Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра интеллектуальных информационных технологий

Отчёт

по курсу «**Естественно-языковой интерфейс интеллектуальных систем**»

Лабораторная работа №2

«Разработка текстового корпуса, его менеджера»

Выполнили студенты группы 121701:	Липский Р. В. Жолнерчик И. А. Стронгин А. В.
Проверил:	Крапивин Ю.Б.

Минск 2023

Цель работы:

1. Изучить принципы построения корпусов текстов, виды разметки и способы аннотирования, инструменты работы с корпусами текстов,
2. Построить корпус текстов и разработать корпусный менеджер.

Задание:

1. Сформировать электронный корпус текстов по выбранной предметной области.

2. Разработать корпусный менеджер, обеспечивающий базовую функциональность работы с созданным корпусом текстов.

Вариант: 13

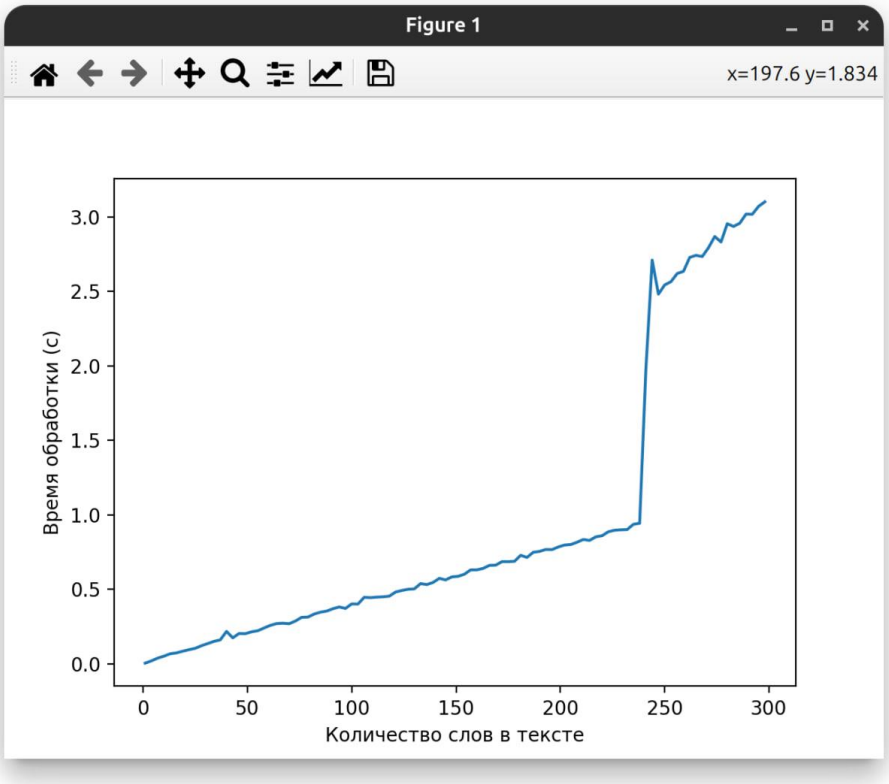
Язык текста: английский

Предметная область: Кинематограф

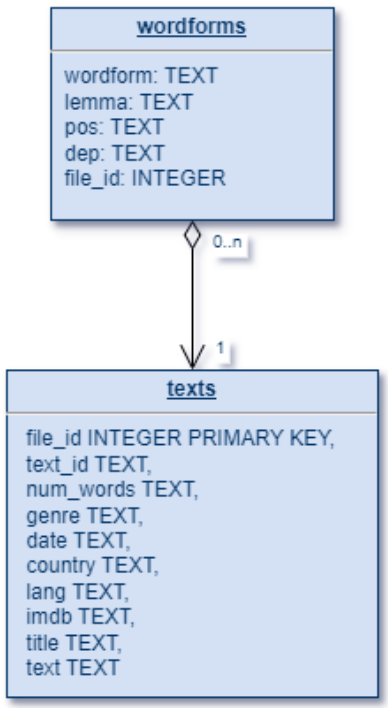
Интерфейс:

Главное окно

Тест производительности:



Структура хранения данных



Структурная схема приложения

Описание алгоритмов:

Разбиение текста на словоформы и получение их морфологической информации:
Начало – Получаем файл формата TXT/RTF – Считываем текст файла – Получаем список словоформ – Получаем всю морфологическую информацию о каждой – Преобразуем в неформатированную строку – Загружаем словоформы и их морфологическую информацию в БД – Конец

Поиск словоформ по подстроке:
Начало – Пользователь вводит строку – Получаем из БД список всех словоформ, которые имеют такую подстроку – Отображаем все словоформы и морфологическую информацию о ней в таблице – Конец.

Поиск примеров использования в контексте:
Начало – Пользователь вводит строку – Получаем из БД список всех словоформ, которые имеют такую подстроку – Получаем из БД тексты источники данных словоформ – Обрезаем тексты до предложения, содержащего данную словоформу – Отображаем словоформы и примеры в таблице – Конец.

Вывод:

В результате выполнения данной лабораторной работы мы изучили понятие текстового корпуса и его менеджера. Текстовый корпус представляет собой набор текстовых документов, собранных для анализа и обработки в рамках определенной задачи или исследования. Он является важным инструментом в области обработки естественного языка, машинного обучения и лингвистического анализа.