МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

Лабораторные работы
по курсу «Информационный поиск»

Информационный поиск и обработка естественно-языковых текстов

Выполнил: И.В. Сектименко
Группа: М8О-410Б-22
Преподаватели: А.А. Кухтичев

Москва, 2025

Условие

Цель работы: познакомиться с существующими поисковыми системами, найти их преимущества и недостатки, после чего разработать свою поисковую систему.

Задачи:

— найти документы, которые будут использоваться в работе;

— подготовить документы к дальнейшему использованию: выделить текст и заголовок, убрать ненужную информацию;

— привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче;

— реализовать токенизацию и стемминг документов;

— реализовать булевый поиск.

Описание данных

В качестве источника данных был выбран набор данных с сайта litres.ru, а именно: заголовком является ссылка на книгу, тексом – название книги, ее описание и комментарии пользователей к книге. Так как сайт адаптирован для удобства пользователей, на странице видны только часть комментариев и часть описания, поэтому через код необходимо взаимодействовать с логической частью сайта, что занимает достаточно много времени при выгрузке данных. Была взята часть книг по теме легкое чтение. Для получения был написан небольшой скрипт на python.

Вся информация, полученная с сайта, была сохранена в текстовый файл. Данные содержат следующие характеристики:

— 31683 документа;

— 1062 слова в документе.

Существующие поисковики

У сайта litres.ru есть свой поисковик, который ищет по названию книги и его автору. Поисковик google позволяет искать информацию по конкретному сайту с помощью ключевого слова: site:<сайт> <текст запроса>.

Примеры запросов приведены на рисунках 1 и 2.

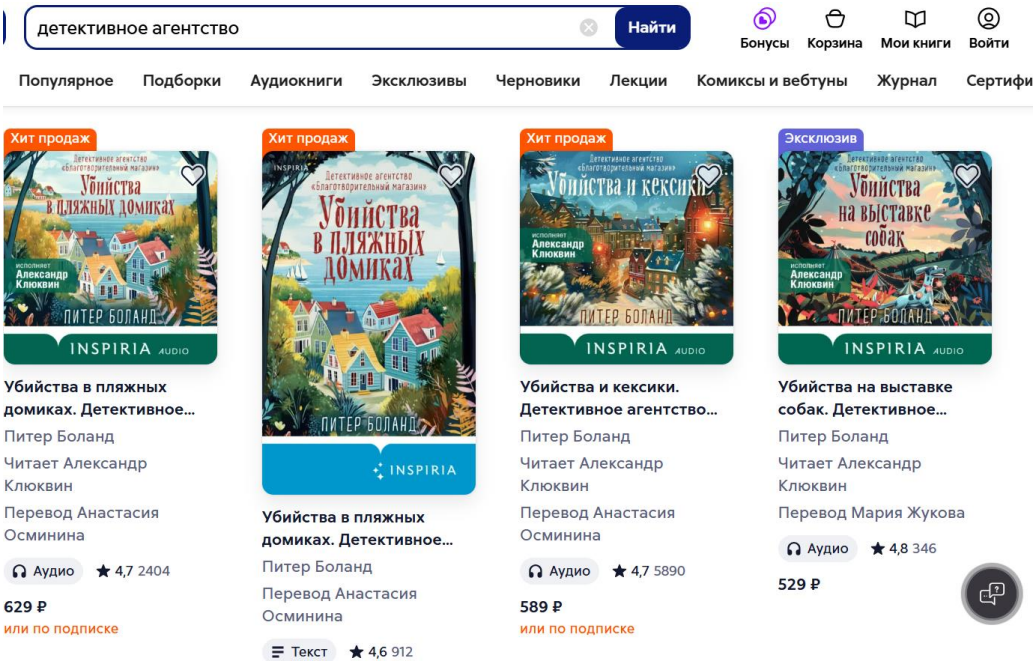

Рисунок 1 – Поиск на сайте litres.ru

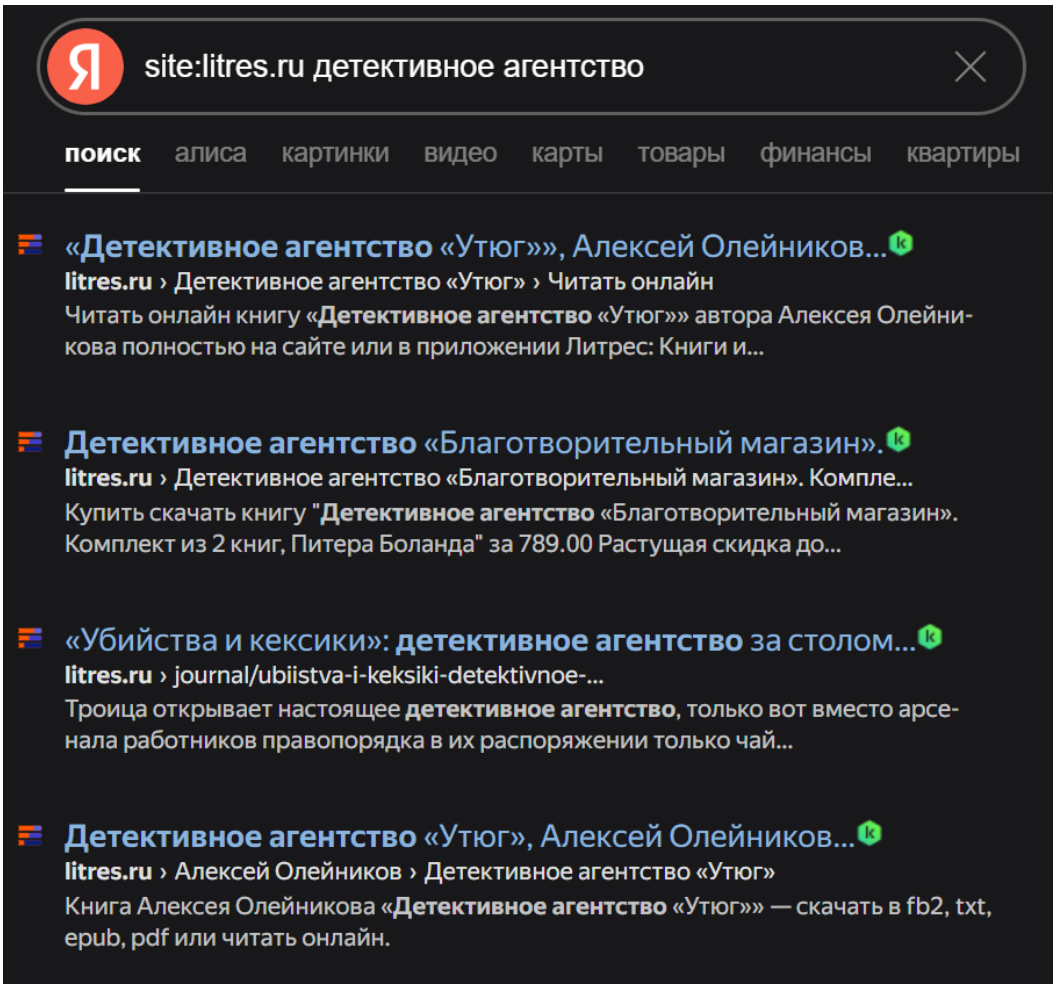

Рисунок 2 – Поиск google по сайту litres.ru

Главные недостатки:

— ограниченность поиска у litres.ru;

— сложность составления запросов: далеко не каждый пользователь догадается с специальной структуре запроса в поисковике google;

— игнорирование потребностей пользователей: не всегда поиск по тексту книги может привести к нужным результатам и помочь найти именно ключевой, поворотный момент сюжета.

Токенизация и стемминг

Для поиска по тексту необходимо произвести токенизацию: разбить текст на отдельные слова без учета знаков препинания. Так можно сократить размер документа и хранить число для каждого токена, которое будет соответствовать количеству его упоминаний в документе.

Для простой обработки естественного языка произведем стемминг: будем хранить основы токенов, то есть обрежем окончания. Так при запросе в разных формах смогут выдаваться одни и те же результаты, ведь при изменении формы слова его суть и смысл не меняется.

Токенизация и стемминг документов были выполнены за 101,256 с.

На рисунке 3 показан график закона Ципфа.

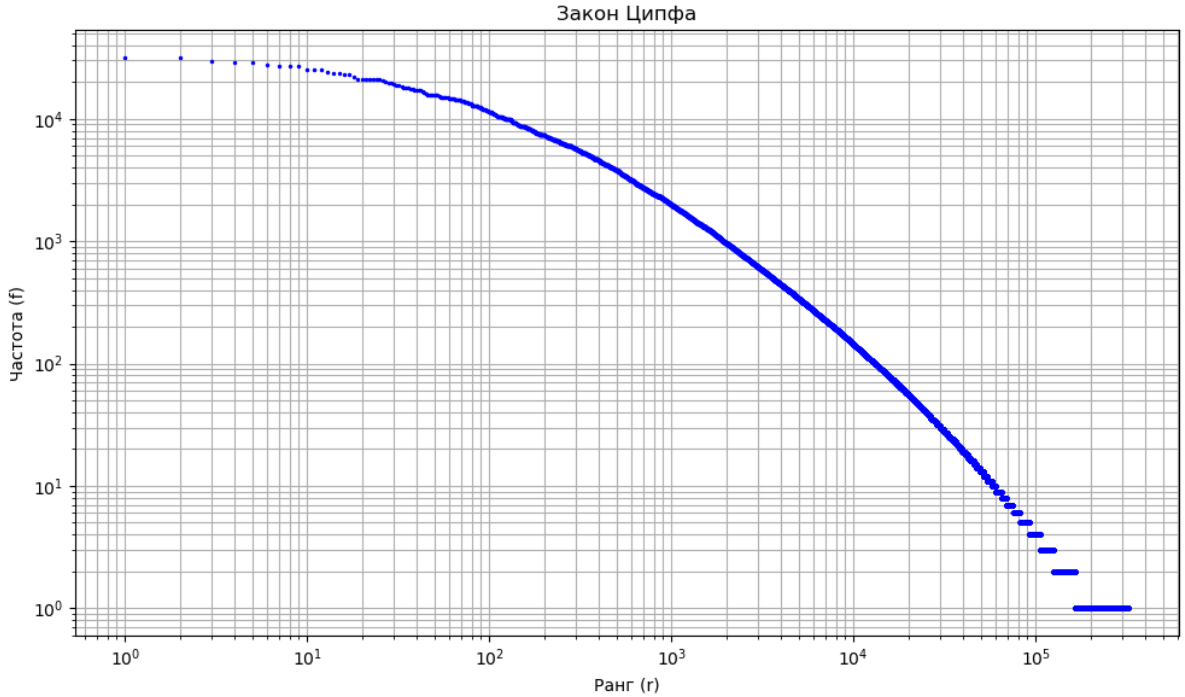

Рисунок 3 – Закон Ципфа

Булевый поиск

Для булевого поиска необходимо сначала спарсить запрос: обработать булевые операторы && (и), || (или), ! (не). В отдельные массивы сохраняется что должно быть в запросе, а что нет. Далее для каждого полученного токена применяется функция стемминга, которая применялась при токенизации документов. После чего формируется фильтр и отправляется на выполнения в базу данных.

В результате пользователь получает список ссылок на документы, соответствующие его запросу.

Поиск выполняется за 8 мс.

Выводы

В ходе выполнения работы реализации алгоритмов поиска информации и обработки естественно-языковых текстов были изучены основные принципы работы поисковых систем: токенизация и парсинг – и способы обработки текстов: стемминг.