиеся слова данного текста в виде где -e есть слово, не совпадающее ни с каким другим словом в данном множестве. Для каждого из этих слов подсчитаем количество его повторов в тексте.  В результате получим где - количество повторений *i*-го слова в тексте, названное Ципфом частотой слова (в данном случае *i*-го слова).

Далее Ципф, располагая слова в порядке убывания их частот, поставил им в соответствие числа натурального ряда, назвав эти числа рангами слов (*R*):слову с максимальной частотой присваивается ранг 1, следующему по частоте — ранг 2 и т.д.  При этом если несколько разных слов имеют одинаковые частоты, то они объединяются в один блок. Наконец Ципф ввел понятие вероятности встречи слова как отношение частоты слова к общему количеству слов в тексте.

Первый закон Ципфа утверждает, что произведение частоты встречи слова в тексте (или вероятности встречи слова по Ципфу) на его ранг есть величина приблизительно постоянная для любых текстов определенного языка, т.е. имеет место:

Второй закон Ципфа схож с 1 и утверждает, что что произведение частоты встречи слова в тексте (или вероятности встречи слова по Ципфу) на их количество (A) в тексте есть величина приблизительно постоянная для любых текстов определенного языка, т.е. имеет место:

На рисунке 1 приведена зависимость частоты слов *f* от его ранга *R*. Кривая представляет собой гиперболу:

*f = C / R*.

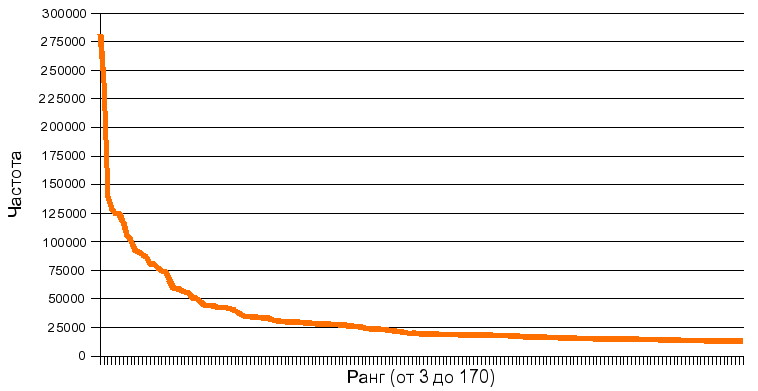


Рисунок 1 – График зависимости частоты слова *f* от его ранга *R*.

Характерно, что все значимые (ключевые) для текста слова размещаются в средней части графика. При этом часто встречающиеся слова (ранг от 1 до 5), как правило, являются вспомогательными (это предлоги, частицы, местоимения и т.д.). Такие слова называются «стоп-словами»; поисковая машина при простом поиске их по умолчанию игнорируют, так как эти слова являются "шумом", помехой, которая затрудняет поиск. Редко встречающиеся слова (правая часть графика) также не имеют существенного значения при анализе текста.

Важно уметь правильно выбрать диапазон значений ранга для извлечения ключевых слов. Если он будет слишком широким, то значимые слова затеряются среди второстепенных. Если же установить очень узкий диапазон, то возможна потеря существенных ключевых слов.

Законы Ципфа используются при создании на поисковых серверах базы данных, в которой хранится индексированная информация; при этом учитывается целый ряд факторов, таких как вес слова, его местоположение в документе, морфологические особенности и др. Они же используются и для оценки релевантности (степени соответствия) документов в процессе поиска.   
Релевантность изменяется от 0 до 1, в зависимости от того, какое количество слов поискового выражения содержится в найденном документе, а также веса каждого из таких слов.

Уточненные законы Ципфа используются также в алгоритмах автоматического распознавания текста программ-экстракторов, которые осуществляют семантический анализ текстов и извлекают ключевые слова и выражения.

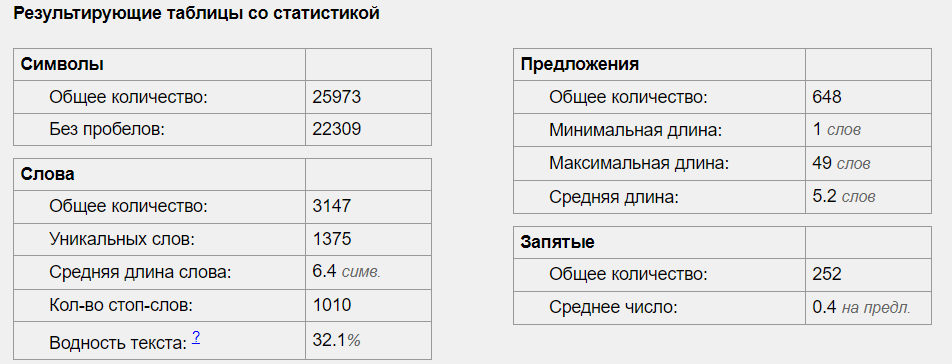
 **3 ПРОГРАММА-ЭКСТРАКТОР** Для анализа текста используется онлайн экстрактор *https://textanalyzer.pro*  
Принцип работы его очень прост: он собирает всю информацию о содержании данного текста и выводит в таблицах разностороннюю статистику.

Рисунок 2.1 – Основная статистика текста.

На рисунке 2.1 можно заметить такие показатели, как:  
количество слов, количество уникальных слов, количество стоп слов, процент «водности» текста и т.п.

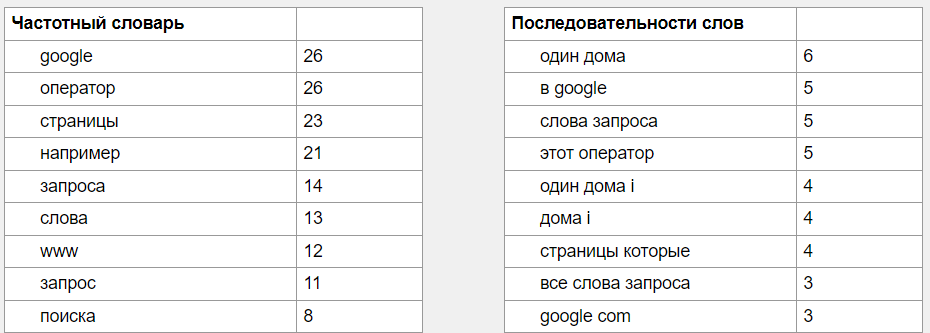
Далее следует более важная составляющая анализатора текста

Рисунок 2.2 – Ключевые слова и фразы в тексте.

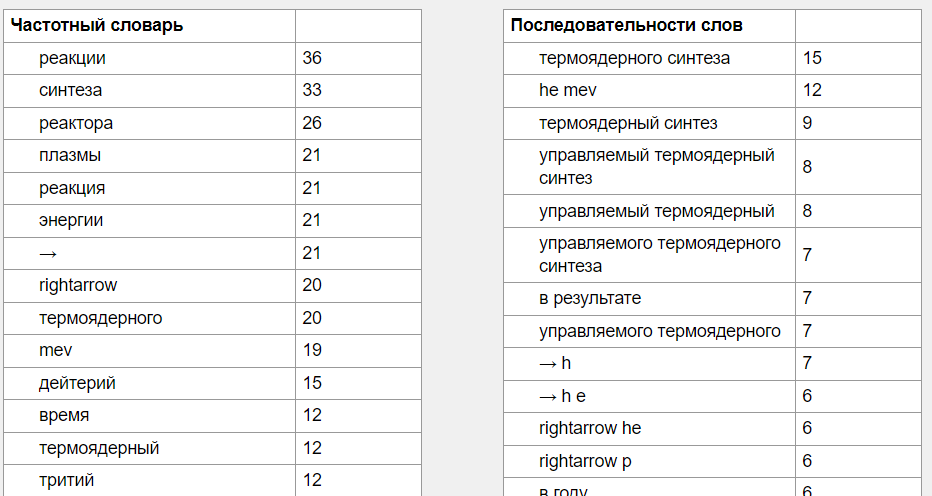
На рисунке 2.2 представлена таблица, где зафиксированы некоторые слова и фразы, а также их количество в порядке убывания.

Исходя из этих сведений можно определить, на каких словах/словосочетаниях делается акцент в тексте, что является основной целью в задаче экстрактора текста.

**4 ПОИСК ПО КЛЮЧЕВЫМ СЛОВАМ**

Чтобы определить, можно ли найти информацию, по ключевым словам, возьмем статью из Википедии и литературное произведение Война и Мир. Успехом будем считать, что по набору ключевых слов можно будет выйти на этот материал в интернете.

4.1 Статья Википедии про управляемый термоядерный синтез

 Введем текст статьи *https://ru.wikipedia.org/wiki/Управляемый\_термоядерный\_синтез* в экстрактор и получим такие сведения:  
  
Рисунок 3.1 – Ключевые слова и фразы статьи из Википедии.

Введем в поисковик *Google* наиболее частые слова.

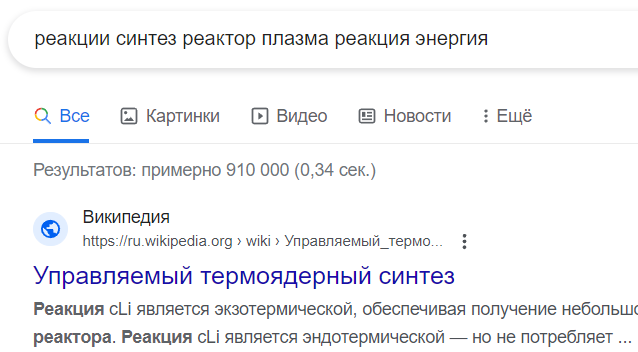


Рисунок 3.2 – Результат поиска по ключевым словам 1-го текста.

Как видно из рисунка 3.2, поиск дал первым же результатом искомую статью.

4.2 Роман Война и Мир

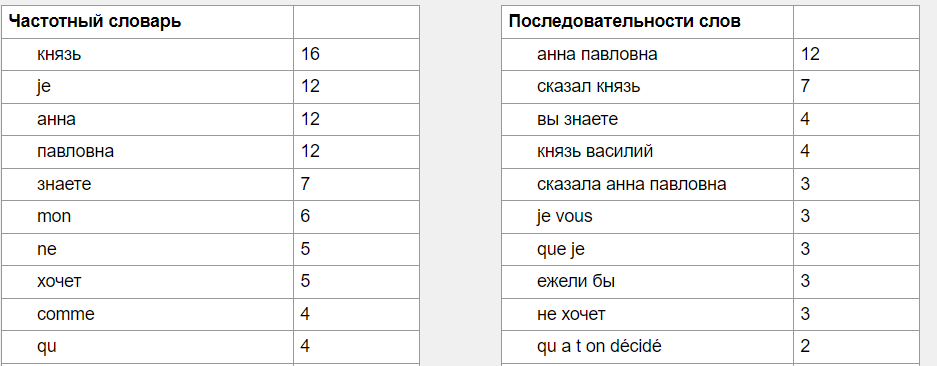
 Аналогично п 4.1 проведем анализ литературного произведения Война и Мир:

Рисунок 3.3 – Ключевые слова и фразы из отрывка 1 главы книги.

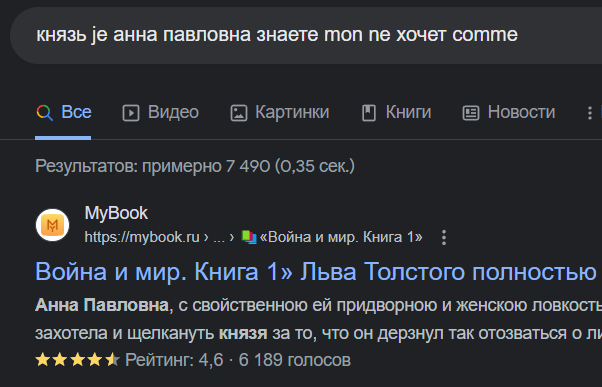
 Введем наиболее частые слова в поисковик.

Рисунок 3.4 – Результат поиска по ключевым словам 2-го текста.

Из рисунка 3.4 можно понять, что ключевые слова дали нам ожидаемый результат.

**5 ВЫВОД**

Проведя исследование по теории Ципфа, используя выведенные им закономерности, теорию, а также проведя несколько практических исследований, основанных на использовании специальных интернет-сайтов для концентрирования информации, ее анализа и организации данных. Можно сделать выводы о полезности метода Ципфа.

В нашем случае, поиск по ключевым словам/фразам, используя метод Ципфа, помогает выделить основные точки опоры текста, определить, на каких понятиях/терминах построен текст, ознакомится с ними для понимая сути текста, а также для поиска более конкретной или дополняющей основной текст информации.