МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский Авиационный Институт»

(Национальный Исследовательский Университет)

Институт: №8 «Информационные технологии и прикладная математика»
Кафедра: 806 «Вычислительная математика и программирование»

Отчет по лабораторным работам
по предмету «Информационный поиск»

Группа: М8О-412Б-22

Студент(ка): Кайдалова А. А.

Оценка:

Дата сдачи:

Москва, 2025

1. Цель работы

Целью данной работы является разработка прототипа поисковой системы, ориентированной на тематический корпус документов. В рамках проекта реализованы следующие ключевые компоненты:

- автоматический сбор корпуса статей с заданных веб-сайтов;
- предварительная обработка текста (токенизация и стемминг);
- построение булева индекса;
- реализация булева поиска с поддержкой логических операций;
- анализ распределения терминов в корпусе на основе закона Ципфа;
- обеспечение автоматического запуска и быстрой загрузки системы.

2. Сбор корпуса документов

В качестве источника информации были выбраны три русскоязычных сайта, посвящённых вопросам материнства и воспитания детей: **7ya.ru**, **mama.ru**, **letidor.ru**. Эти ресурсы содержат большое количество структурированных текстов, что делает их пригодными для автоматического анализа.

Чтобы удостовериться, что выбранные источники подходят для лабораторных, в поисковой строке Google я ввела запрос: «первый прикорм схема site: letidor.ru OR site:7ya.ru OR site:mama.ru». Google выдал страницы, соответствующие запросу. Значит, мой корпус документов уже проиндексирован и доступен для поиска через мощную существующую систему.

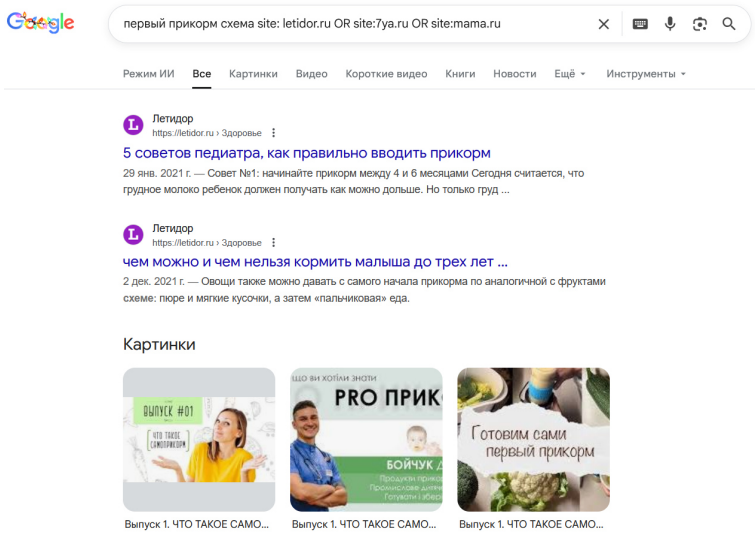

Рисунок 1. Поиск в Google

Недостатки существующих поисковиков:

- Поисковые системы сложны в использовании для тех, кто не владеет навыками составления запросов;
- Чаще показывается то, что популярно или хорошо раскручено, а не то, что лучше всего отвечает на ваш конкретный вопрос;

Для сбора данных был разработан поисковой робот, выполняющий следующие функции:

1. **Обход целевых доменов**. Обход осуществляется двумя способами:
   - Рекурсивный обход внутренних ссылок с начальных URL-адресов;
   - Парсинг sitemap.xml, с автоматическим извлечением всех вложенных карт сайта.
2. **Фильтрация и валидация URL.** Ссылка считается допустимой только при соответствии шаблонам. Для каждого сайта свой шаблон, полученный путем анализа. Все остальные ссылки игнорируются. Также проверяется соответствие robots.txt с кэшированием на 24 часа.
3. **Извлечение структурированных данных**. Для каждой статьи извлекаются:
   - заголовок, автор, дата публикации;
   - основной текст;
   - исходный HTML.
4. **Очистка и фильтрация контента**. Перед сохранением:
   - Удаляются нерелевантные элементы;
   - Проверяется наличие ошибок (404, страница не найдена) по заголовку, h1, содержимому;
   - Пропускаются документы с менее чем 10 словами или пустым основным текстом.
5. **Структурированное хранение**. Все данные сохраняются в базу данных PostgreSQL в таблицу documents с полями:
   - url, normalized_url — исходный и нормализованный URL;
   - title, author, publish_date, clean_text, html;
   - source — имя домена;
   - fetch_timestamp, last_fetch_attempt — метки времени загрузки и последней попытки обновления;
   - last_modified_header, etag_header — HTTP-заголовки для эффективного обновления.
6. **Обновление ранее собранных документов**. После завершения начального сбора запускается фаза обновления:
   - Выбираются документы, у которых last_fetch_attempt старше 7 дней;
   - Для каждого выполняется условный HTTP-запрос;
   - При ответе 304 Not Modified обновляется только дата попытки;
   - При изменении контента — документ перезаписывается с новыми данными;
   - При редиректе — проверяется, не существует ли уже новая версия; при конфликте — обновление отменяется.

7. **Устойчивость и прерывание**
   - Состояние обхода сохраняется в файл crawler_state.pkl после каждого документа;
   - При получении сигнала Ctrl+C состояние сохраняется, и программа завершается корректно.

```
Подключение к бд успешно

Запускаю обновление

Пакет 1
Не изменилась: https://www.7ya.ru/article/6-navykov-bezopasnosti-kotorym-nuzhno-nauchit-malysha-2-3-let/
Не изменилась: https://www.7ya.ru/article/Mama-pyati-detej-kazhdyj-zasypaet-v-svoej-krovati/
Не изменилась: https://www.7ya.ru/article/Kak-nauchit-rebenka-ostavatsya-bez-mamy/
```

Рисунок 2. Пример работы поискового робота после сбора всех документов

В результате работы краулера был сформирован корпус, содержащий 54 543 документов.

	source	article_count
1	7ya.ru	8 417
2	letidor.ru	40 685
3	mama.ru	5 441

Рисунок 3. Распределение кол-ва документов по сайтам

Название	#	Тип данных
id	1	bigserial
url	2	text
normalized_url	3	text
source	4	text
created_at	5	timestamptz
html	6	text
clean_text	7	text
title	8	text
author	9	text
publish_date	10	text
last_modified_header	11	text
etag_header	12	text
fetch_timestamp	13	int8
last_fetch_attempt	14	int8

Рисунок 4. Поля в таблице documents

Количество документов	54 543
Источники	7ya.ru, mama.ru, letidor.ru
Средний размер сырого HTML	304.625 КБ
Средний размер извлечённого текста	3 192 символов, 5.66 КБ
Общий объём бд	8 ГБ

Таблица 1. Информация по данным в бд (значения получены путем запросов к 1000 рандомным документам из бд).

3. Предварительная обработка текста

3.1. Экспорт текстов из базы данных

Перед токенизацией тексты извлекаются из таблицы documents с помощью вспомогательного скрипта export_clean_text.py. Каждый документ сохраняется в отдельный файл docs/id.txt, где id — уникальный идентификатор документа в PostgreSQL.

3.2. Токенизация

Токенизация реализована в программе tokenizer.exe, написанной на C++17 с использованием UTF-8 обработки кириллицы. Программа читает все файлы из docs/, применяет правила фильтрации и сохраняет результат в tokens/id.tokens.

Основные правила токенизации:

- **Разрешены только русскоязычные слова** и **чистые числа** до 4 цифр. Английские слова, смеси символов, спецсимволы и HTML-сущности отбрасываются.
- **Слова ограничены по длине**, максимум 20 UTF-8-символов.
- **Дефис разрешён только один**, и только если по обе стороны от него корректные русские части, каждая не менее 2 кириллических символов, например: физико-математический.
- **Повторы символов** автоматически отфильтровываются.
- **Регистр приводится к нижнему**.
- **Известные аббревиатуры** сохраняются в исходном виде из файла known_abbrevs.txt.

```
Обработано 50000 документов, токенов: 19669370, скорость: 356.91 КБ/сек
Обработано 51000 документов, токенов: 20743564, скорость: 368.32 КБ/сек
Обработано 52000 документов, токенов: 21769820, скорость: 377.96 КБ/сек
Обработано 53000 документов, токенов: 22840348, скорость: 387.92 КБ/сек
Обработано 54000 документов, токенов: 23834032, скорость: 395.35 КБ/сек

Документов обработано: 54542
Всего токенов: 24388937
Средняя длина токена: 5.76 символов
Время выполнения: 770.80 сек
Скорость токенизации: 399.65 КБ/сек
```

Рисунок 5. Результаты запуска токенизации

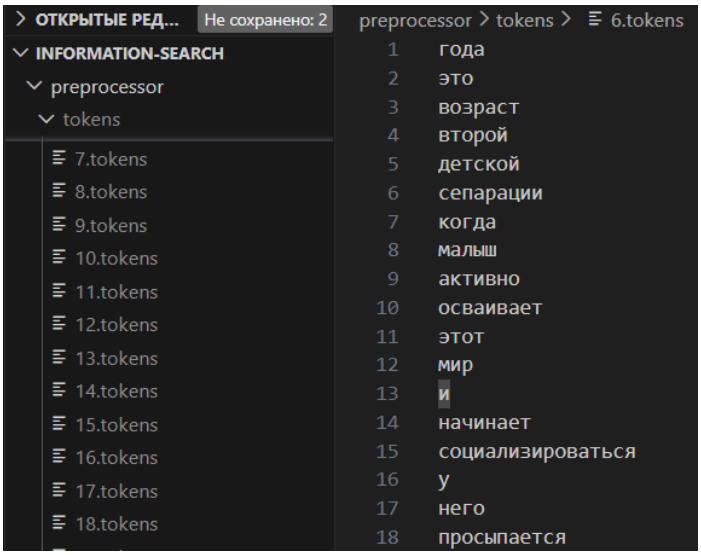

Рисунок 6. Пример токенов

3.3. Стемминг

После токенизации запускается программа stemmer.exe, которая читает файлы из tokens/ и применяет правила нормализации к каждому токену, сохраняя результат в stems/id.stems.

Особенности стеммера:

- **Удаление стоп-слов** перед стеммингом, список загружается из stop_words.txt. Это позволяет сократить размер индекса и улучшить релевантность.
- **Обработка возвратных глаголов**: суффиксы -ся / -сь корректно удаляются.
- **Специальные правила для русского языка**.

- **Исключения**: слова-глаголы высокой частотности (быть, мочь) остаются без изменений.
- **Минимальная длина стемы**: стемы короче 2 UTF-8 символов отбрасываются.
- **Все суффиксы отсортированы по убыванию длины**, чтобы сначала применять наиболее специфичные правила.

```
Обработано 50000 файлов, стем: 14585288 (отфильтровано: 5083168), скорость: 1.79 МБ/сек
Обработано 51000 файлов, стем: 15382135 (отфильтровано: 5361423), скорость: 1.84 МБ/сек
Обработано 52000 файлов, стем: 16139755 (отфильтровано: 5629793), скорость: 1.89 МБ/сек
Обработано 53000 файлов, стем: 16920224 (отфильтровано: 5919638), скорость: 1.94 МБ/сек
Обработано 54000 файлов, стем: 17652793 (отфильтровано: 6180504), скорость: 1.99 МБ/сек

Файлов обработано: 54543
Всего стем: 18056909
Отфильтровано стоп-слов: 6332030
Время выполнения: 144.41 сек
Средняя скорость: 2.01 МБ/сек
```

Рисунок 7. Результаты запуска стеммера

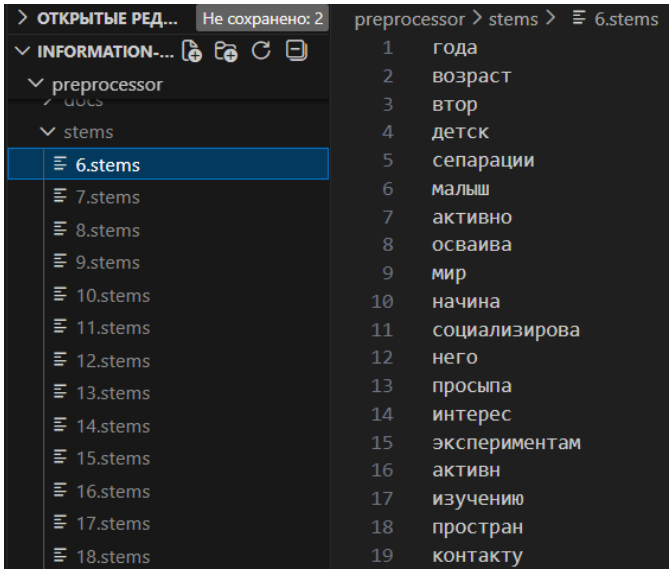

Рисунок 8. Пример стем

4. Закон Ципфа

Закон Ципфа — наблюдение, согласно которому частота встречаемости слова в тексте обратно пропорциональна его рангу в убывающем списке частот.

Для проверки закона был составлен список всех токенов корпуса с их частотами. Токены отсортированы по убыванию частоты. Построен график зависимости частоты токена от его ранга в логарифмических координатах.

```
Всего токенов: 24388939
Уникальных токенов: 461522
```

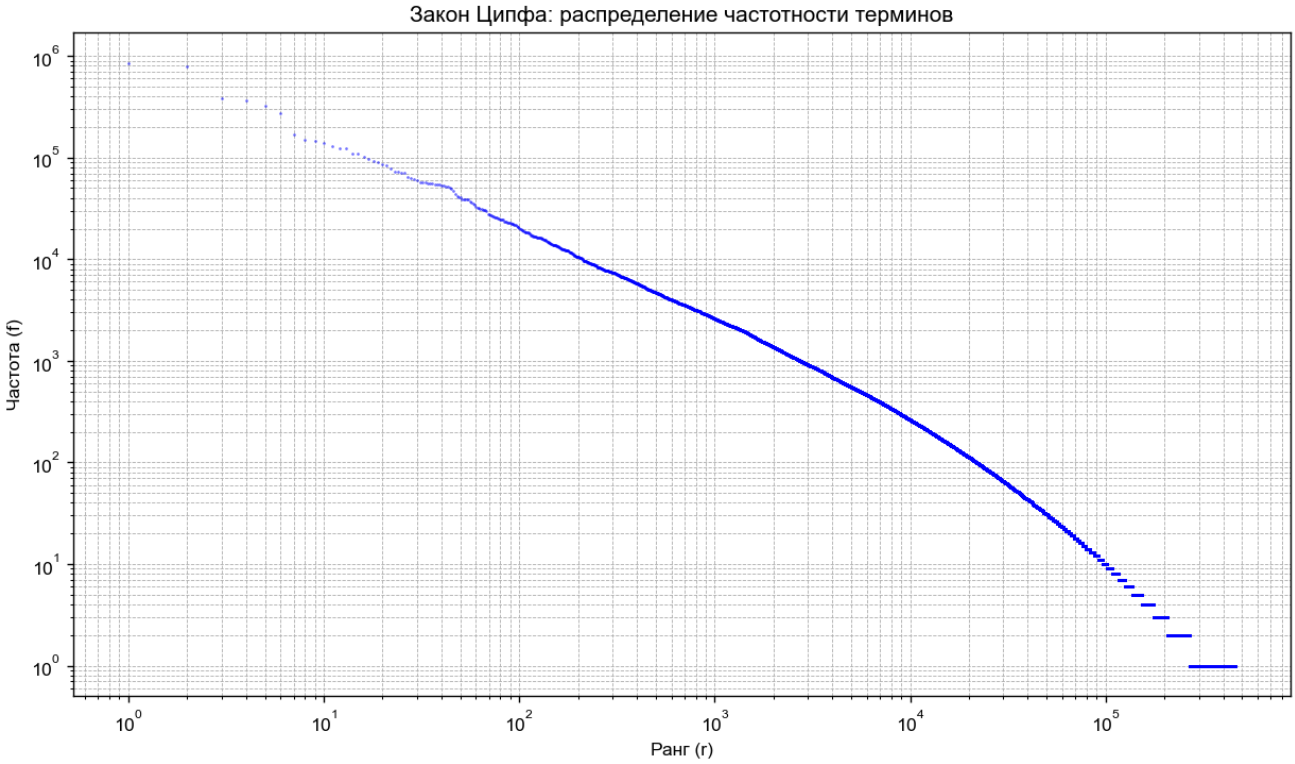

График 1. Закон Ципфа для собранного корпуса

График демонстрирует чёткую линейную зависимость в логарифмическом масштабе, это является подтверждением закона Ципфа. Это означает, что:

- Небольшое количество слов встречаются очень часто (верхняя часть графика);
- Огромное число слов встречается крайне редко (нижняя часть графика).

Наличие небольших отклонений в области высоких рангов (правая часть графика) — нормальное явление, обусловленное шумом и конечным объёмом корпуса.

5. Булев индекс

Булев индекс реализован как единая обратная индексная структура, оптимизированная для поддержки булевых запросов. Индекс строится на основе уже нормализованных и отфильтрованных стем из папки stems/. Для каждого уникального термина формируется posting-лист — отсортированный список идентификаторов документов, в которых встречается данный термин.

Особенности реализации:

- **Удаление дубликатов на уровне документа**: если термин встречается в тексте несколько раз, в posting-лист он добавляется только один раз.
- **Лексикографическая сортировка терминов**: все термины в индексе упорядочены по алфавиту, что упрощает последующий поиск и валидацию.
- **Сортировка posting-листов по возрастанию ID**: каждый список документов отсортирован с помощью алгоритма сортировки вставками.
- **Удаление дубликатов в posting-листах**: после сортировки из списков удаляются повторяющиеся идентификаторы.
- **Сохранение в текстовом формате**: индекс записывается в файл boolean_index.txt в виде термин:doc1,doc2,doc3... .

Индекс строится однократно и сохраняется на диск (boolean_index.txt). При последующих запусках поисковой системы он загружается напрямую.

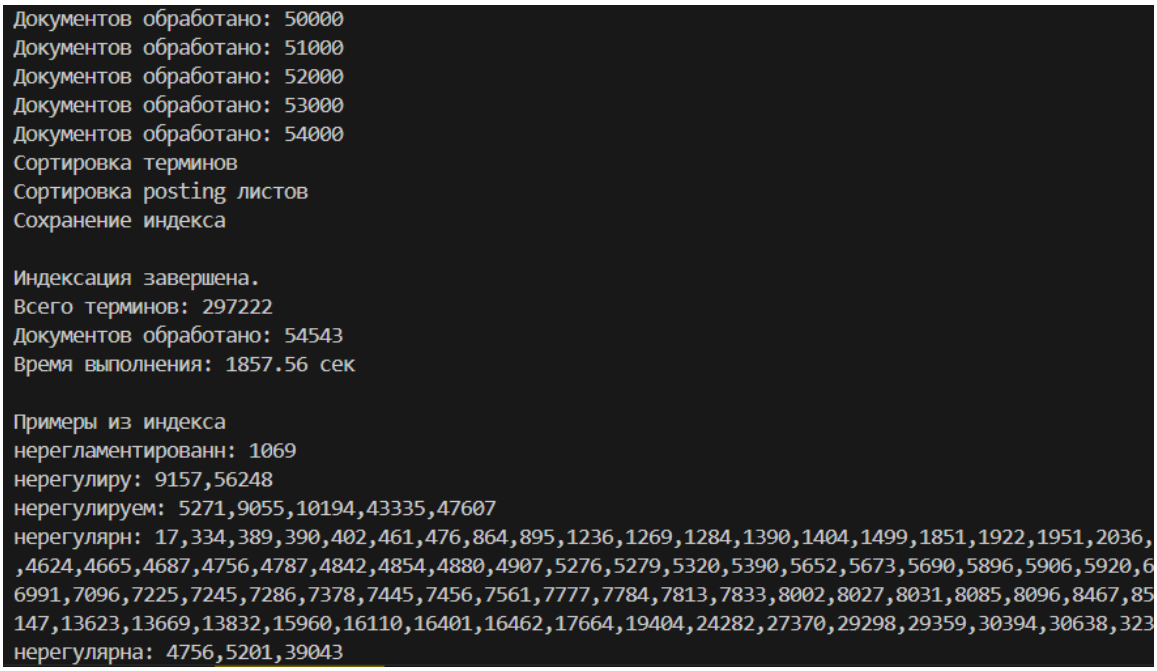

Рисунок 9. Результаты запуска булевого индекса

6. Булев поиск

Реализованная поисковая система поддерживает булевы запросы с использованием следующих операторов:

- and — логическое И (документ должен содержать все указанные термины);
- or — логическое ИЛИ (документ должен содержать хотя бы один термин);
- and not — исключение (документ содержит первый термин, но не содержит второй).

Запросы вводятся через консоль, например: мама and ребёнок.

Этапы алгоритма поиска:

1. **Преобразование запроса**: все термины приводятся к нижнему регистру для соответствия формату индекса.
2. **Поиск терминов**: для каждого термина выполняется бинарный поиск в списке boolean_index.txt.
3. **Извлечение posting-листов**: если термин найден, из индекса извлекается отсортированный список идентификаторов документов.
4. **Применение логических операций**:
    - Для оператора and выполняется пересечение двух списков (алгоритм двух указателей, линейная сложность);
    - Для or — объединение с удалением дубликатов;
    - Для and not — разность множеств (документы из первого списка, отсутствующие во втором).
5. **Формирование результата**:
    - В режиме --ids-only выводятся только идентификаторы документов;
    - В обычном режиме система подключается к базе данных PostgreSQL и извлекает заголовок и нормализованный URL каждого найденного документа.

Все списки документов в индексе отсортированы по возрастанию, что позволяет выполнять операции пересечения и объединения за линейное время без дополнительной сортировки.

```
Загрузка индекса.

Введите запрос:
ма-ма
Найдено: 17 документов
[id: 81] Календарь игр  - четвертый месяц — https://mama.ru/articles/kalendar-igr-4/
[id: 167] Обучение чтению: методики — https://mama.ru/articles/kak-uchat-chteniu/
[id: 174] Обучение чтению: методика Евгения Чаплыгина — https://mama.ru/articles/chtenie-po-bukvam-metodika-evgeniya-chaplygina/
[id: 4160] Почему при детях нельзя сидеть в телефоне — https://mama.ru/articles/pochemu-pri-detyah-nelzya-sidet-v-telefone/
[id: 6051] Крестить ли ребенка некрещеным родителям? — https://www.7ya.ru/article/Krestit-li-rebenka-nekrewenym-roditelyam/
[id: 7219] Зарядка для язычка: артикуляционная и дыхательная гимнастика для малышей — https://www.7ya.ru/article/Zaryadka-dlya-yazychka-artikulyacionnaya-i-dyhatelna
[id: 9273] Выход есть, или Моя история о послеродовой депрессии — https://www.7ya.ru/article/Vyhod-est-ili-Moya-istoriya-o-poslerodovoj-depressii-9433/
[id: 9486] Как превратить кубики в познавательную игру — https://www.7ya.ru/article/Kak-prevratit-kubiki-v-poznavatelnuyu-igru/
[id: 12850] Ребенок родился. Что главное в самом начале? — https://www.7ya.ru/article/Rebenok-rodilsya-Chto-glavnoe-v-samom-nachale-2017/
[id: 13559] Наш первый год — https://www.7ya.ru/article/Nash-pervyj-god-15009/
[id: 16836] "Возвращение Буратино" - старый мультфильм на новый лад — https://letidor.ru/otdyh/quot-vozvraschenie-buratino-quot-staryy-multfilm-na-novyy-lad.htm
[id: 18716] Методика обучения чтению: слоги, "ребус-метод" и карточки Домана — https://letidor.ru/obrazovanie/metodika-obucheniya-chteniyu-slogi-quot-rebus-metod-quo
[id: 20091] Все в порядке, мама: развитие младенца по месяцам — https://letidor.ru/dom-i-rebenok/vse-v-poryadke-mama-razvitie-mladenca-po-mesyacam.htm
[id: 23544] Как заставить ребенка заговорить — https://letidor.ru/zdorove/kak-zastavit-rebenka-zagovorit.htm
[id: 25791] Как развивать речь малыша от нуля до года: рассказывает логопед — https://letidor.ru/obrazovanie/kak-razvivat-rech-malysha-ot-0-do-goda-rasskazyvaet-logo
[id: 28075] Почему ребёнка нужно учить читать и писать одновременно — https://letidor.ru/obrazovanie/pochemu-rebyonka-nuzhno-uchit-chitat-i-pisat-odnovremenno.htm
[id: 56465] Юлия Меньшова рассказала, о чем врала в детстве — https://letidor.ru/novosti/yuliya-menshova-rasskazala-o-chem-vrala-v-detstve-17-06-2020.htm
Найдено: 17 документов



Введите запрос:
учитель
Найдено: 1 документов
[id: 53691] Стало известно об уровне стресса российских учителей — https://letidor.ru/novosti/stalo-izvestno-ob-urovne-stressa-rossiiskikh-uchitelei-23-03-2020.htm
Найдено: 1 документов



Введите запрос:
ма-ма AND учитель
Ничего не найдено.


Введите запрос:
ма-ма OR учитель
Найдено: 18 документов
[id: 81] Календарь игр  - четвертый месяц — https://mama.ru/articles/kalendar-igr-4/
[id: 167] Обучение чтению: методики — https://mama.ru/articles/kak-uchat-chteniu/
[id: 174] Обучение чтению: методика Евгения Чаплыгина — https://mama.ru/articles/chtenie-po-bukvam-metodika-evgeniya-chaplygina/
```

```
Введите запрос:
ОГЭ AND класс
Найдено: 18 документов
[id: 5821] 10 приложений для изучения английского в школе, с 1 по 11 класс — https://www.7ya.ru/article/10-prilozhenij-dlya-izucheniya-anglijskogo-v-shkole-s-1-po-11-klass/
[id: 7473] Подготовка к ОГЭ и ЕГЭ: чем отличается - и зачем писать олимпиады — https://www.7ya.ru/article/Podgotovka-k-OGJe-i-EGJe-chem-otlichaetsya-i-zachem-pisat-olimpiady/
[id: 7738] Вы уверены, что выбрали подходящую школу для ребенка? — https://www.7ya.ru/article/Vy-uvereny-chto-vybrali-podhodyawuyu-shkolu-dlya-rebenka/
[id: 8248] Путешествуем круглый год. А как же школа для ребенка? — https://www.7ya.ru/article/Puteshestvuem-kruglyj-god-A-kak-zhe-shkola-dlya-rebenka/
[id: 8857] Как правильно готовиться к ЕГЭ? — https://www.7ya.ru/article/Kak-podgotovitsya-k-EGJe-Tri-stadii-podgotovki-k-jekzamenam/
[id: 9837] Почему родители боятся дистанционного обучения — https://www.7ya.ru/article/Pochemu-roditeli-boyatsya-distancionnogo-obucheniya/
[id: 9887] Как сдать ЕГЭ по обществознанию? Результат учителя обществознания - 90 баллов — https://www.7ya.ru/article/Kak-sdat-EGJe-po-obwestvoznaniyu-Rezultat-uchitelya-obwes
[id: 11211] Как сдают ОГЭ и поступают в колледжи дети из детских домов — https://www.7ya.ru/article/Kak-sdayut-OGJe-i-postupayut-v-kolledzhi-deti-iz-detskih-domov/
[id: 12307] Куда пойти учиться? 7 плюсов колледжа после 9 класса — https://www.7ya.ru/article/Kuda-pojti-uchitsya-7-plyusov-kolledzha-posle-9-klassa/
[id: 12587] Как семейное обучение учит не бояться экзаменов — https://www.7ya.ru/article/Kak-semejnoe-obuchenie-uchit-ne-boyatsya-jekzamenov/
[id: 21258] 5 обучающих онлайн-сервисов: как подготовить ребенка к школе — https://letidor.ru/obrazovanie/5-obuchayuschih-onlayn-servisov-kak-podgotovit-rebenka-k-shkole.htm
[id: 27253] «Если не можете заинтересовать ребёнка чтением, не портите ему лето»: интервью с Димой Зицером — https://letidor.ru/otdyh/esli-ne-mozhete-zainteresovat-rebyonka-ch
o-intervyu-s-dimoy-zicer.htm
[id: 37182] В Волгограде ОГЭ за школьника написали учитель математики и десятиклассник — https://letidor.ru/novosti/v-volgograde-oge-za-shkolnika-napisali-uchitel-matematiki-i
18.htm
[id: 38518] Вопрос экспертам: должны ли дети выполнять на каникулах домашнее задание — https://letidor.ru/obrazovanie/vopros-ekspertam-dolzhny-li-deti-vypolnyat-na-kanikulakh-
[id: 47293] 7 страхов, которые одолевают мам по мере приближения 1 сентября — https://letidor.ru/psihologiya/7-strakhov-kotorye-odolevayut-mam-po-mere-priblizheniya-1-sentyabr
[id: 49759] Почему самостоятельная подготовка по сборникам ЕГЭ не дает результатов — https://letidor.ru/obrazovanie/pochemu-samostoyatelnaya-podgotovka-po-sbornikam-ege-ne-dae
[id: 54271] Для российских выпускников школ отменили весенний призыв в армию — https://letidor.ru/novosti/dlya-rossiiskikh-vypusknikov-shkol-otmenili-vesennii-prizyv-v-armiyu-
[id: 55381] 7 неоднозначных решений министра просвещения, принятых из-за коронавируса — https://letidor.ru/obrazovanie/7-neodnoznachnykh-reshenii-ministra-prosvesheniya-prinya
tm
Найдено: 18 документов



Введите запрос:
ОГЭ AND 9 AND класс
Неподдерживаемый запрос.
Ничего не найдено.
```

```
Введите запрос:
учеба AND NOT мама
Найдено: 184 документов
[id: 75] Одаренный ребенок — https://mama.ru/articles/pikovit/
[id: 638] Интенсивные нагрузки на глаза  – причина потери зрения — https://mama.ru/articles/intensivnye-nagruzki-na-glaza-prichina-poteri-zreniya/
[id: 1526] Готовим ребенка к школе — https://mama.ru/articles/gotovim-rebenka-k-shkole/
[id: 1529] Скоро 1 сентября: готовим первоклашку — https://mama.ru/articles/skoro-1-sentyabrya-gotovim-pervoklashku/
[id: 2267] «Чудо детки» – новый детский бренд от марки «Чудо» — https://mama.ru/articles/chudo-detki-novyi-detskii-brend-ot-marki-chudo/
[id: 2377] Как улучшить детскую память — https://mama.ru/articles/kak_uluchshit_detskuju_pamjat/
[id: 3057] Как найти репетитора по английскому для ребенка: 5 советов — https://mama.ru/articles/kak-najti-repetitora-po-anglijskomu-dlya-rebenka-5-sovetov/
[id: 3391] "Покидая Неверленд" - это учебное пособие для родителей — https://mama.ru/articles/pokidaya-neverlend-eto-uchebnoe-posobie-dlya-roditelej/
[id: 3563] 11 упражнений, которые помогут подготовить руку ребенка к письму — https://mama.ru/articles/11-uprazhnenij-kotorye-pomogut-podgotovit-ruku-rebenka-k-pismu/
[id: 4035] Первый раз в первый класс: как помочь ребенку адаптироваться к школе? — https://mama.ru/articles/pervyj-raz-v-pervyj-klass-kak-pomoch-rebenku-adaptirovatsya-k-shkole/
[id: 4115] Вредные привычки подростка: как быть родителям? — https://mama.ru/articles/vrednye-privychki-podrostka-kak-byt-roditelyam/
[id: 4368] Почему нельзя шлепать ребенка — https://mama.ru/articles/pochemu-nelzya-shlepat-rebenka/
[id: 4410] Гаджетомания: как победить зависимость — https://mama.ru/articles/gadzhetomaniya-kak-pobedit-zavisimost/
[id: 4430] Уровень стресса школьников повысился на дистанционном обучении — опрос — https://mama.ru/articles/uroven-stressa-shkolnikov-povysilsya-na-distantsionnom-obuchenii-opros/
[id: 5830] Где учатся с интересом от первого до выпускного? — https://www.7ya.ru/article/Gde-uchatsya-s-interesom-ot-pervogo-do-vypusknogo/
[id: 6273] Как учиться за рубежом бесплатно? 4 варианта — https://www.7ya.ru/article/Kak-uchitsya-za-rubezhom-besplatno-4-varianta/
```

Рисунки 10, 11, 12. Примеры запросов в обычном режиме

```
Загрузка индекса.
Режим: только ID. Введите запрос:
ИВ
Найдено: 33 документов
2672
3170
3234
4675
4911
5257
5901
6218
6872
6923
7335
8415
8888
9393
11183
12008
12094
12682
13734
13840
16221
17629
20255
27816
29060
32796
35216
38588
42397
48657
50954
53388
53529
Найдено: 33 документов

Введите запрос:
ИВ AND мама
Найдено: 13 документов
3170
3234
4675
5257
5901
8415
11183
12682
```

Рисунок 13. Примеры запросов в --ids-only режиме

7. Заключение

В ходе выполнения лабораторной работы был реализован полноценный прототип поисковой системы, включающий:

- автоматический сбор тематического корпуса;
- лингвистическую обработку текста (токенизация и стемминг);
- построение и использование булева индекса;
- поддержку булевых запросов с логическими операциями;
- анализ корпуса на соответствие закону Ципфа;
- механизм быстрого запуска за счёт сохранения состояния.