

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:

Corpus Manager

Contact with developers

Query:

name

Search

Number of occurrences:

1210

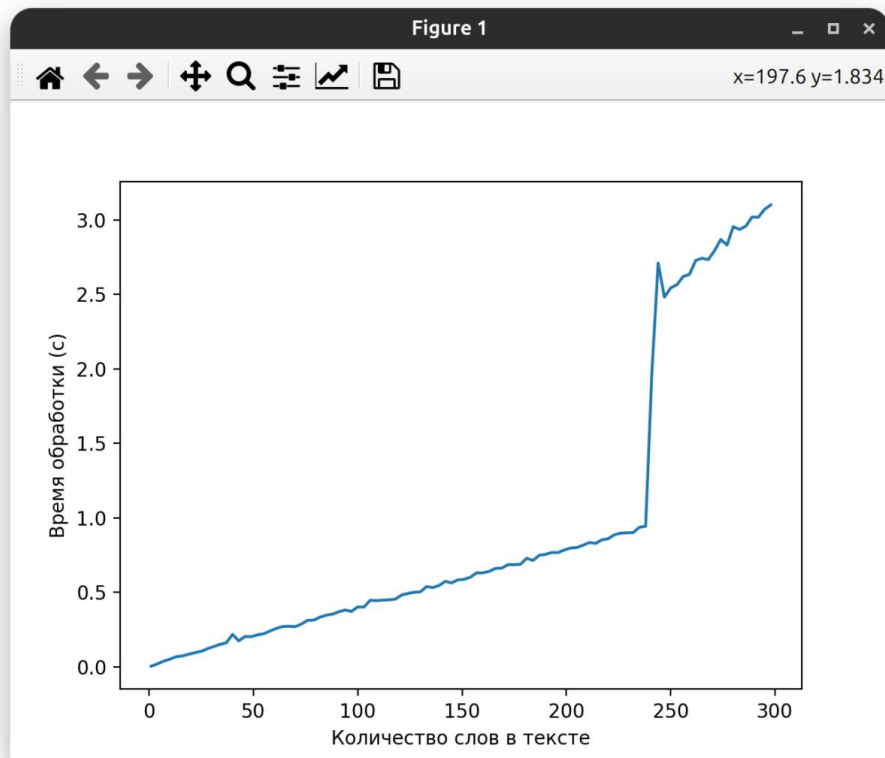
Search results:

Wordform	Lemma	POS	Link
name	name	noun	Anna Christie (USA, 1930)
name	name	noun	Over the Hill (USA, 1931)
names	name	noun	The Littlest Rebel (USA, 1935)
name	name	noun	The Littlest Rebel (USA, 1935)
name	name	noun	The Robber Kitten (USA, 1935)
name	name	noun	Charlie Chan at the Race Track (USA, 1936)
names	name	noun	The Prisoner of Shark Island (USA, 1936)
name	name	noun	The Prisoner of Shark Island (USA, 1936)
name	name	noun	The Lady Vanishes (UK, 1938)
name	name	noun	Holiday (USA, 1938)
named	name	verb	Holiday (USA, 1938)
names	name	noun	Holiday (USA, 1938)
name	name	verb	Holiday (USA, 1938)
name	name	noun	Sun Valley Serenade (USA, 1941)
name	name	noun	The Outlaw (USA, 1943)
name	name	noun	The Seventh Victim (USA, 1943)
name	name	noun	Between Two Worlds (USA, 1944)
name	name	noun	The Unsuspected (USA, 1947)
ornaments	ornament	noun	The Unsuspected (USA, 1947)
name	name	noun	Sands of Iwo Jima (USA, 1949)

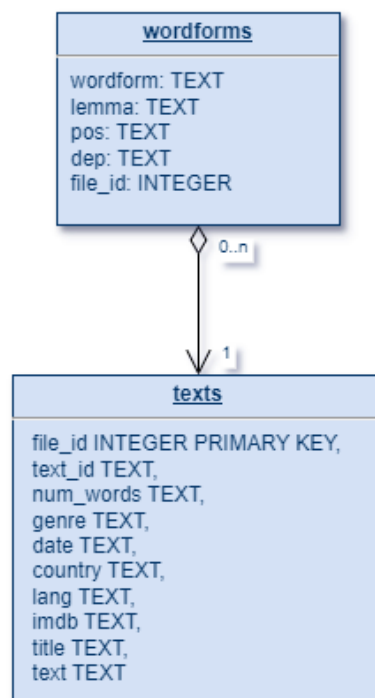
Example	Link
- That 's his full name .	Anna Christie (USA, 1930)
Thou shalt not take the name of the Lord thy God in vain .	Over the Hill (USA, 1931)
Well , the world is full of names .	The Littlest Rebel (USA, 1935)
That 's no name for me .	The Robber Kitten (USA, 1935)
Only three beside humble self knew name of death weapon .	Charlie Chan at the Race Track (USA, 1936)
[Audience_Laughing] , Faint [Actor] I expect I 'm liable to call myself some awful	The Prisoner of Shark Island (USA, 1936)
Being in my right mind , I shall take the veil and the orange blossom and changi	The Lady Vanishes (UK, 1938)
- What 's her name ?	Holiday (USA, 1938)
I 'd like to , but Mr Murray is n't considering anything but a big-name band for S	Sun Valley Serenade (USA, 1941)
My name 's Holiday .	The Outlaw (USA, 1943)

Главное окно

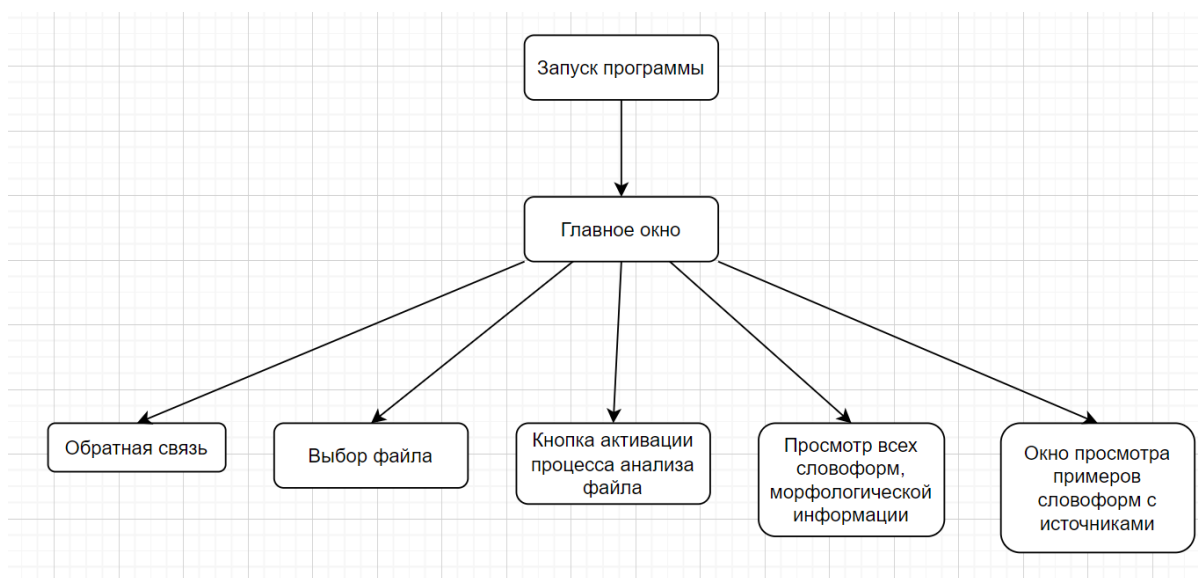
Тест производительности:



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.