Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №2

По дисциплине: «Естественно-языковой интерфейс ИС»

# Тема: «Построение и использование корпусов текстов естественного языка»

Выполнил:

Студент 3 курса

Группы ИИ-21

Романко Н. А.

Проверила:

Якимук А. В.

Брест 2024

**Цель:** 1. изучить принципы построения корпусов текстов, виды разметки и способы аннотирования, инструменты работы с корпусами текстов, 2. построить корпус текстов и разработать корпусный менеджер.

**Ход работы:**

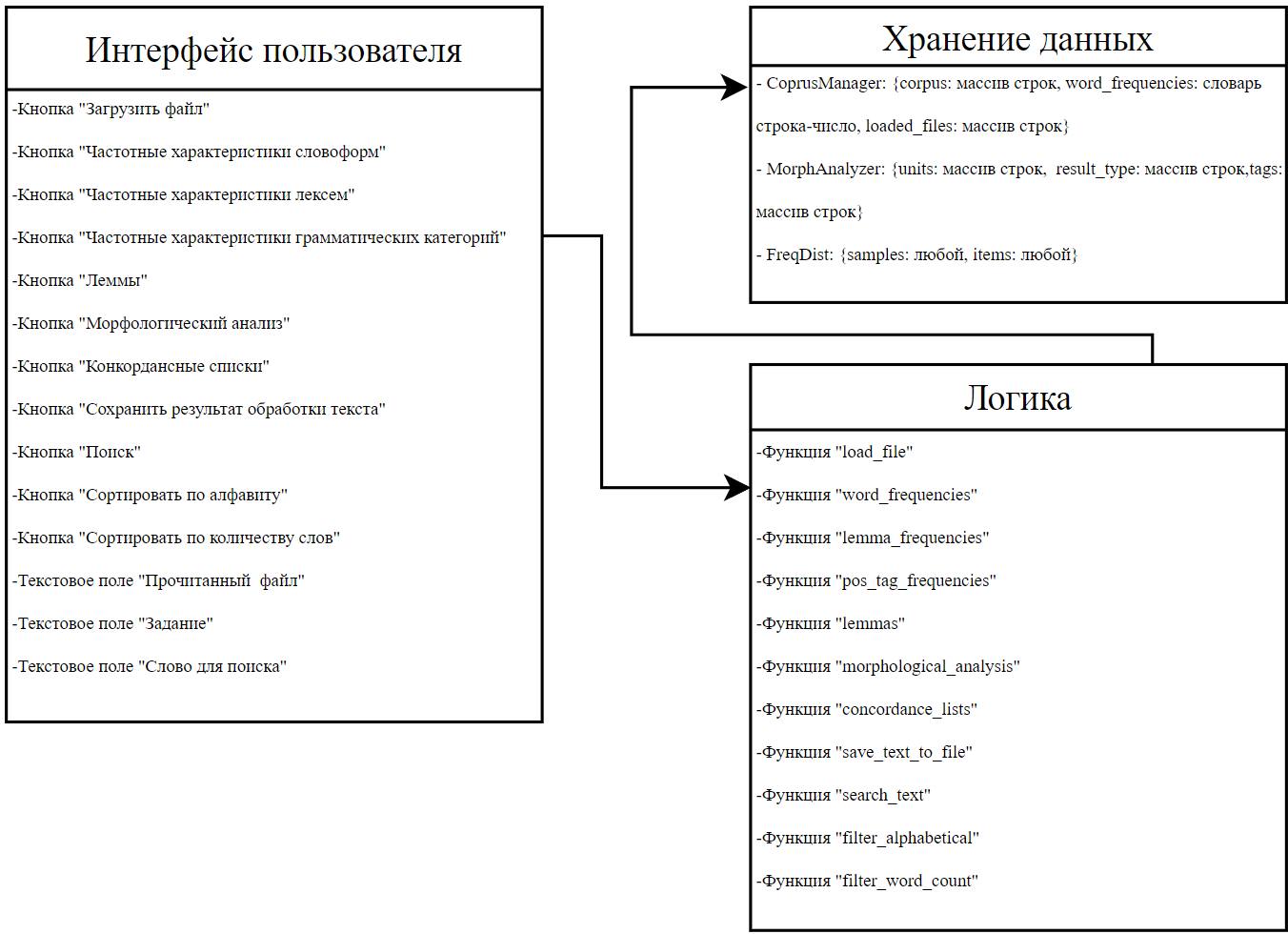
|  |  |  |
| --- | --- | --- |
| **Номер**  **варианта** | **Язык текста** | **Предметная область** |
| 14 | Английский | Кулинария |

**Методические указания:**

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

* входные данные – фрагмент текста (фраза или слово) на естественном языке – запрос корпусному менеджеру;
* выходные данные – частотные характеристики словоформ, лексем, грамматических категорий, леммы, морфологические характеристики словоформ и их метаданные (библиографические, типологические), конкордансные списки, согласно согласованным с преподавателем требованиям к функциональности корпусного менеджера;
* 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 текста и/или его фрагмента в соответствии с реализуемой функциональностью корпусного менеджера;
* поддержка различных форматов представления входных данных (TXT, RTF, PDF, DOC, DOCX).

**Структурно-функциональная схема:**



**Листинг:**

import nltk

from nltk.tokenize import word\_tokenize

from nltk.probability import FreqDist

from nltk.corpus import wordnet as wn

import tkinter as tk

from tkinter import ttk

from tkinter import scrolledtext

word\_freq\_dist = None

def analyze\_text():

global word\_freq\_dist

input\_text = text\_entry.get("1.0",'end-1c')

tokens = word\_tokenize(input\_text)

pos\_tags = nltk.pos\_tag(tokens)

lemmatizer = nltk.WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(token) for token in tokens]

word\_freq\_dist = FreqDist(tokens)

output\_text.delete(1.0, tk.END)

output\_text.insert(tk.END, "Частотные характеристики словоформ:\n")

output\_text.insert(tk.END, f"{word\_freq\_dist}\n\n")

output\_text.insert(tk.END, "Частеречная разметка:\n")

output\_text.insert(tk.END, f"{pos\_tags}\n\n")

output\_text.insert(tk.END, "Лемматизация:\n")

output\_text.insert(tk.END, f"{lemmatized\_tokens}\n\n")

output\_text.insert(tk.END, "Морфологические характеристики и метаданные:\n")

for token in tokens:

synsets = wn.synsets(token)

if synsets:

output\_text.insert(tk.END, f"Слово: {token}\n")

for synset in synsets:

output\_text.insert(tk.END, f" Определение: {synset.definition()}\n")

output\_text.insert(tk.END, f" Примеры: {', '.join(synset.examples())}\n")

output\_text.insert(tk.END, f" Часть речи: {synset.pos()}\n\n")

output\_text.insert(tk.END, "-----------------------------------------------------\n")

root = tk.Tk()

root.title("Кулинарный анализатор по Американски")

label = ttk.Label(root, text="Введите текст для анализа:")

label.grid(row=0, column=0, padx=10, pady=10)

text\_entry = scrolledtext.ScrolledText(root, width=40, height=10)

text\_entry.grid(row=1, column=0, padx=10, pady=10)

analyze\_button = ttk.Button(root, text="Анализировать", command=analyze\_text)

analyze\_button.grid(row=2, column=0, padx=10, pady=10)

output\_text = scrolledtext.ScrolledText(root, width=80, height=20)

output\_text.grid(row=3, column=0, padx=10, pady=10, columnspan=2)

root.mainloop()

# Результат:

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Автоматически созданное описание

**Вывод:** в ходе выполнения лабораторной работы освоил принципы построения корпусов текстов, виды разметки и способы аннотирования, инструменты работы с корпусами текстов.