Разработка API-сервиса и информационно-поисковой системы для встраивания поиска на сайтах

Сидоров С.В (galserge.sidorov@gmail.com),
Черкасова В.А. (valyc@mail.ru)

Астраханский государственный университет им. В. Н. Татищева

«Математика. Компьютер. Образование – 2023»

Цели и задачи

Целью данной работы является разработка API-сервиса полнотекстовой поисковой системы. Актуальность работы обусловлена необходимостью оперативного получения информации пользователями официального портала Астраханского государственного университета. В процессе разработки требовалось решить следующие задачи:

1. изучить существующие технологии полнотекстового поиска
2. реализовать алгоритм поиска
3. спроектировать и разработать API-сервис для встраивания поиска на сайтах

Полнотекстовый поиск и индексирование

Определение

Полнотекстовый поиск — это автоматизированный поиск документов, при котором поиск ведётся не по именам документов, а по их содержимому.

Поиск осуществляется на основе заранее построенного индекса, что позволяет ускорить процесс оценки и извлечения информации, соответствующей запросу.

Алгоритм индексирования

На вход построителю индекса подается коллекция документов D (в данном случае это страницы сайта, полученные из базы). В процессе построения индекса осуществляется:

1. очистка документов от тегов и html-сущностей, приведение строк в нижний регистр
2. удаление служебных слов
3. токенизация (разбиение на слова) и лемматизация (определение словарной формы слов)
4. построение мешка слов с помощью TF-IDF

Векторы TF-IDF

Определение 1.

Мешок слов - представление коллекции текстов в виде таблицы, в которой строки - документы, а столбцы - слова. В ячейках таблицы содержатся числа, соответствующие количеству вхождений конкретного слова в документ.

Определение 2.

TF-IDF - (TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции.

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D),$$

где $tf(t, d)$ - частота термина t в документе d, $idf(t, D)$ - обратная частота термина t в коллекции документов D.

Оценка релевантности

На вход поисковику подается строка запроса q. Для получения подмножества релевантных документов применяются следующие операции:

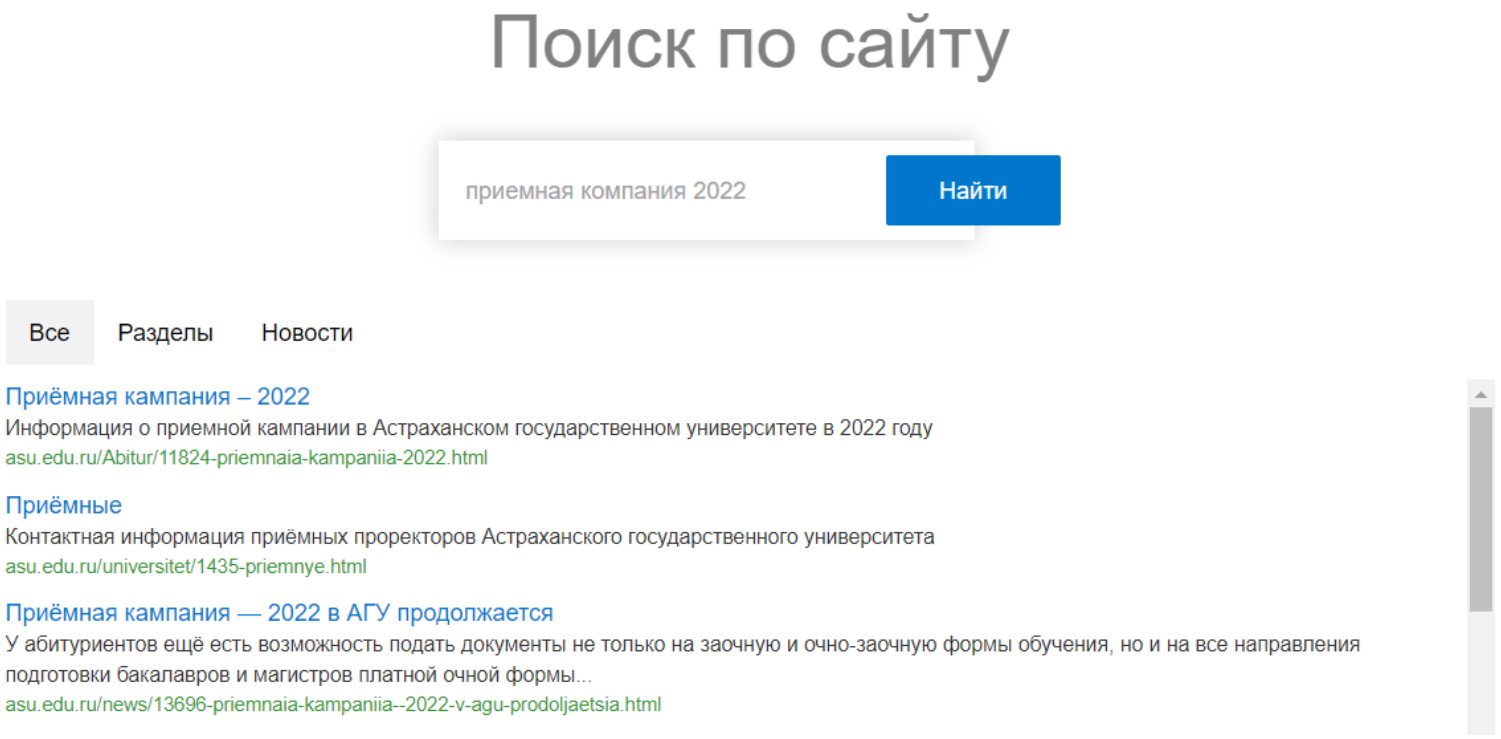
1. запрос q проходит в точности все этапы очистки преобразования, что и коллекция документов при индексировании
2. запрос q преобразуется в вектор TF-IDF относительно словаря, составленного про индексировании
3. каждый документ коллекции сравнивается с запросом q

$$score(q, d) = cos(\mathbf{q}, \mathbf{d})$$

На последнем этапе оценки сортируются по возрастанию величины косинусного расстояния между векторами документов $\mathbf{d}$ и вектором запроса $\mathbf{q}$.

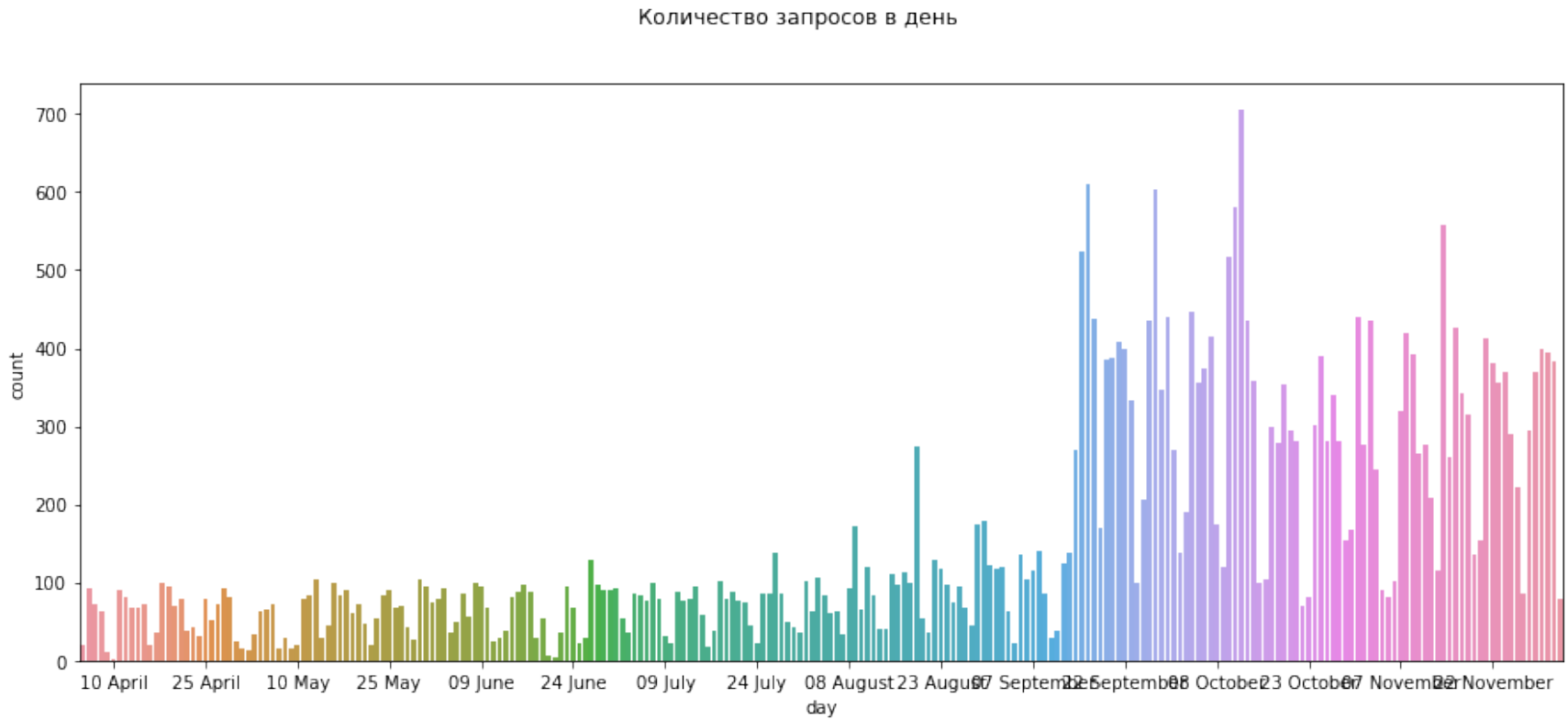
Поиск на сайте АГУ

Результатом данной работы стала реализация поискового API, на базе фреймворка fastAPI, его прототип используется на сайтах Астраханского государственного университета.



Использование поиска

Внедрение поиска было проведено в конце сентября 2022 года. После чего активность использования поиска на сайте возросла в 4 раза в сравнении с поиском от Яндекса.



Заключение

Внедрение нового поиска на сайте повысило удобство пользования сайтом, что подтверждается возросшим числом запросов.
Разработанный API-сервис позволяет подключить к поиску несколько сайтов.
В дальнейшем планируется доработка интерфейса API (упрощение принципа работы) и реализация функций автодополнения, реферирования текста и исправления ошибок и опечаток в запросах.