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 №3

Тема: Синтаксический анализ текстов естественного языка

Выполнили: Демидовец Д.В.

Козырев Д.А.

гр. 221703

Проверил: Крапивин Ю.Б.

Минск 2025

**Цель работы:**

Освоить принципы разработки прикладных сервисных программ для решения

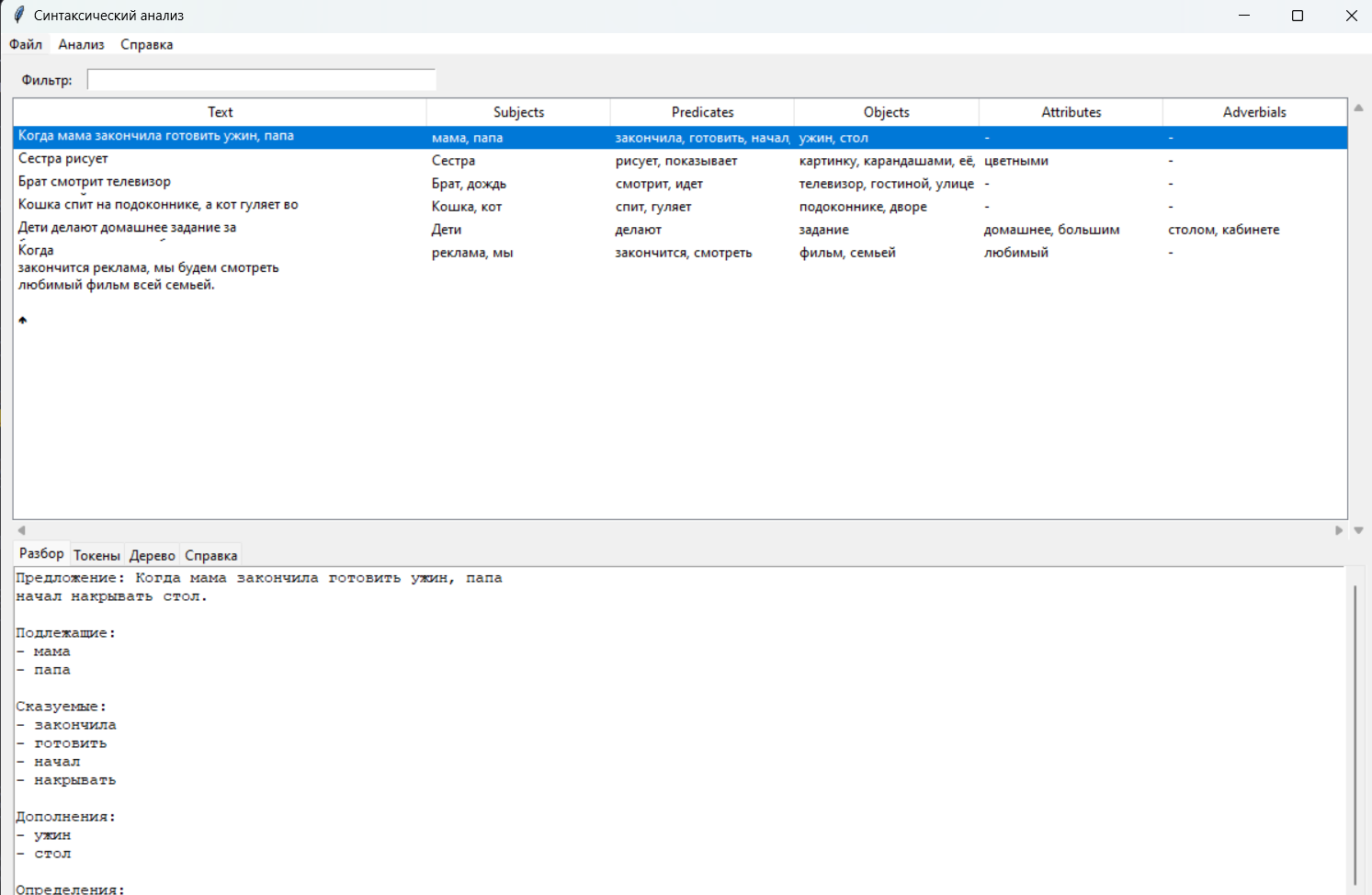
задачи анализа текста естественного языка.

**Задание:**

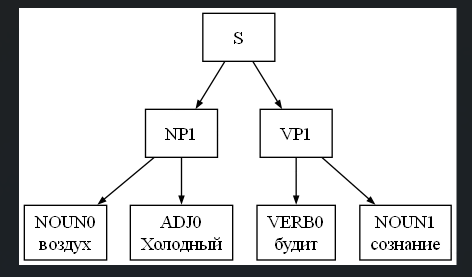
1. Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического синтаксического анализа ТЕЯ.

2. Закрепить навыки программирования при решении задач автоматической обработки ТЕЯ.

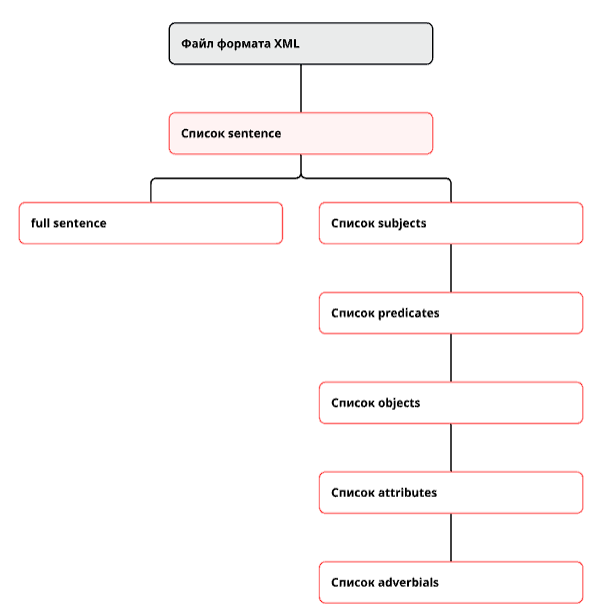
**Интерфейс:**

****

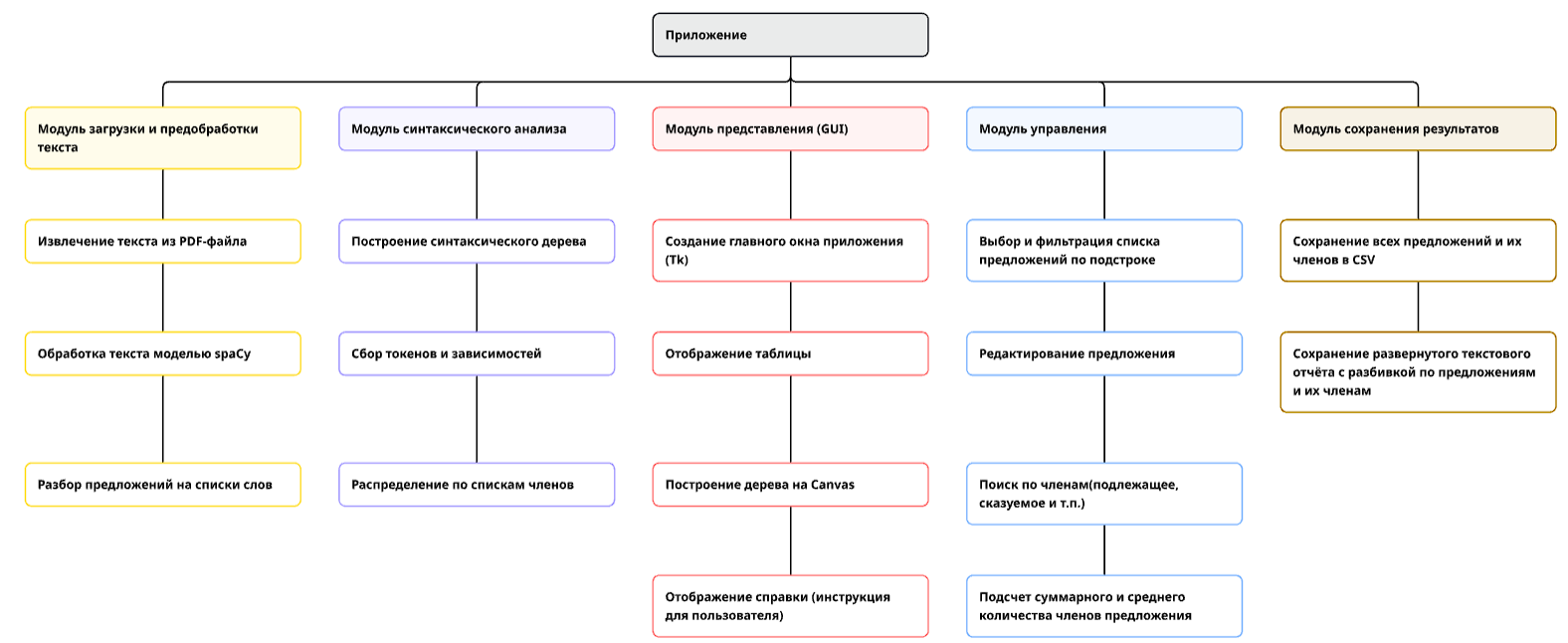
**Интерфейс синтаксического дерева предложения:**

****

**Структура хранения данных:**

****

**Структурная схема приложения:**

****

**Использованные библиотеки:**

* Библиотека **spaCy** применяется для глубокого лингвистического анализа текста. С её помощью: разбивается текст на токены и предложения, определяются части речи (POS), выявляются синтаксические зависимости и строится объект Doc, по которому далее происходит извлечение субъектов, предикатов, объектов и т.д.
* Библиотека **tkinter** используется для создания графического интерфейса пользователя. С её помощью реализованы окно программы, меню, кнопки, поля для ввода, таблица для отображения токенов, а также обработка пользовательских событий, например, двойной клик по строке таблицы для редактирования данных.
* Модуль **pdfminer.high\_level** используется для извлечения «сырого» текста из PDF-файла перед его передачей в spaCy. Это позволяет работать с PDF-документами без необходимости ручного копирования содержимого.
* Библиотека **xml**.etree.ElementTree (ET) применяется для создания и сохранения результатов синтаксического анализа в формате XML. Она позволяет экспортировать таблицу в виде структурированного файла с разделителями.
* Модуль **typing** (List, Optional) применяется для явного указания типов полей и возвращаемых значений, повышая читаемость и позволяя статическим анализаторам проверять корректность кода

**Описание алгоритмов:**

**1. Обработка текста из файла:**

Основной алгоритм для извлечения текста из PDF реализован в функции extract\_info:

1. Начало.
2. Пользователь выбирает PDF через диалоговое окно (tkinter.filedialog).
3. Функция extract\_text() из pdfminer.high\_level считывает текстовое содержимое файла.
4. Текст разбивается на строки, удаляются пустые и лишние пробелы.
5. Очищенный текст передаётся в NLP-пайплайн spaCy для дальнейшей обработки.
6. Конец.

**2. Синтаксическая NLP-обработка текста:**

1. Начало.
2. Очищенный текст (уже извлечённый из PDF) передаётся в языковую модель spaCy.
3. Создаётся объект doc, содержащий информацию о всём тексте.
4. С помощью встроенного механизма doc.sents текст разбивается на отдельные предложения.
5. Для каждого предложения выполняется:
   1. Токенизация (разделение на отдельные слова).
   2. Лемматизация (приведение слова к начальной форме).
   3. Определение частей речи.
   4. Анализ морфологических признаков (число, род, падеж и т.д.).
   5. Построение синтаксических зависимостей между словами.
6. Из каждого предложения выделяются подлежащее, сказуемое, дополнения и обстоятельства на основе синтаксических связей.
7. Сохранённые данные используются для отображения в таблице и дальнейшего анализа.
8. Конец.

**3. Построение синтаксического дерева:**

1. Начало.
2. Пользователь выбирает предложение из списка.
3. Построение иерархии:
   1. Строится дерево зависимостей, начиная с главного слова (ROOT).
   2. К каждому токену рекурсивно добавляются его зависимые слова.
4. Для корректного отображения вычисляются координаты каждого узла.
5. На холсте (Canvas) рисуются узлы и соединяющие их линии.
6. Отображаются слова и типы связей.
7. Дерево автоматически обновляется при выборе нового предложения.
8. Конец.

**4. Редактирование предложения**

1. Начало.
2. Пользователь двойным щелчком выбирает предложение в таблице.
3. Открывается новое окно tkinter, где текст можно изменить вручную.
4. Пользователь вносит изменения и нажимает "Сохранить".
5. Обновлённое предложение сохраняется в основной список.
6. Повторно выполняется синтаксическая обработка редактированного предложения.
7. В таблице и на холсте отображаются актуальные данные.
8. Закрытие окна редактирования.
9. Конец.

**Тест производительности:**

****

В ходе эксперимента было проведено измерение времени выполнения синтаксического анализа при обработке текстов с различным количеством предложений. Для тестирования использовался один исходный текст, из которого автоматически формировались выборки с увеличивающимся числом предложений: от 10 до 500.

Полученный график демонстрирует линейную зависимость: с ростом количества предложений увеличивается и общее время обработки текста. Такая тенденция объясняется тем, что для каждого предложения выполняется последовательная обработка — токенизация, определение частей речи, синтаксических зависимостей и других лингвистических признаков. Эти операции имеют линейную временную сложность, что приводит к пропорциональному увеличению времени анализа при увеличении объёма текста.

**Вывод:**

В ходе выполнения лабораторной работы было разработано приложение для синтаксического анализа текста на русском языке, позволяющее обрабатывать документ формата .pdf и автоматически извлекать информацию о каждом слове. В результате работы программа формирует словарь, содержащий предложения и слова, из которых они состоят, а также подлежащие, сказуемые и другие грамматические признаки. Разработанная структура хранения данных обеспечивает наглядное и удобное представление информации, что значительно упрощает последующую обработку и анализ текста. В перспективе разработанный функционал можно расширять и адаптировать для других языков, что увеличит область применения в образовательной, исследовательской и редакторской практике.