Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления

Кафедра Интеллектуальных информационных технологий

**ОТЧЁТ**

**По лабораторной работе № 1**

по дисциплине «Естественно-языковой интерфейс интеллектуальных систем»

по теме: “Разработка автоматизированной системы формирования словаря естественного языка”

Выполнил: Самута Д.В.

Семенов Е. Г.

гр. 221703

Проверил: Крапивин Ю. Б.

Минск 2025

**Цель работы:**

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

**Задание:**

Список слов, упорядоченный по алфавиту и включающий как лексемы, так и словоформы, с указанием частоты встречаемости каждой из форм. Для словоформ пользователю должна быть предоставлена возможность вводить дополнительную морфологическую информацию, а именно, отнесение слова к соответствующей части речи, указание рода, числа, падежа и т.п. При этом морфологическая информация может быть оформлена как отдельная неформатированная запись, т.е. это просто текст, который пользователь может оформлять произвольным образом.

**Интерфейс:**

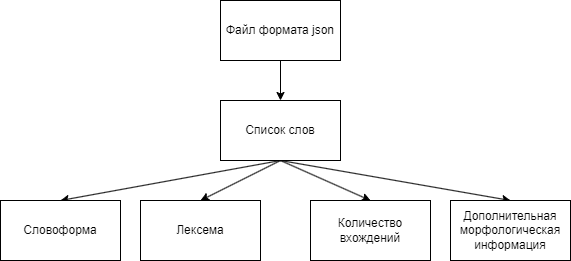
****

**Тест производительности:**

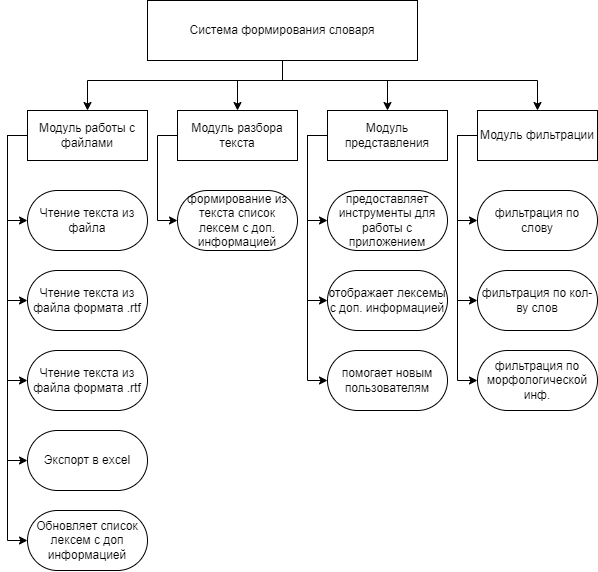
****

На данном графике показана зависимость времени обработки текста от количества слов. Видно, что с увеличением количества слов время обработки также возрастает. Это указывает на **линейную зависимость** между количеством слов и временем обработки, хотя есть небольшие отклонения. Общая тенденция — увеличение времени обработки при увеличении объема текста.

**Структура хранения данных:**



**Структурная схема приложения:**

****

**Использованные библиотеки:**

**spaCy** — для обработки текста, включая токенизацию, лемматизацию и определение частоты слов. Библиотека используется в функции get\_info\_from\_text, где с её помощью анализируется текст и извлекаются данные о словах, их лексемах и частоте.

**rtf\_to\_text** (из библиотеки striprtf) — для преобразования содержимого RTF-файлов в текст. Используется в функции read\_text\_from\_file, чтобы поддерживать обработку файлов формата .rtf наряду с .txt.

**openpyxl -** для работы с файлами Excel (.xlsx). Она позволяет читать и записывать данные в формат Excel, а также модифицировать существующие файлы.

**tkinter** - для создания графических пользовательских интерфейсов (GUI). Она предоставляет инструменты для создания окон, кнопок, полей ввода, таблиц и других элементов интерфейса.

**textwrap** - для работы с текстовыми строками, предназначенная для обёртки текста в заданной ширине, чтобы текст не выходил за пределы заданной ширины строки

**Описание алгоритмов:**

**1. Алгоритм поиска слов в файле:**

1. Пользователь выбирает файл с помощью диалогового окна.
2. Проверяется формат файла:
   * Если это .txt, текст считывается стандартным способом.
   * Если это .rtf, текст преобразуется в обычный с помощью rtf\_to\_text.
3. Полученный текст анализируется библиотекой spaCy:
   * Текст токенизируется (разбивается на отдельные слова и символы).
   * Каждый токен проверяется регулярным выражением, чтобы определить, является ли он русским словом.
   * Из отфильтрованных токенов создаётся словарь, содержащий:
     + Слово (как оно есть в тексте),
     + Лемму (основная форма слова),
     + Количество вхождений,
     + Дополнительную информацию (если есть).
4. Сформированный список слов возвращается для дальнейшей обработки.

**2. Алгоритм объединения данных из нового анализа и существующего файла:**

1. Загружаются данные из файла word\_data.json, если он существует.
2. Новый список слов из анализа текста сравнивается с уже загруженными данными:
   * Если слово уже присутствует, его счётчик увеличивается на значение из нового анализа.
   * Если слово отсутствует, оно добавляется как новая запись.
3. Результирующий список объединённых данных сохраняется обратно в файл word\_data.json, чтобы обновить существующие данные.

**3. Алгоритм отображения данных в графическом интерфейсе:**

1. Очищается текущее содержимое таблицы (древовидного виджета).
2. Каждая запись из обновлённого списка данных добавляется в таблицу:
   * Столбцы содержат информацию о слове, лемме, количестве вхождений и дополнительной морфологической информации.
3. Если пользователь вносит изменения (например, добавляет морфологическую информацию), данные обновляются в файле и повторно отображаются в таблице.

**Вывод:**

Данная программа представляет собой инструмент для создания списка словоформ, включающих как лексемы, так и словоформы с указанием частоты их встречаемости и их морфологическими характеристиками. Она также предоставляет возможность пользователю редактировать морфологическую информацию для словоформ. Данная программа имеет

перспективы в области анализа текстов и книг для подведения различных статистик.