NaМинистерство образования Республики Беларусь

Учреждение образования

“Белорусский государственный университет

информатики и радиоэлектроники”

Факультет информационных технологий и управления

Кафедра интеллектуальных информационных технологий

Лабораторная работа №1

на тему

“Разработка автоматизированной системы формирования словаря естественного языка”

по дисциплине «Естественно-языковой интерфейс интеллектуальных систем»

|  |  |
| --- | --- |
| Выполнил: | Код: Поплавский Э. Э.  Отчет: Колосовский Е. С.  Отчет: Смелов А. А. |
| Группа: | 021703 |
| Проверил: | Name N. N. |

Минск 2023

Содержание

[1. Цель работы:](#_qgwal63jo2yb)

[2. Задачи лабораторной работы:](#_v2xm9djnywb7)

[3. Структурно-функциональная схема разработанного приложения.](#_z701oh6bkqde)

[4. Описание структур хранения данных, алгоритмов их обработки, необходимых для реализации базовых требований к разработанной программе.](#_ayvov9v3c3d6)

[5. Оценка быстродействия приложения.](#_qzit0noawv8k)

[6. Вывод.](#_h3i38hmtu3aj)

# Цель работы:

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

# Задачи лабораторной работы:

Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического лексического и лексико-грамматического анализа ТЕЯ.

Закрепить навыки программирования при решении задач автоматической обработки ТЕЯ.

Вариант лабораторной работы №14.

# Структурно-функциональная схема разработанного приложения.

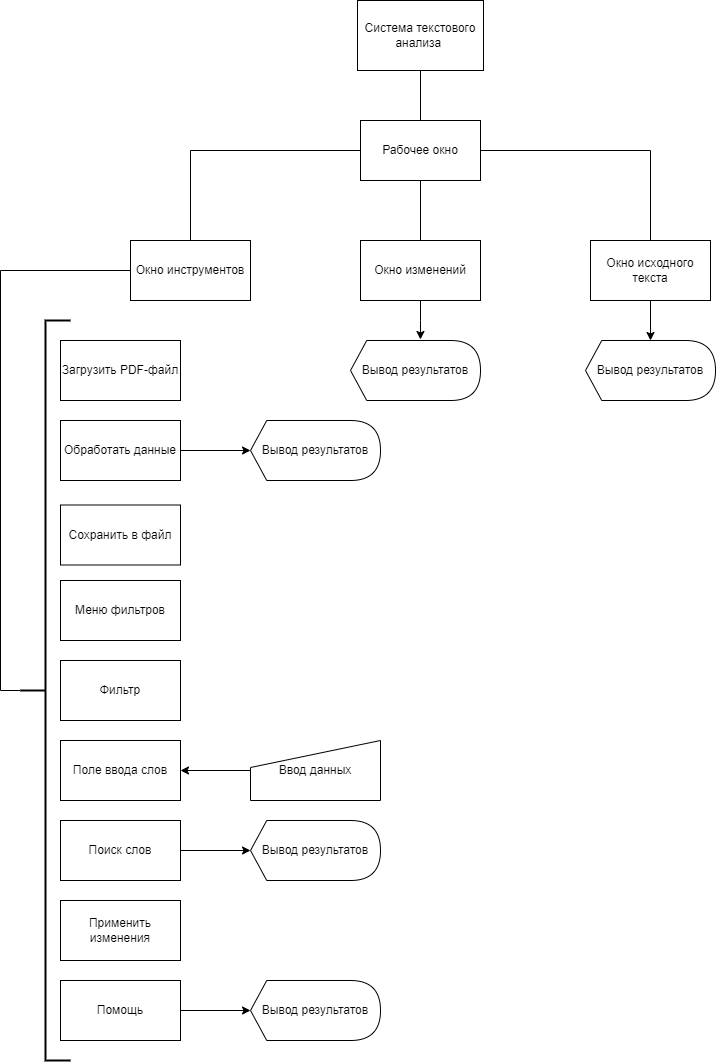


Рисунок 1. Структурно-функциональная схема разработанного приложения.

# Описание структур хранения данных, алгоритмов их обработки, необходимых для реализации базовых требований к разработанной программе.

Согласно варианту задания форматом входного документа является PDF файл с текстом на английском языке.

Для хранения данных о лексемах в программе был создан дата-класс WordForm с полями содержащими информацию о словоформе, части речи, роде, числе, падеже и частоте.

За логику обработки данных отвечает класс WordForms в котором реализованы такие методы как добавить, удалить, изменить лексему, сортировать словоформы по алфавиту, сохранить словарь в файл.

Форматом выходного документа является csv-файл содержащий словарь лексем в виде таблицы.

Для анализа полученных лексем использовалась библиотека NLTK. Для создания графического пользовательского интерфейса использовалась библиотека Tkinter. Скриншоты работы программы приведены ниже на рисунках 2-4.

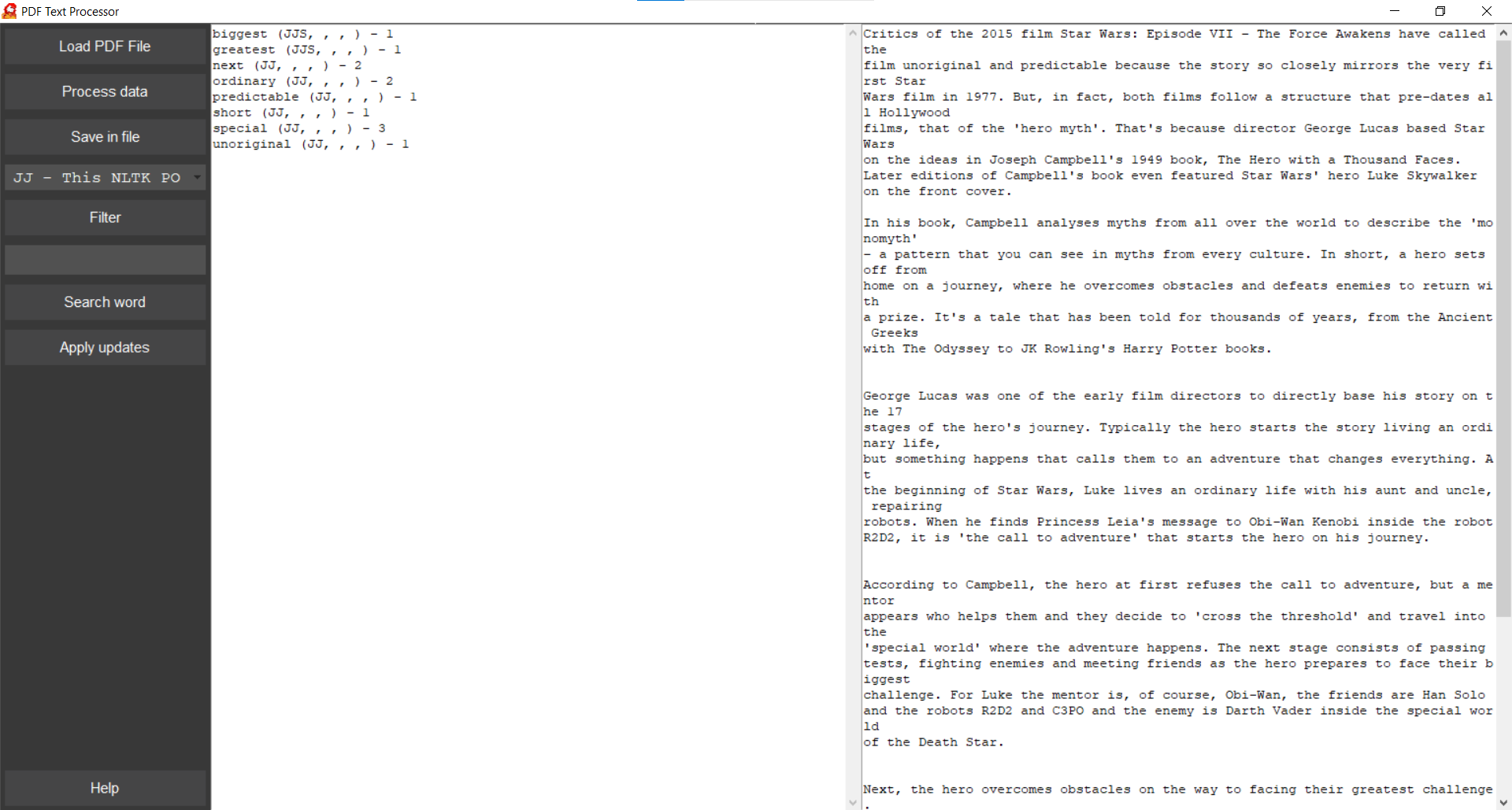


Рисунок 2. Главное окно приложения.

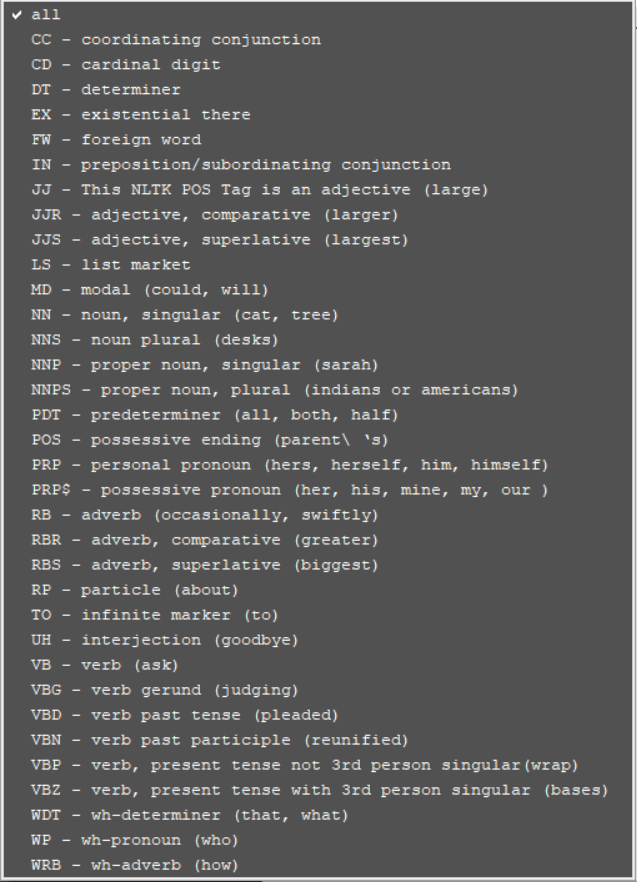


Рисунок 3. Доступные фильтры частей речи для словоформ.

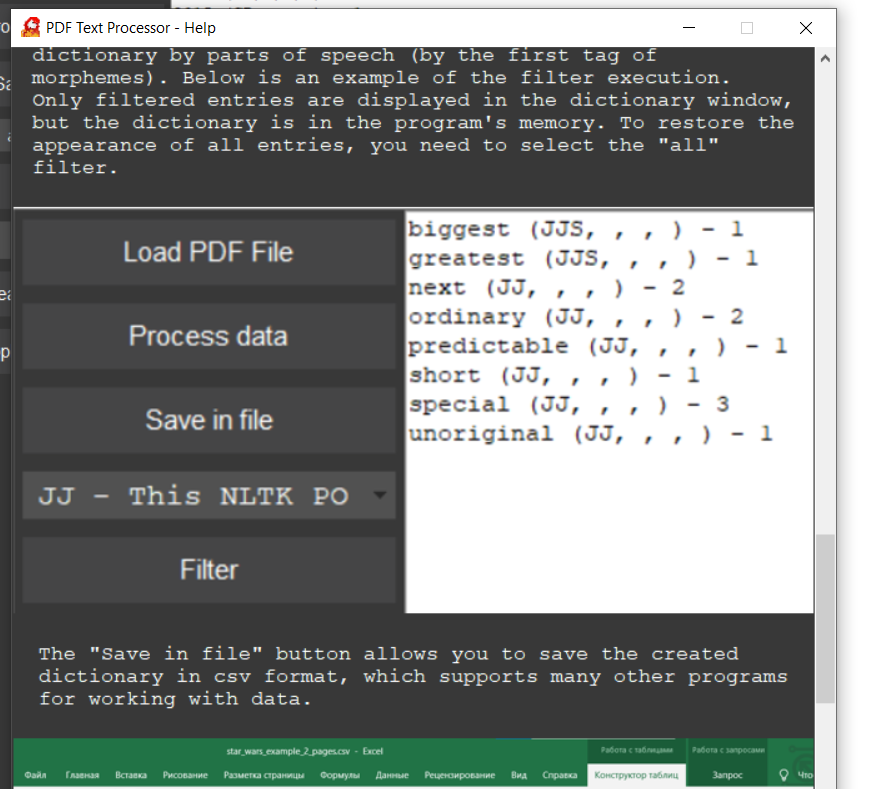


Рисунок 4. Окно помощи разработанного приложения.

# Оценка быстродействия приложения.

В приложении все основные функции выполняются за время O(n). Только операция сортировки выходного словаря выполняется за время O(n log n).

**O(n)** — линейная сложность.

Такой сложностью обладает, например, алгоритм поиска наибольшего элемента в не отсортированном массиве. Нам придётся пройтись по всем n элементам массива, чтобы понять, какой из них максимальный.

**O(n log n)** — логарифмическая сложность

~~Простейший пример — бинарный поиск. Если массив отсортирован, мы можем проверить, есть ли в нём какое-то конкретное значение, методом деления пополам. Проверим средний элемент, если он больше искомого, то отбросим вторую половину массива — там его точно нет. Если же меньше, то наоборот — отбросим начальную половину. И так будем продолжать делить пополам, в итоге проверим log n элементов.~~

Используется Timsort (Рисунок 5) - стандартная сортировка языка Python.

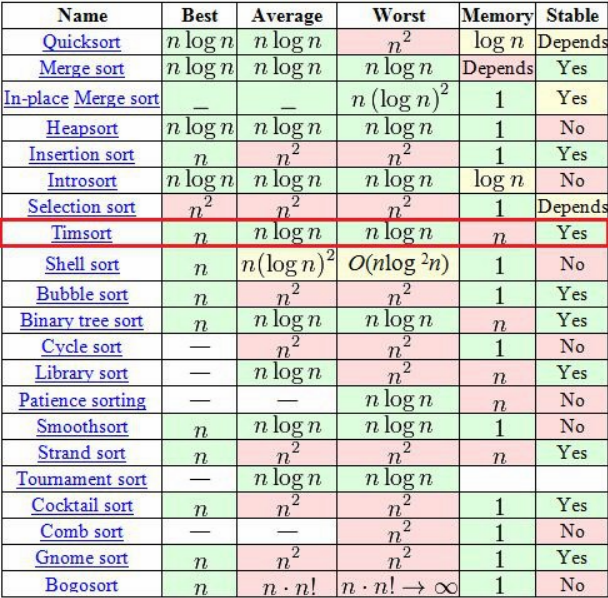


Рисунок 5. Сравнение быстродействия алгоритмов сортировки.

# [Вывод](#_h3i38hmtu3aj)

В ходе выполнения лабораторной работы были изучены назначение, структура и функциональность, предоставляемая базовым ЯП для решения задачи автоматического лексического и лексико-грамматического анализа ТЕЯ.

Результатом лабораторной работы является приложение для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка. Данное приложение позволяет проводить анализ ТЕЯ для получения словаря характерных лексем.