Министерство образования Республики Беларусь

Учреждение образования

“Белорусский государственный университет информатики и радиоэлектроники”

Факультет информационных технологий и управления

Кафедра интеллектуальных информационных технологий

Лабораторная работа №1

по дисциплине «Естественно-языковые интерфейсы интеллектуальных систем»

Выполнили Тесловский А.П.

студенты группы 721702 Ахраменко М.Г.

Проверила Крапивин Ю.Б.

Минск 2020

Описание системы:

Основной моделью системы является модель токена. Она содержит непосредственно сам токен в строковом представлении, коэффициент значимости, и url источника.

Система имеет следующие сервисы:

1) Сервис токенизации. Удаляет из текста знаки препинания, ненужные слова, извлекает токены.
2) Сервис индексации. Для каждого источника производится индексация и создаются токены, также производится расчет коэффициента значимости и числа вхождений в файл.
3) Сервис поиска. Ищет пересечения токенов из поискового запроса с токенами из файлов, коэффициент значимости больше или равен конфигурируемому минимальному значению, формирует результат поиска.

В системе присутствуют прочие элементы, функционал которых направлен на обеспечение работоспособности приложения и пользовательского интерфейса (HTML-шаблоны).

Используемые библиотеки и фреймворки:

Для реализации системы использовалась библиотека lemmatizer, которая позволяет работать с естественным языком (в данном случае с английским). Эта библиотека используется для выделения токенов из словоформ.

Также, в основном для реализации пользовательского графического интерфейса и работой с базой данных, был использован фреймворк Rails. Это веб-фреймворк, имеющий удобный шаблонизатор и ORM реализующий шаблон Active Record.

Тестирование системы:

Для тестирования системы были взяты 8 источников: 4 про Дональда Трампа и 4 про котов.

Результат поиска по запросу “trump”:

https://edition.cnn.com/2020/10/24/politics/donald-trump-joe-biden-oil-fracking-election-2020/index.html

https://en.wikipedia.org/wiki/Donald_Trump

https://www.whitehouse.gov/people/donald-j-trump/

Результат поиска по запросу “cat”

https://www.polygon.com/2019/7/18/20696129/cats-2019-musical-trailer-explained

https://www.purina.co.uk/cats/behaviour-and-training/understanding-cat-behaviour/fun-facts-about-cats

https://www.purina.com/articles/cat/facts/10-fascinating-facts-about-cats

https://en.wikipedia.org/wiki/Cat

Результат поиска по запросу “president”

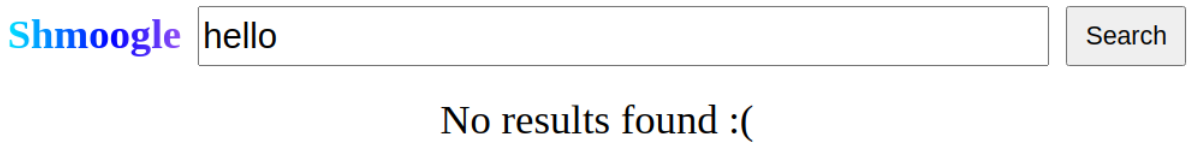
https://en.wikipedia.org/wiki/Donald_Trump

https://www.whitehouse.gov/people/donald-j-trump/

Результаты поиска по запросу “purr”:

https://en.wikipedia.org/wiki/Cat

Результаты поиска по запросу “hello”:



Вывод:

За счет использования библиотеки lemmatizer удалось осуществить поиск по словоформам (например, если в тексте есть слово mentioned, то этот текст будет результатом поиска по запросу “mention”). Благодаря применению алгоритма индексации удалось сделать процесс поиска значительно быстрее процесса поиска без применения такого алгоритма. Однако сама индексация и проводится довольно долго, ее можно проводить в фоновом режиме.