Министерство образования Республики Беларусь

Учреждение образования “Белорусский государственный университет информатики и радиоэлектроники”

Факультет информационных технологий и управления

Кафедра интеллектуальных информационных технологий

Отчёт по дисциплине

“Естественно-языковой интерфейс интеллектуальных систем”

Лабораторная работа №1

“Разработка автоматизированной системы формирования словаря естественного языка”

Выполнили студенты группы 221702: Хлуд А.Н

Чаюк Т.А.

Набешко Д.С.

Проверил: Крапивин Ю.Б.

Минск 2024

Лабораторная работа №1

**Цель работы:**

Освоить принципы разработки прикладных сервисных программ для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка.

**Задание (№4, вариант 10):**

Список слов, упорядоченный по алфавиту и включающий

только лексемы с дополнительно оформленными записями о месте и роли

данного слова в составе предложения. К такой информации относится

описание того, каким членом предложения может быть данное слово и в какой форме (падеж, число, время и т.п.). Например, если это существительное в именительном падеже, то оно может выступать в роли подлежащего; если это существительное в родительном падеже, то оно может быть дополнением; если это прилагательное, то оно может быть определением и т.п.

**Язык текста:**

Русский

**Формат входного документа:**

RTF, TXT.

**Используемые средства разработки:**

Для решения поставленной задачи было разработано приложение на языке программирования **Python** с использованием некоторых библиотек.

Библиотека **Pymorpfy3** - предназначена для проведения морфологического разбора каждого слова — определение его роли (подлежащее, сказуемое и т.д.) и грамматических форм (падеж, род, число и т.д.).

Библиотека **tkinter** – это стандартная библиотека для создания графических интерфейсов (GUI) в Python. Она является оболочкой над библиотекой Tk, которая используется для разработки кроссплатформенных интерфейсов.

Библиотека **re** - библиотека предназначенная для работы с регулярными выражениями. Разбивает текст на предложения, а их на слова.

Библиотека **json -** библиотека предназначенная для сохранения и загрузки данных в формате JSON, а именно для сохранения словаря слов с ролями и формами.

Библиотека **os -** библиотека предназначена для работы с операционной системой — файлы, директории, проверки на существование. Она проверяет существует ли файл перед загрузкой.

**Используемые структуры для хранения данных**

Основной словарь естественного языка реализован с использованием стандартной структуры данных Python — словаря. В нем хранится информация о словах, их частях речи, а также о словосочетаниях и их типах.

Система поддерживает работу со следующими частями речи:

* **"VERB": "глагол",**
* **"AUX": "вспомогательный глагол",**
* **"NOUN": "существительное",**
* **"PROPN": "имя собственное",**
* **"ADV": "наречие",**
* **"ADJ": "прилагательное",**
* **"PRON": "местоимение",**
* **"ADP": "предлог",**
* **"DET": "определитель",**
* **"PUNCT": "знак препинания",**
* **"CCONJ": "сочинительный союз",**
* **"SCONJ": "подчинительный союз",**
* **"NUM": "числительное",**
* **"PART": "частица",**
* **"INTJ": "междометие",**
* **"SYM": "символ"**

Для реализации структур хранения базы данных было использовано **СУБД** **SQlite**

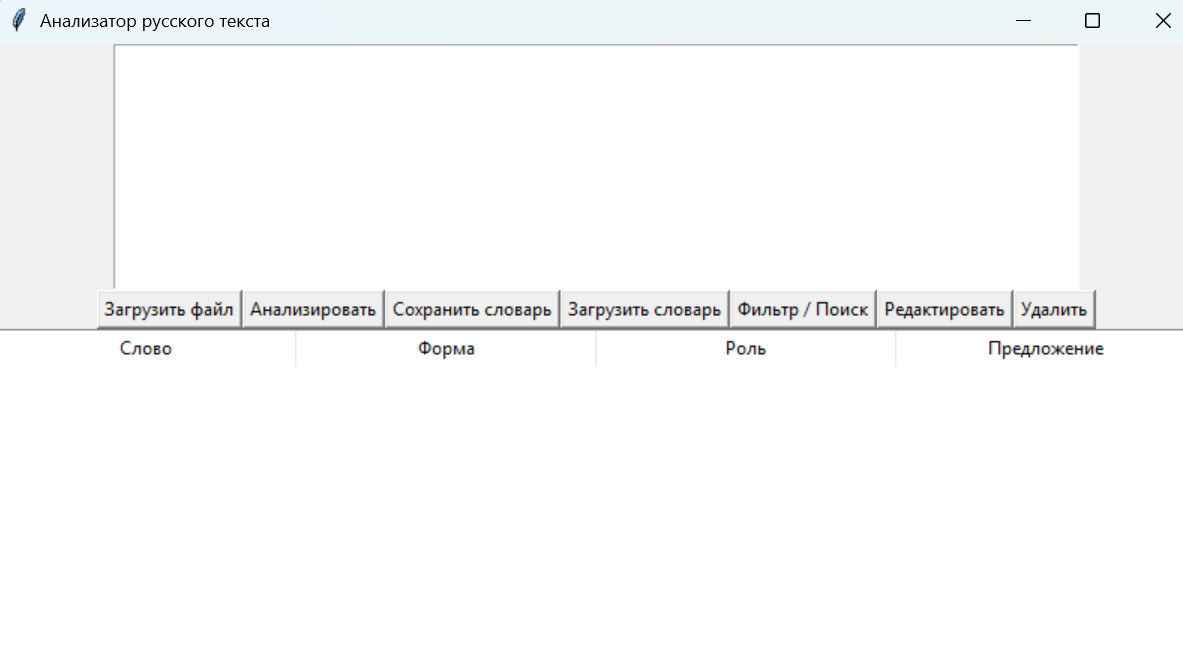
SQLite — это легковесная, встраиваемая база данных, которая хранит данные в одном .sqlite или .db файле и не требует отдельного сервера.

Структура таблицы:

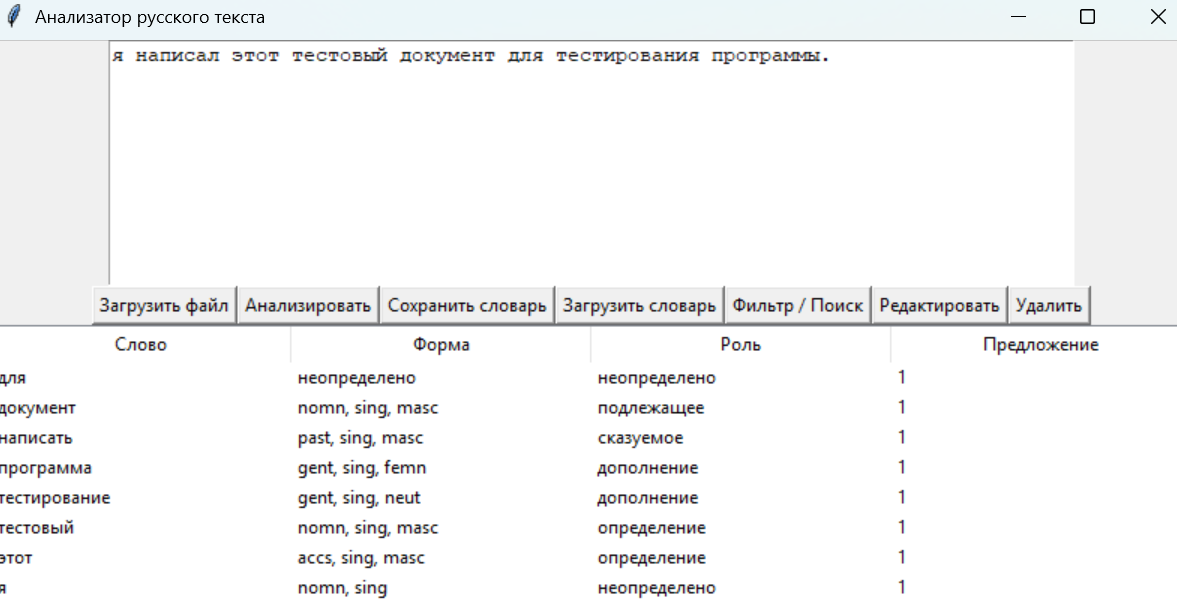
|  |  |  |  |
| --- | --- | --- | --- |
| id (INTEGER) | word (TEXT) | role (TEXT) | morphological\_features (TEXT) |
| 1 | "cat" | "subject" | "morphological\_features" |
| 2 | "run" | "verb" | "morphological\_features" |

**Демонстрация интерфейса разрабатываемого приложения**

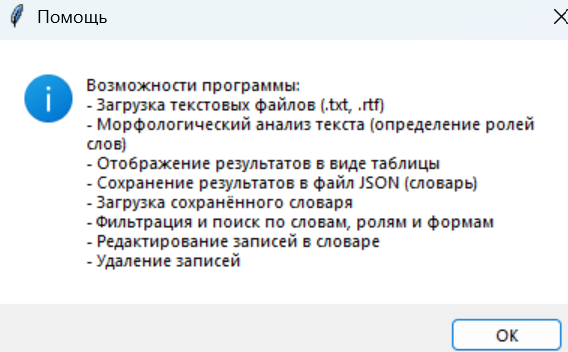
**Главное окно приложения:**

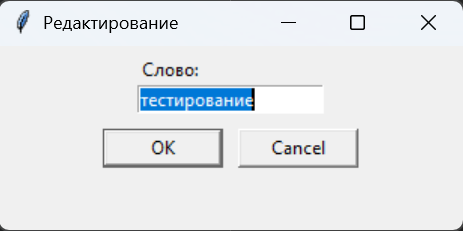
  
Рисунок 1 - главное окно приложения

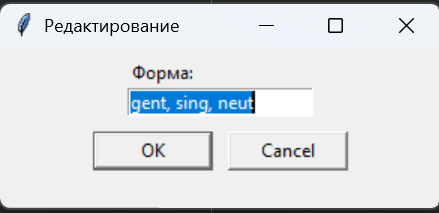
**Окно после загрузки текста:**

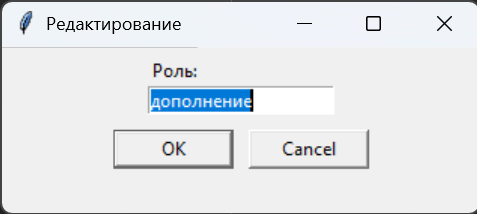
  
Рисунок 2 – окно после загрузки текста

**Окно с инструкциями:**

  
Рисунок 3 – окно просмотра инструкций

**Окно редактирования записи:**





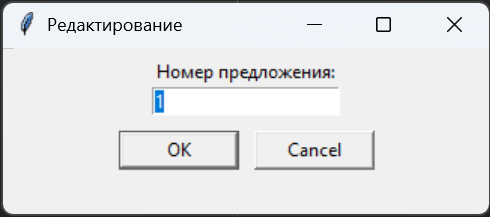


Рисунок 5 – окно редактирования записи

**Структурно-функциональная схема разрабатываемого приложения:**

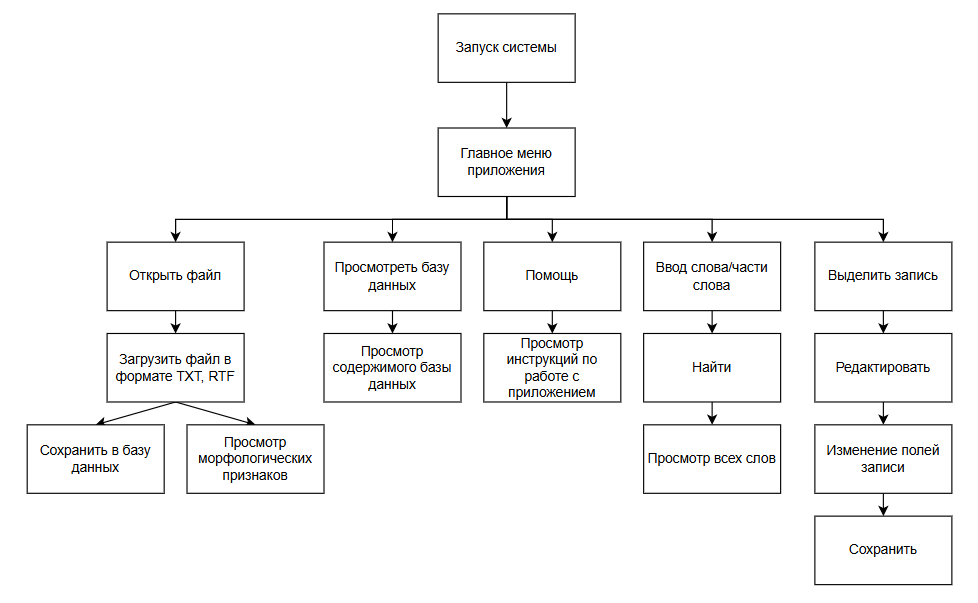
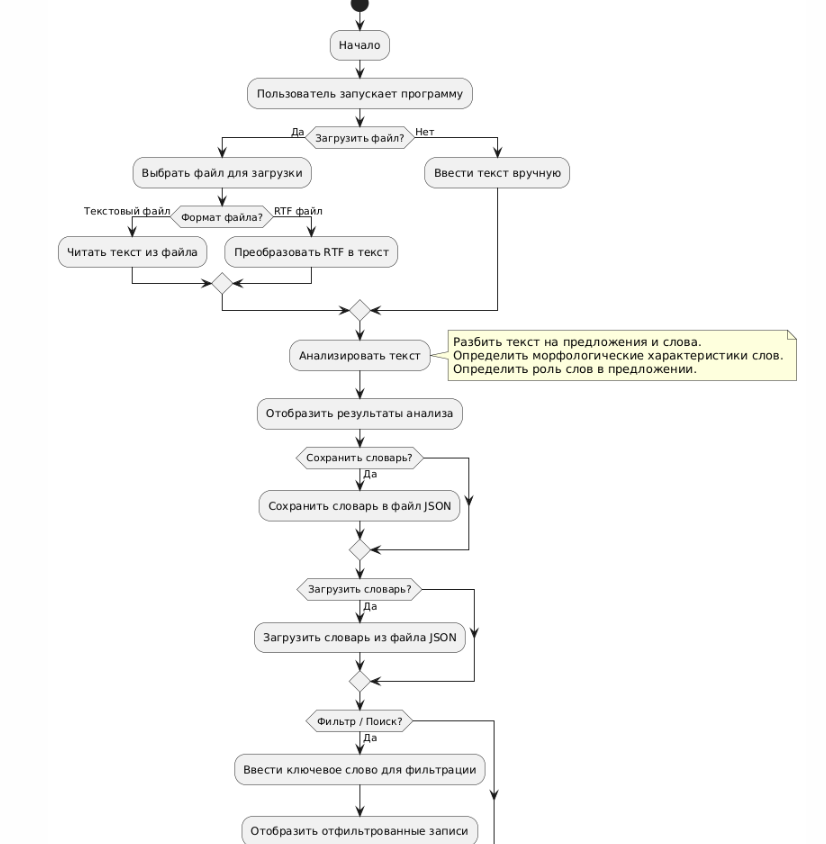


Рисунок 6 - структурно-функциональная схема разрабатываемого приложения

**Алгоритм определения части речи слова в тексте и его морфологические признаки:**



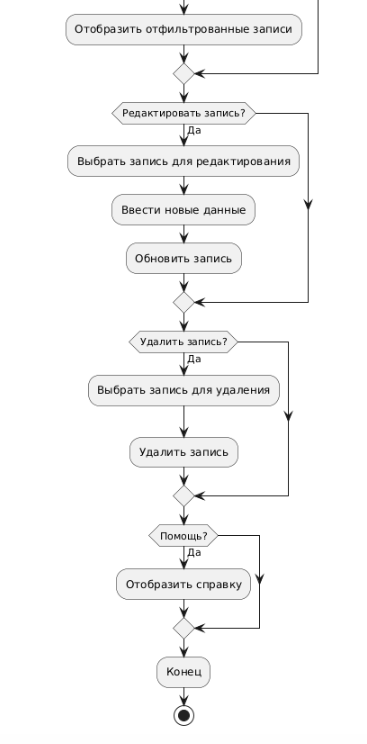


Рисунок 7 - Алгоритм определения части речи слова в тексте и его морфологические признаки

**Тестирование скорости программы**

Для тестирования быстродействия приложения были использованы файлы формата .txt, содержащие текст, размер которого составляет 50, 150, 500, 1000 и 5000 слов.

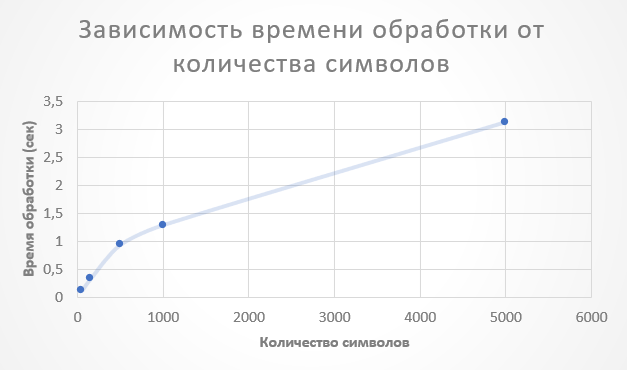


Рис. 8 - график зависимости времени обработки от количества слов в анализируемом тексте

**Выводы по работе и по перспективам развития приложения:**

Программа представляет собой автоматизированную систему формирования словаря естественного языка, представляющего собой список упорядоченных по алфавиту слов вместе с описанием того, каким членом предложения может быть данное слово и в какой форме.

В качестве перспективы развития можно выделить модификацию алгоритма, определяющего каким членом предложения может быть определённое слово – на данном этапе алгоритм хоть и имеет высокую точность, но может точно предсказать лишь наиболее часто встречающиеся случаи.