Министерство образования Республики Беларусь

Учреждение образования

“Белорусский государственный университет информатики и радиоэлектроники”

Кафедра информационных интеллектуальных технологий

Лабораторная работа 2

“Построение и использование корпусов текстов естественного языка.”

Выполнил Мулярчик Д.С.

гр.121701 Лемантович Д.К.

Проверил Крапивин Ю. Б

Минск 2024

Задание: Сформировать электронный корпус текстов по выбранной предметной области. Используя результаты лабораторной работы №1 (возможность получения лингвистических сведений для произвольной лексемы естественного языка) разработать корпусный менеджер, обеспечивающий базовую функциональность работы с созданным корпусом текстов.

Используемые инструменты: Python с PyQt и Natural Language Toolkit.

Структуры хранения: TXT- и RTF-файлы

Структурно-функциональная схема

Структурно-функциональная схема приложения представлена чёрным ящиком:

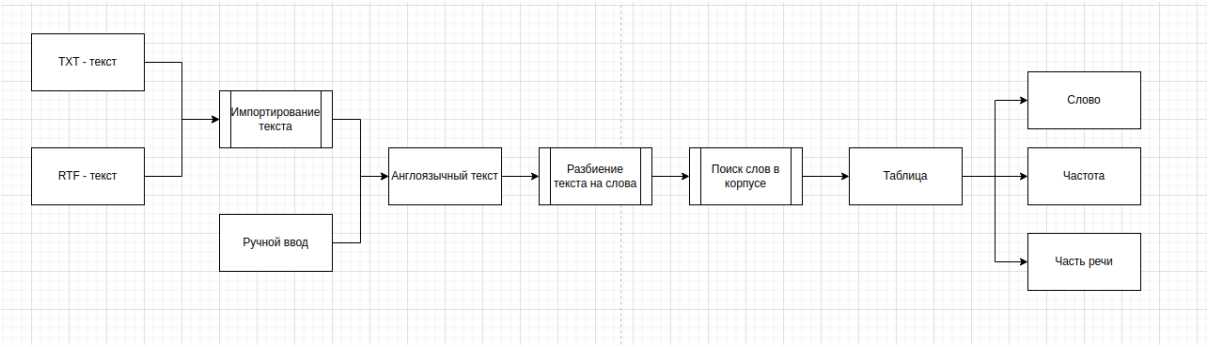

Рис.1 - Схема

Структуры хранения

После поиска слов в корпусе мы получаем словарь(ключ-значение), где ключом является лексема/словоформа, а значение другой словарь из двух ключей: “частота” и “дополнительная информация”. При нажатии на кнопку сохранить результат мы вводим имя файла и весь наш словарь записывается в удобном формате JSON, который очень удобно использовать в дальнейшем.

Размеченный корпус представляет набор XML файлов, где наш текст разбивается на предложения, они содержаться в теге <sentence>, слова в предложении содержатся в тегах <word> и имеет дополнительную морфологическую информацию.

Алгоритм обработки

1. Вход: текст на естественном языке.
2. Вызов функции **main(text)**, где text - текст на естественном языке.
3. Разбиение текста на слова по пробелам.
4. Помещаем все слова, которые не являются стоп-словами в список words.
5. Загружаем корпус текста в переменную corpus.
6. Берем **i (i=0)** элемент в списке words.
7. Поиск i элемента в corpus.
8. Записываем всю информацию в **words_with_info**.
9. Увеличиваем **i** на 1, если **i** > количества слов переходим к пункту 11.
10. Переходим к пункту 6.
11. Возвращаем итоговый словарь.
12. Выход: словарь(ключ-значение), где ключом является лексема/словоформа, а значение другой словарь из двух ключей: “частота” и “дополнительная информация”.

Алгоритм фильтрации и поиска

1. Вход: флаг, который указывает тип фильтрации/поиска, данный - словарь, который получен после обработки текста, запрос - в случае фильтрации по слову или части речи - строка, в случае частоты - максимальное и минимальное значение.
2. Динамическое создание функций фильтрации по типу флага.
    a. Если флаг равен “word”, поиск осуществляется по словам с помощью встроенной функции **filter(lambda x: search_type in x[0], data.items())**. Проходимся по словам в словаре и проверяем вхождение строки запроса в слово.
    b. Если флаг равен “frequency”, поиск осуществляется по словам с помощью встроенной функции"frequency": **filter(lambda x: frequency[0] <= x[1]["frequency"] and frequency[1] >= x[1]["frequency"], data.items())**. Проходимся по частоте в словаре и смотрит, чтобы частота не выходила запределы рамок.
    c. Если флаг равен “extra information”, поиск осуществляется по словам с помощью встроенной функции **"extra information": filter(lambda x: search_type in x[1]["additional information"], data.items())**. Проходимся по частям речи в словаре и проверяем вхождение строки запроса в части речи.
3. Получаем объект класса **filter** и преобразуем всё в словарь
4. Возвращаем пользователю
5. Выход: словарь(ключ-значение), где ключом является лексема/словоформа, а значение другой словарь из двух ключей: “частота” и “дополнительная информация”.

Пример работы приложения

1. При запуске приложения нас встречает интерфейс:

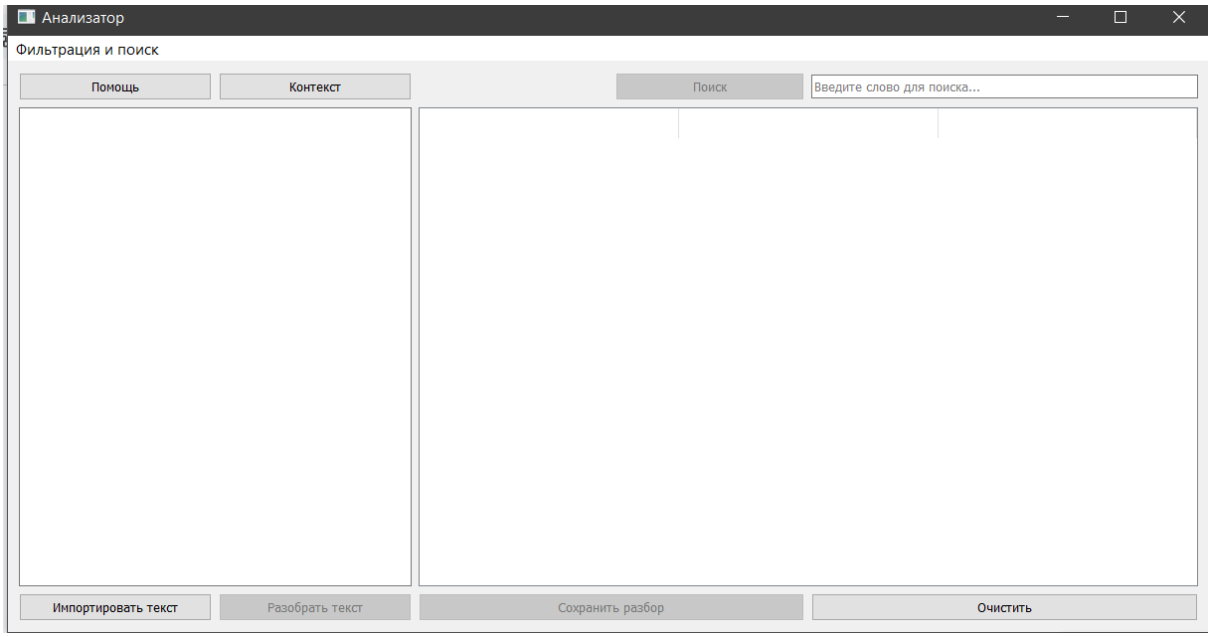

Рис.2 - Интерфейс

2. В него можно написать текст или импортировать текст из TXT- или RTF-файла, нажав на кнопку «Импортировать файл». Пример текста:

   Have you ever thought about what your future life is going to be like? What are you going to do when you finish school? It is never too early or late to start thinking about your future career. Maybe you enjoy some of the subjects at school more than others. If you do, this is a good sign, because they will guide you to your future profession.

3. Пример разбора данного текста после нажатия кнопки «Разобрать текст»:

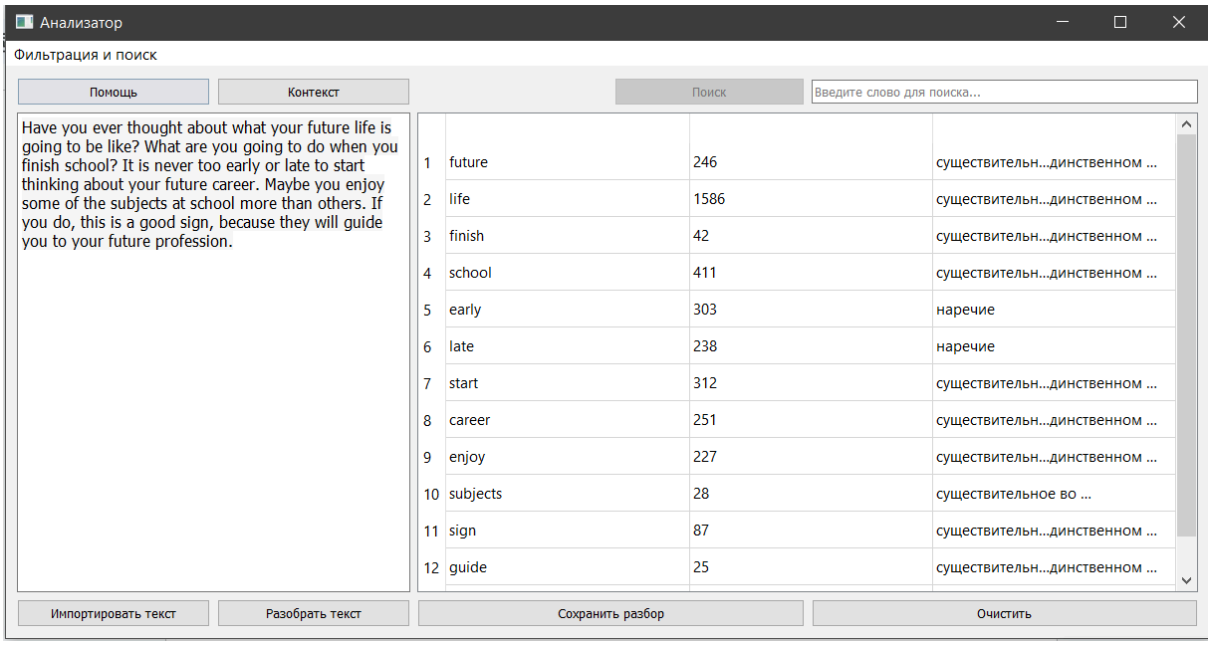

Рис.4 - Пример разбора

4. При нажатии на кнопку “Фильтрация и поиск” мы можем выбрать критерий по которому будет производится фильтрация и поиск: “Поиск по словам”, “Фильтрация по частоте”, “Фильтрация по дополнительной информации”

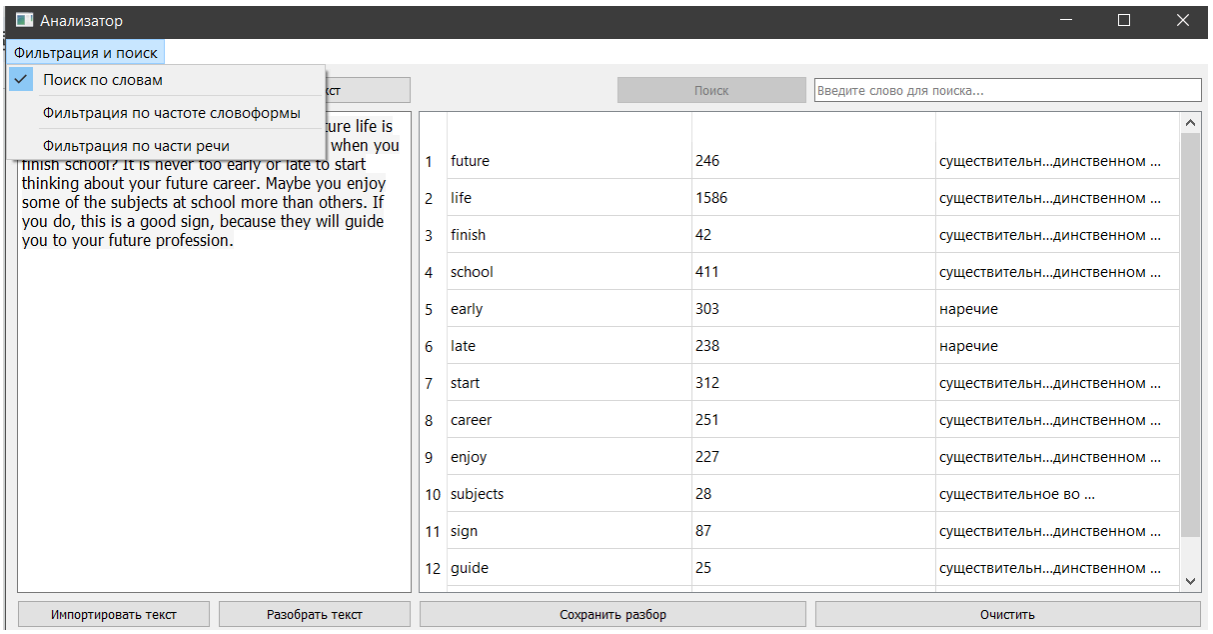

Рис.5 - Пример поиска

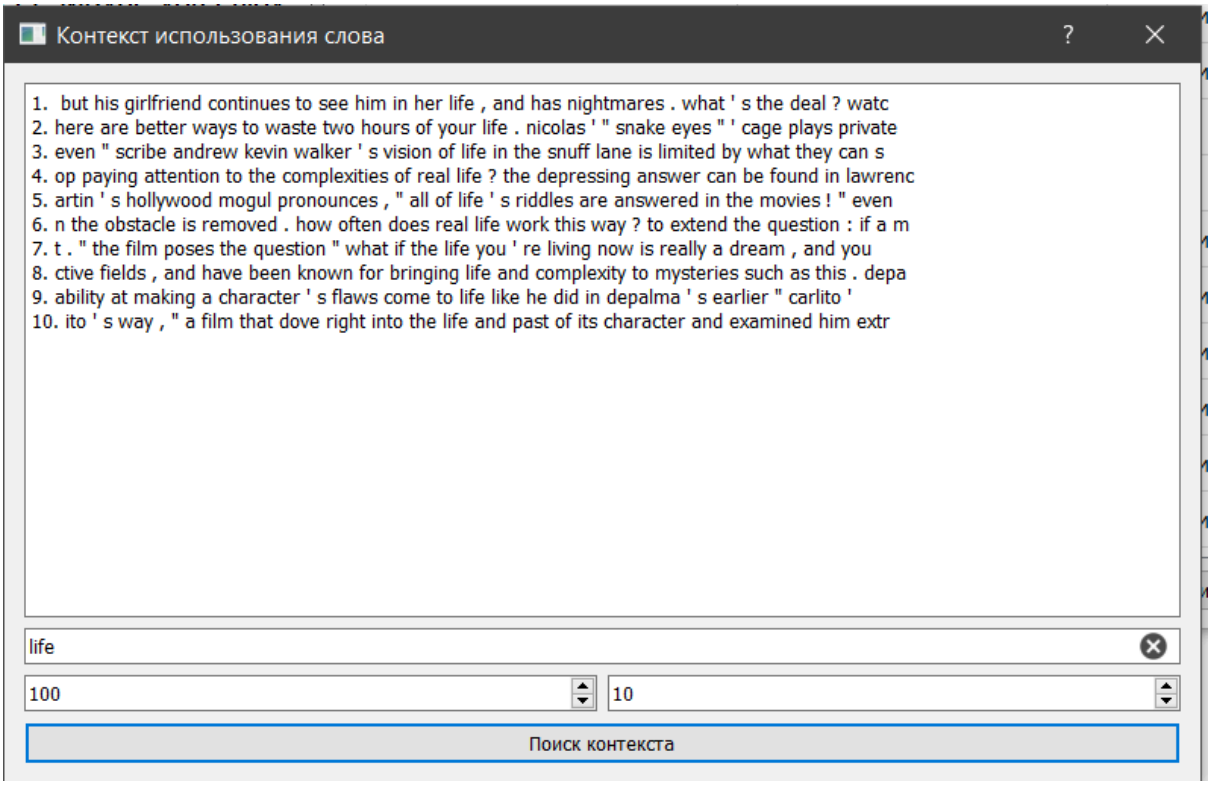

Рис. 6 - Пример поиска контекста