Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №1

По дисциплине: «Естественно-языковой интерфейс ИС»

Тема: «Разработка автоматизированной системы формирования словаря естественного языка»

Выполнил:

Студент 3 курса

Группы ИИ-21

Карагодин Д. Л.

Проверила:

Якимук А. В.

Брест 2024

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

Ход работы:

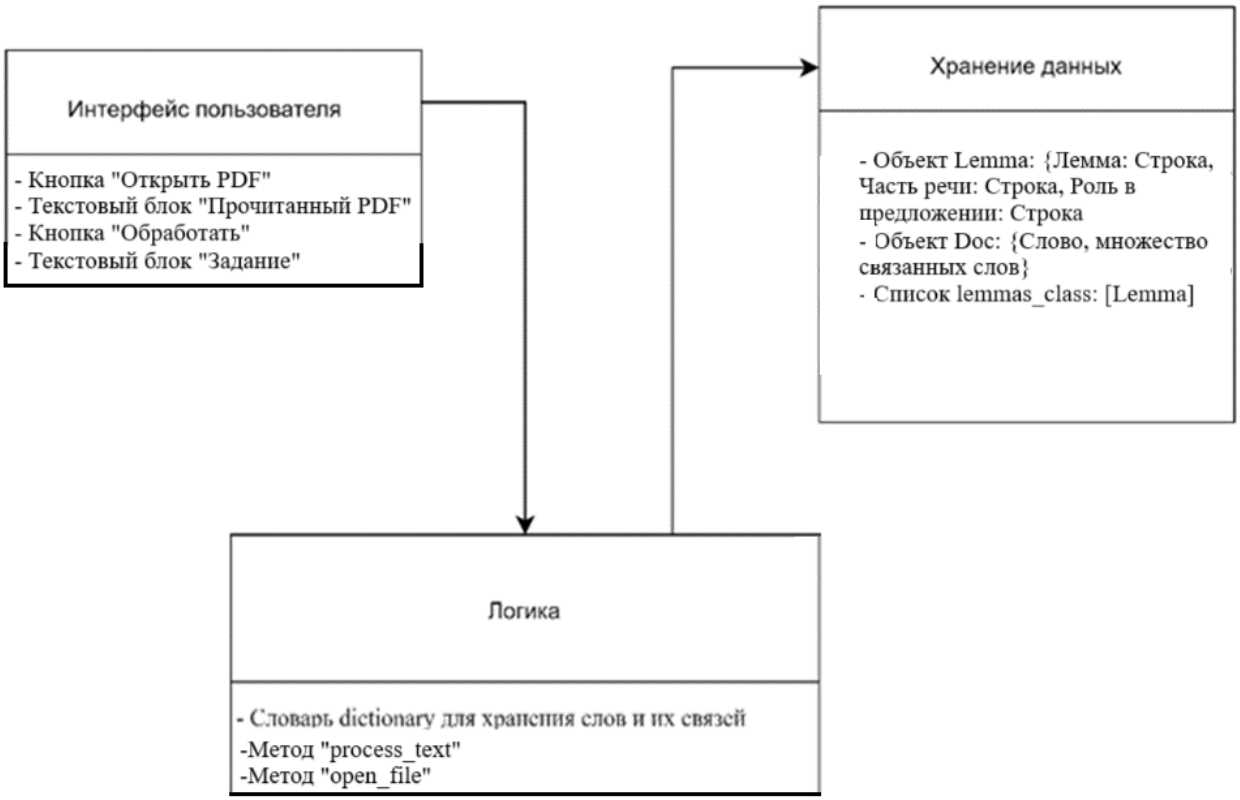
№	Язык текста	Формат входного документа	Вариант задания
11	Русский	PDF	Задание 4

Задание 4. 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 какой части речи. 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.

Методические указания:

Требуется спроектировать и программно реализовать структуры хранения данных, алгоритмы их обработки, необходимые в рамках следующих базовых требований к разрабатываемому приложению:

- входные данные – текст заданного естественного языка;
- выходные данные – перечень лексем с дополнительной информацией согласно заданию;
- взаимодействие с пользователем посредствам графического интерфейса (интерфейс должен быть интуитивно-понятным и дружественным пользователю);
- наличие системы средств помощи пользователю;
- обеспечение возможности построения, сохранения, просмотра, редактирования, пополнения, фильтрации и поиска по заданному условию, документирования автоматически получаемого словаря либо заданной его части;
- поддержка форматов представления входных данных (TXT, RTF, PDF, DOC, DOCX).

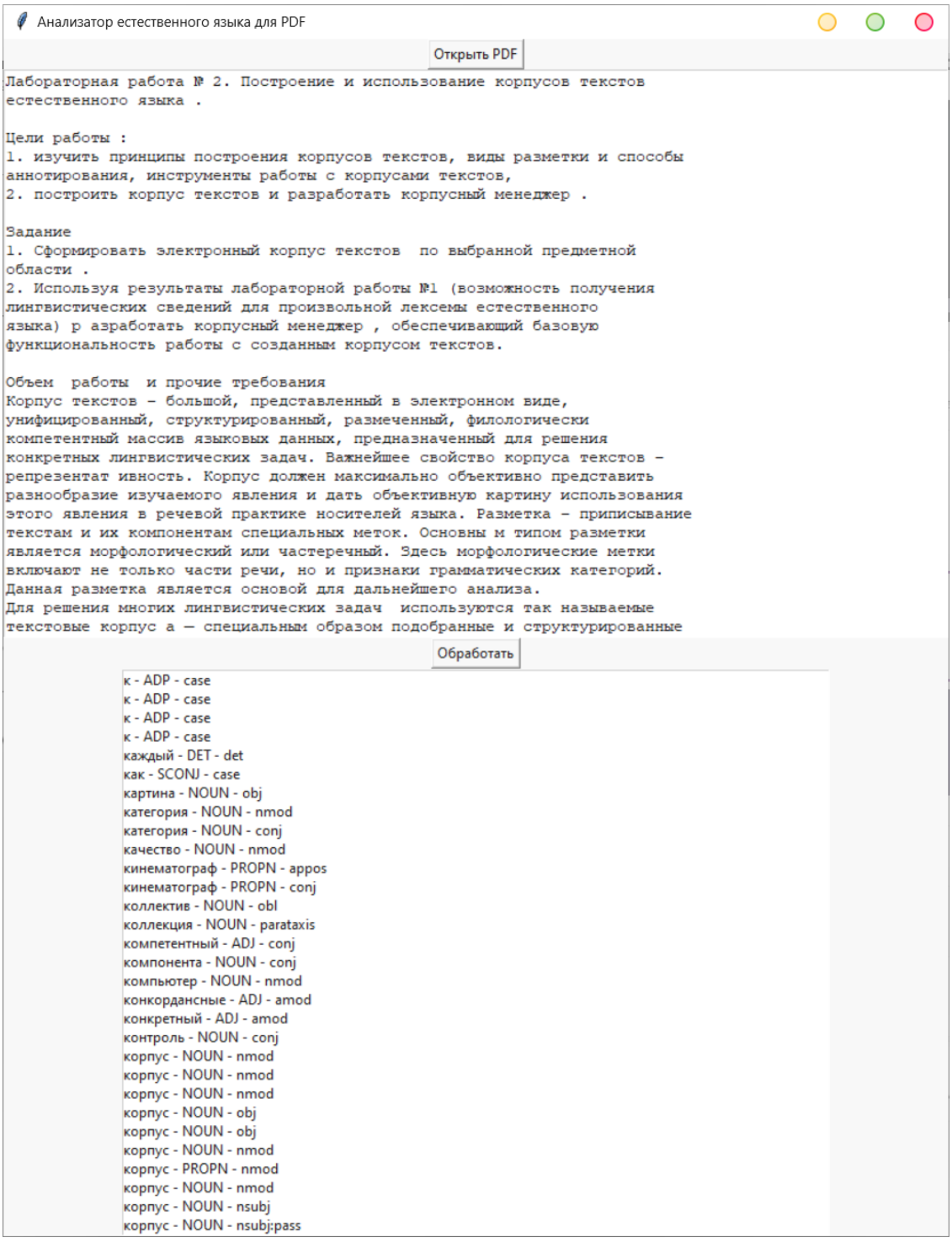


Листинг:

```
import spacy
import PyPDF2
from tkinter import *
import tkinter.filedialog
# Класс для хранения информации о лексемах
class Lemma:
    def __init__(self, lemma,morphology="",role=""):
        self.lemma:str = lemma
        self.morphology:str = morphology
        self.role:str = role
    def __iter__(self):
        return iter(self.lemma)
       def set_morphology(self, tag:str):
        self.morphology = tag
    def set_role(self, role):
        self.role = role
# Обработка текста
def process_text():
    # Извлечение текста из PDF
    with open(file_path, 'rb') as f:
        pdf_reader = PyPDF2.PdfReader(f)
        text = ""
        for page in pdf_reader.pages: text+=page.extract_text()
    # Морфологический и Синтаксический анализ
    nlp = spacy.load('ru_core_news_sm')
    doc = nlp(text)
    lemmas_class: list[Lemma] = []
    for token in doc:
        lemmas_class.append(Lemma(token.lemma_, token.pos_, token.dep_))
    lemmas_class.sort(key=lambda x: x.lemma.lower())
    # Вывод результатов
    result_list.delete(0, END)
    for lemma in lemmas_class:
        result_list.insert(END, f"{lemma.lemma} - {lemma.morphology} - {lemma.role}")
# Открытие PDF-документа
def open_file():
    global file_path
    file_path = tkinter.filedialog.askopenfilename(filetypes=[("PDF files", "*.pdf")])
    if file_path:
        # Извлечение текста из PDF
        with open(file_path, 'rb') as f:
            pdf_reader = PyPDF2.PdfReader(f)
            text = ""
            for page in pdf_reader.pages: text+=page.extract_text()
        # Отображение извлеченного текста
        text_input.delete('1.0', END)
        text_input.insert('1.0', text)
# Графический интерфейс
root = Tk()
root.title("Анализатор естественного языка для PDF")
# Кнопка для открытия PDF-документа
open_file_button = Button(root, text="Открыть PDF", command=open_file)
open_file_button.pack()
# Текстовое поле для отображения извлеченного текста
text_input = Text(root, width=100, height=30)
text_input.pack()
# Кнопка для запуска обработки
process_button = Button(root, text="Обработать", command=process_text)
process_button.pack()
```

```
# Область вывода для отображения списка лексем с дополнительной информацией
result_list = Listbox(root,width=100,height=30)
result_list.pack()
root.mainloop()
```

Результат:



Вывод: в ходе выполнения лабораторной работы освоил принципы разработки прикладных сервисных программ для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка.