Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра интеллектуальных информационных технологий

Отчёт

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

Лабораторная работа №4

«Семантико-синтаксический анализ текстов естественного языка»

|  |  |
| --- | --- |
| Выполнили:  студенты гр. 221701: | Крупский А. В. Глёза Е. Д. |
| Проверил: | Крапивин Ю.Б. |

Минск, 2025

# **Цель работы:**

Освоить принципы разработки прикладных сервисных программ для решения задачи семантико-синтаксического анализа текста естественного языка.

# **Задание**:

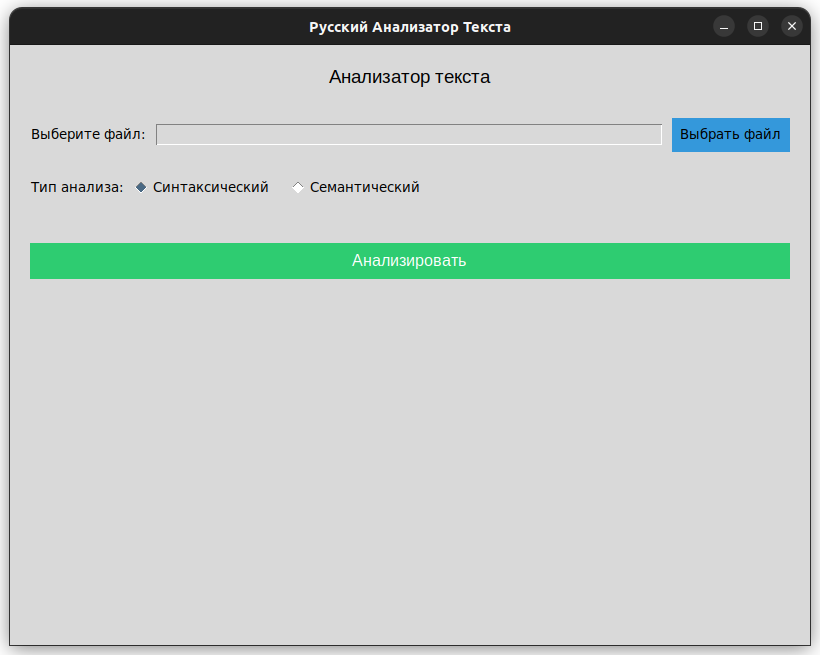
1. Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического семантико-синтаксического анализа ТЕЯ.
2. Закрепить навыки программирования при решении задач автоматической обработки ТЕЯ.

**Используемые бибилиотеки:**

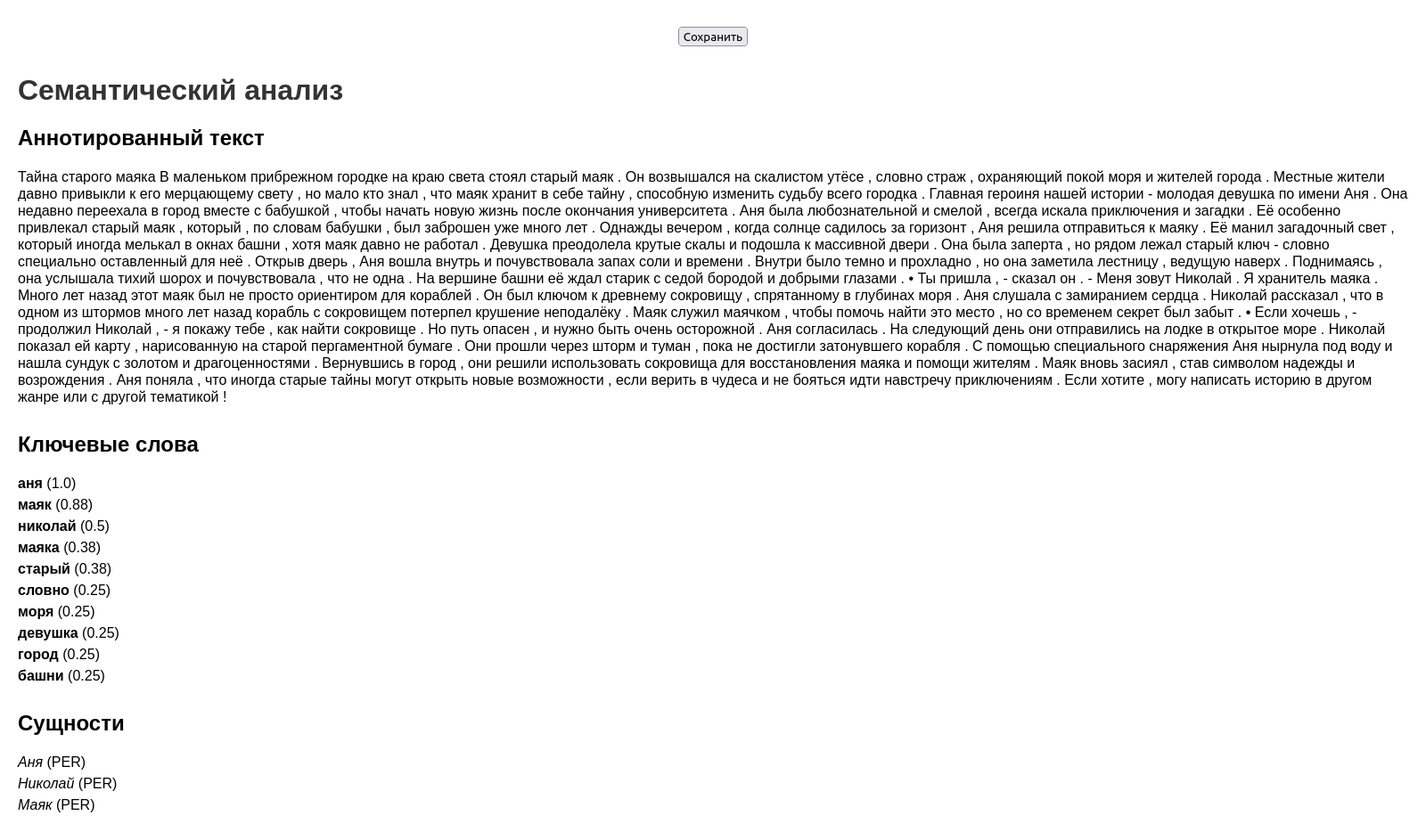
1. **natasha**  
   **Описание:** Библиотека для обработки естественного языка на русском языке.  
   **Использование:** Токенизация, лемматизация, извлечение имён, географических объектов и других сущностей, синтаксический и морфологический разбор русскоязычных текстов.
2. **pypandoc**  
   **Описание:** Обёртка для Pandoc — универсального конвертера документов.  
   **Использование:** Преобразование документов между множеством форматов (например, Markdown → DOCX, HTML → PDF), удобная автоматизация конвертации в Python-программах.
3. **setuptools**  
   **Описание:** Стандартный инструмент для упаковки и распространения Python-проектов.  
   **Использование:** Создание setup-файлов, управление зависимостями, сборка и публикация Python-библиотек на PyPI.
4. **stop-words**  
   **Описание:** Простая библиотека для получения списков стоп-слов на различных языках.  
   **Использование:** Фильтрация часто встречающихся, но малозначимых слов (например, "и", "в", "на") при обработке текста в задачах NLP.
5. **textract**  
   **Описание:** Библиотека для извлечения текста из различных типов документов.  
   **Использование:** Получение текста из PDF, DOC, DOCX, PPT, изображений и других форматов с помощью единого интерфейса (использует внешние зависимости, такие как Tesseract, antiword и др.).
6. **tkinter**

**Описание:** Стандартная библиотека Python для создания графических интерфейсов на основе Tcl/Tk.  
**Использование:** Разработка оконных приложений: создание кнопок, полей ввода, холстов, меню и других GUI-компонентов.

**Интерфейс:**

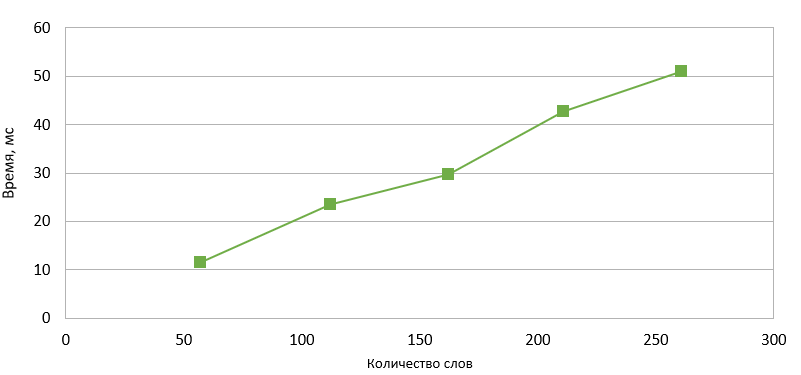


Окно программы

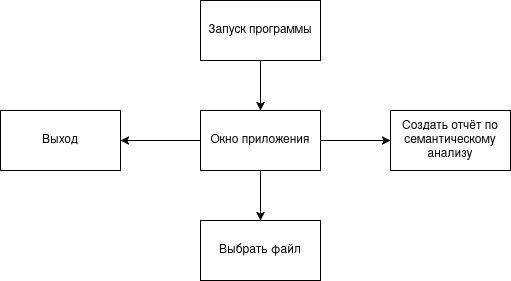


Страница отчёта анализа

**Тест производительности:**



## Структурная схема приложения



**Описание алгоритма:**

*Анализ текста:*

1. **Начало**
2. Получаем файл формата TXT/RTF
3. Считываем текст файла
4. Производим сегментацию текста
5. Производим морфологический анализ
6. Производим синтаксический анализ
7. Для каждого токена производим лемматизацию
8. Извлекаем именованные сущности
9. Отображаем результат в HTML-документе
10. **Конец**

**Вывод:**

Данная программа представляет собой инструмент семантико-синтаксического анализа предложений. Данная программа имеет перспективы в области машинного обучения или обработке текстов.