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 №2

“Построение и использование корпусов текстов естественного языка”

Выполнили студент группы 221702: Целуйко Д.А.

Гринь Н.А.

Проверил: Крапивин Ю.Б.

Минск 2025

Лабораторная работа №2

**Цель работы:**

1. Изучить принципы построения корпусов текстов, виды разметки и способы аннотирования, инструменты работы с корпусами текстов.
2. Построить корпус текстов и разработать корпусный менеджер.

**Задание (Вариант 13):**

1. Сформировать электронный корпус текстов по выбранной предметной области (кинематограф, язык корпуса - английский).
2. 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.

**Язык текста:**

Английский

**Предметная область:**

Кинематограф

**Формат входного документа:**

TXT, RTF, PDF, DOC, DOCX

**Используемые средства разработки:**

Для решения поставленной задачи было разработано десктоп-приложение на языке программирования **Python** с использованием библиотеки **Tkinter**, а также некоторые другие библиотеки

**Tkinter —** это пакет для Python, предназначенный для работы с библиотекой Tk. Библиотека Tk содержит компоненты графического интерфейса пользователя (graphical user interface – GUI). Эта библиотека написана на языке программирования Tcl.

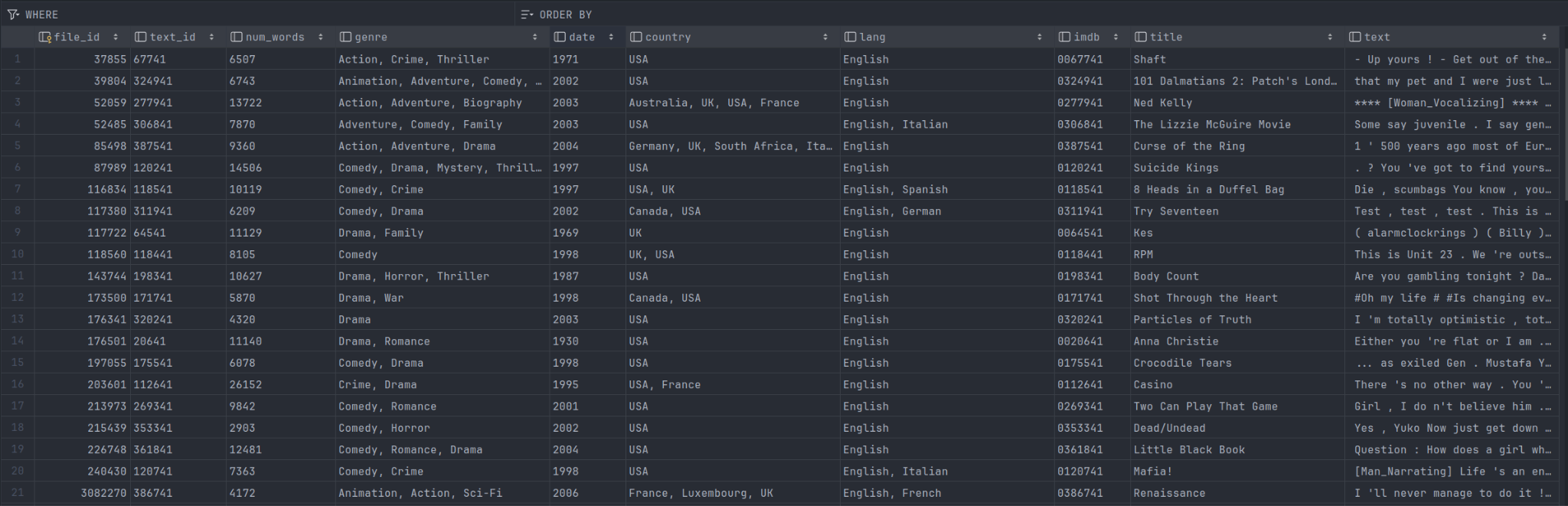
**Spacy (spaCy)** — это программная библиотека с открытым исходным кодом для расширенной обработки естественного языка, написанная на языках программирования Python и Cython. Были догружены следующие ресурсы:

* **spacy.load('en\_core\_web\_sm') -** легкая английская языковая модель, оптимизированная для использования центральным процессором. Эта модель позволяет выполнять присвоение частей речи, синтаксический анализ зависимостей и лемматизацию токенов во входном тексте. Разработано компанией Explosion AI.

**Json** — используется в качестве стандартного текстового формата для хранения и извлечения структурированных данных, таких как информация об анализируемых словах и их морфологические данные.

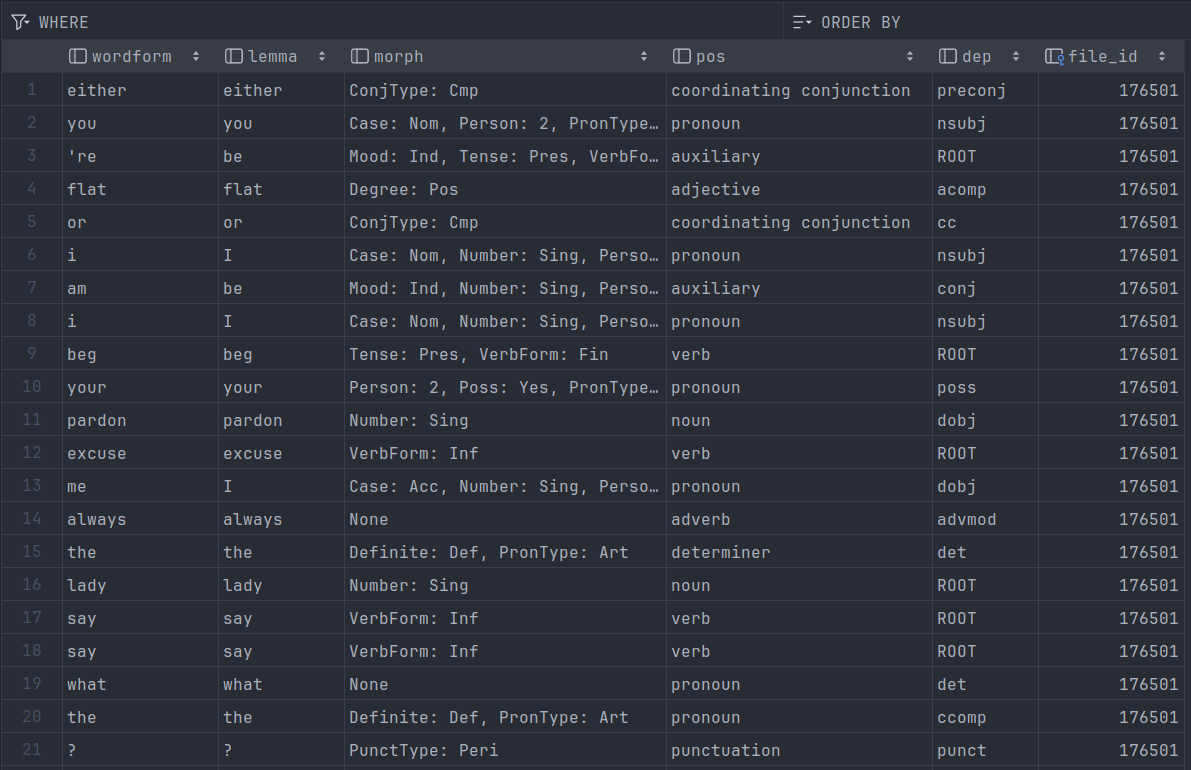
Для хранения всей необходимой информации были использованы две таблицы:

* **texts -** хранит такую информацию о фильмах как: количество слов, жанр, год, страна, язык, id на [IMDb](https://www.imdb.com/), название, контент фильма.



*Рис. 1 - таблица* ***texts***

* **wordforms** - хранит в себе такую информацию о словоформах как: словоформа, лемма, морфологическая информация, часть речи, зависимость, id фильма)

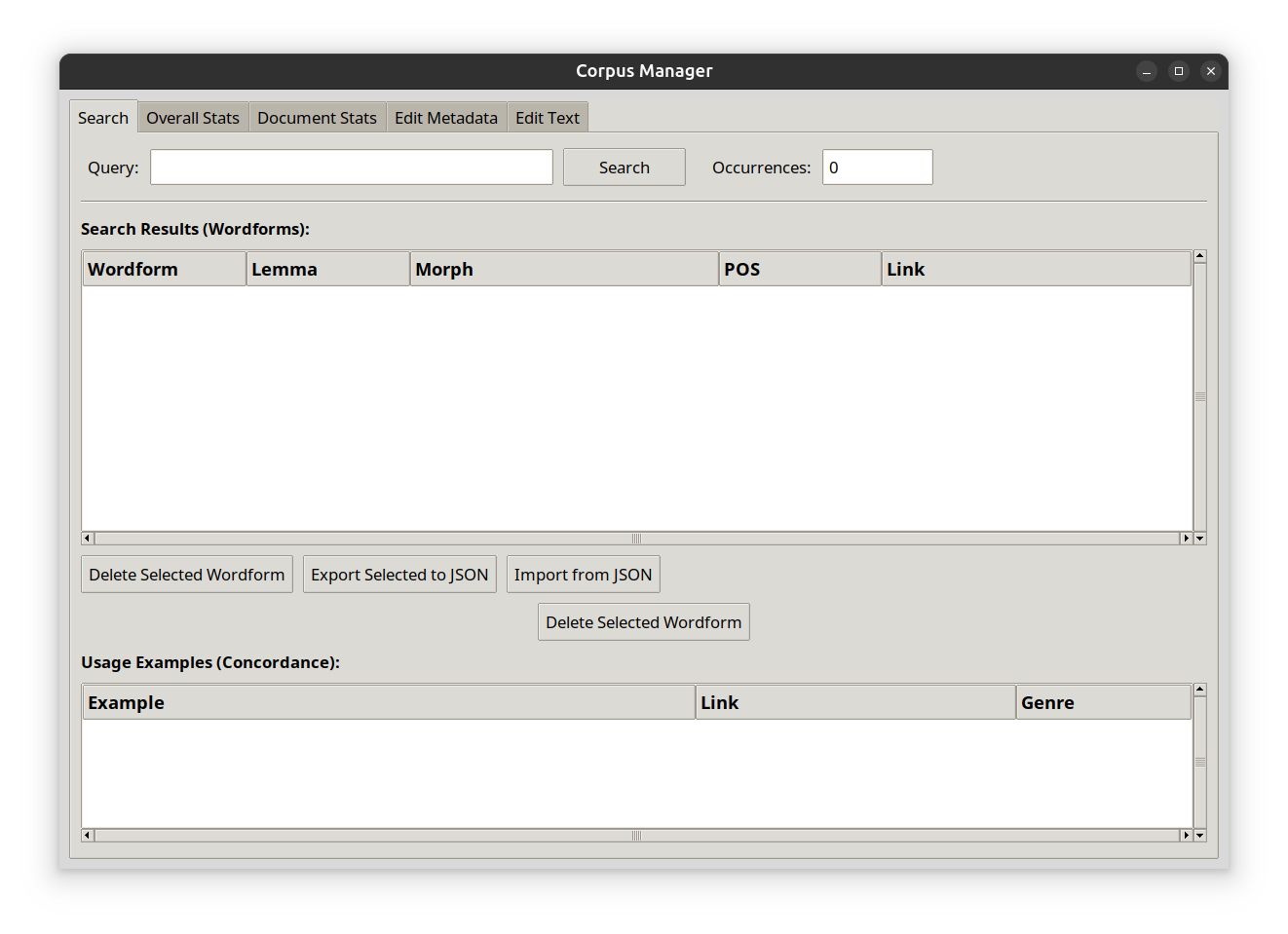


*Рис. 2 - таблица* ***wordforms***

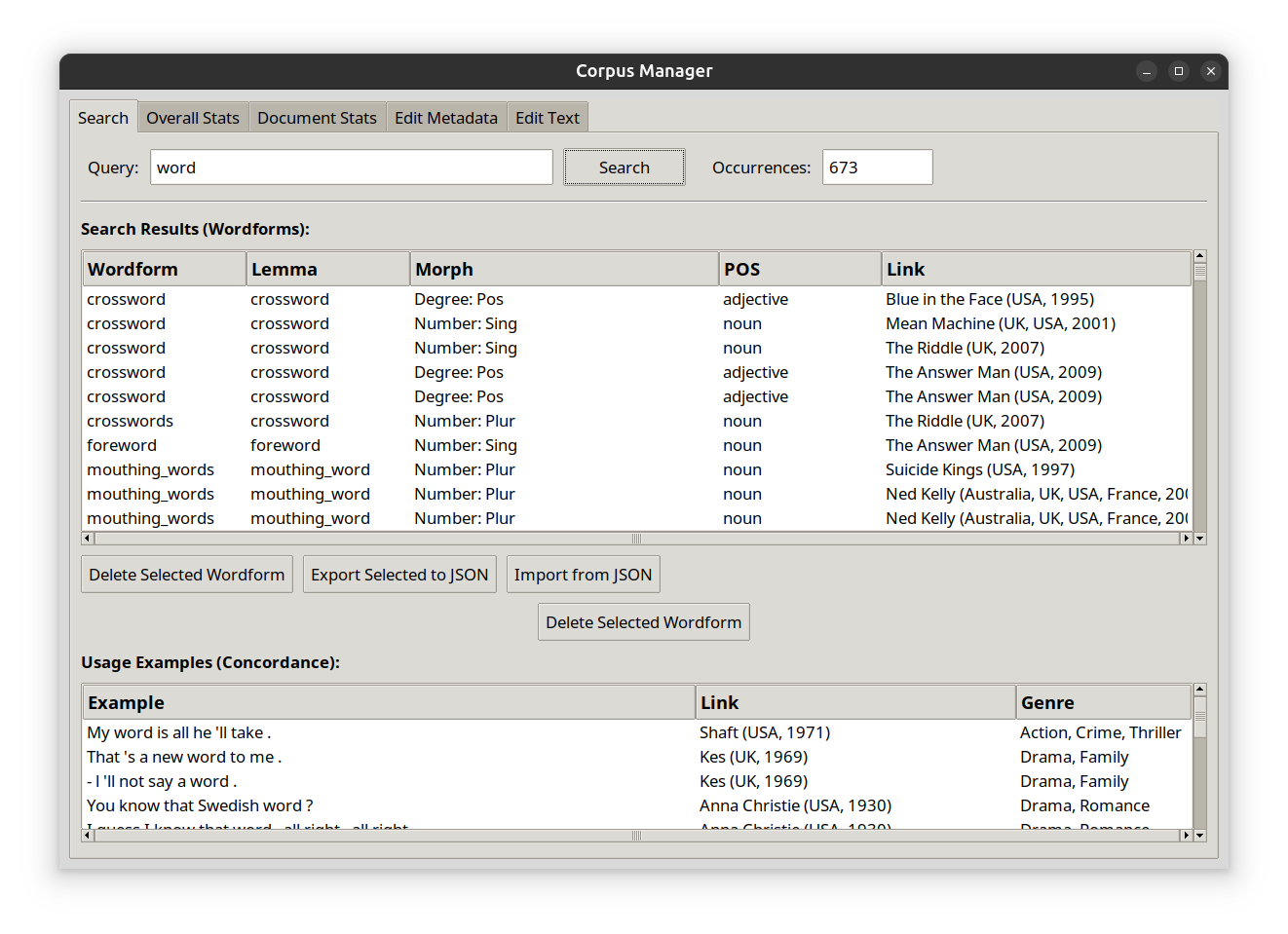
Система работает со следующими частями речи:

1. **VERB**: глагол
2. **PRON**: местоимение
3. **ADJ**: прилагательное
4. **NOUN**: существительное
5. **ADV**: наречие
6. **ADP**: предлог
7. **DET**: артикль/детерминатив
8. **PROPN**: имя собственное
9. **CCONJ**: сочинительный союз
10. **AUX**: вспомогательный глагол
11. **SCONJ**: подчинительный союз
12. **NUM**: числительное
13. **PART**: частица
14. **INTJ**: междометие
15. **SYM**: символ
16. **PUNCT**: знак препинания

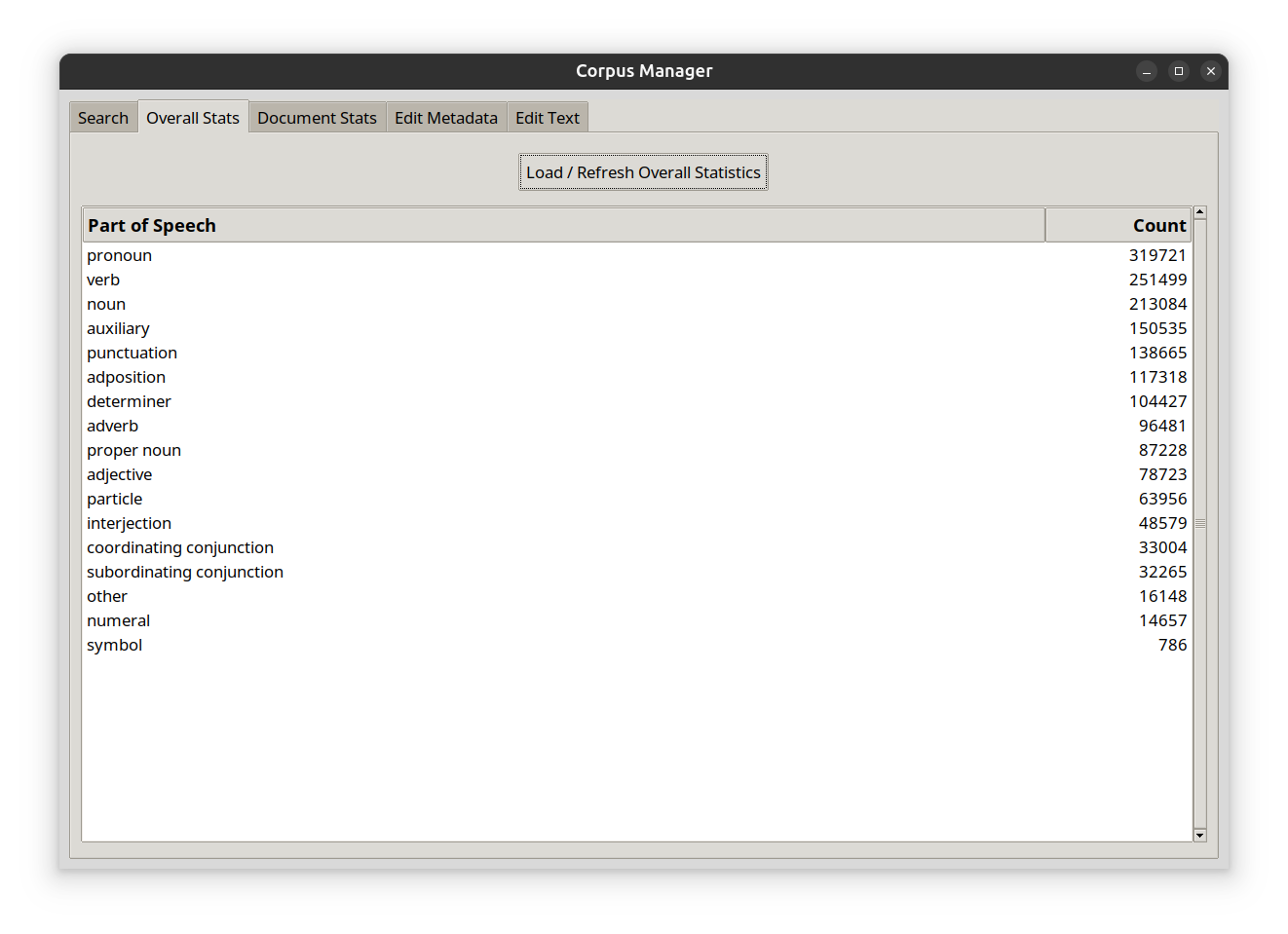
**Скриншоты разработанной системы:**



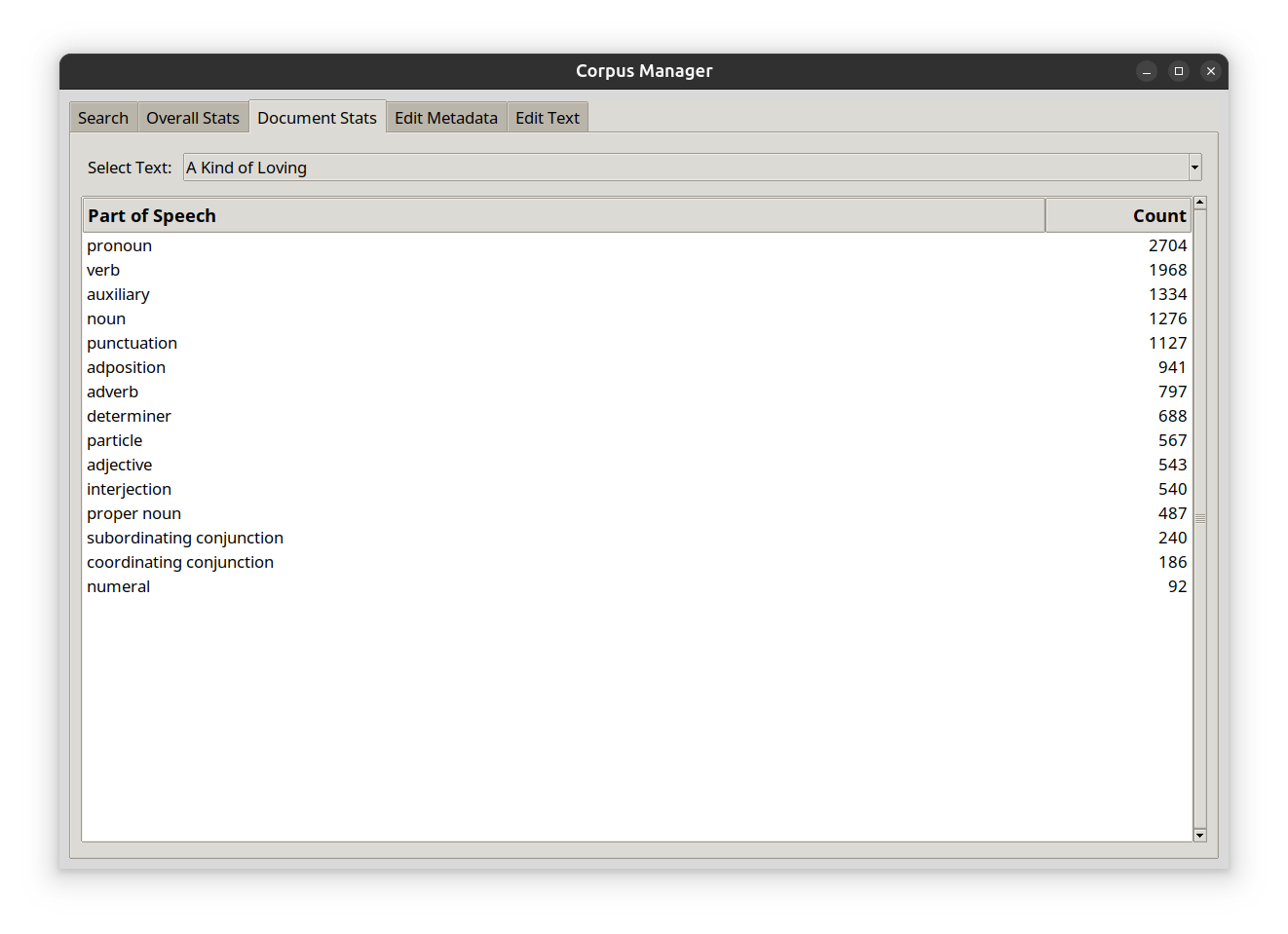
*Рис. 3 - начальная страница*

**

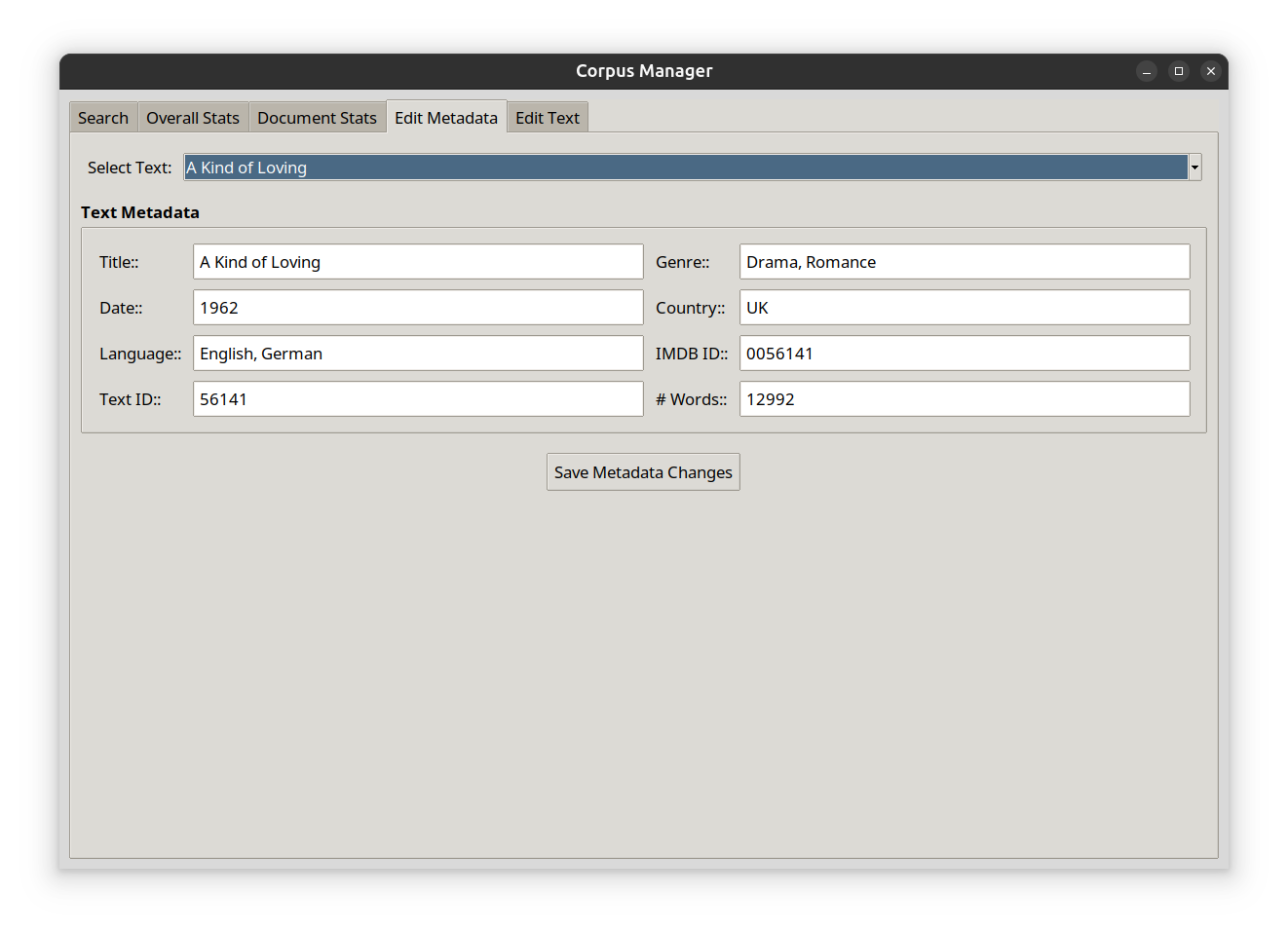
*Рис. 3 - результат поиска словоформы “word”*

**

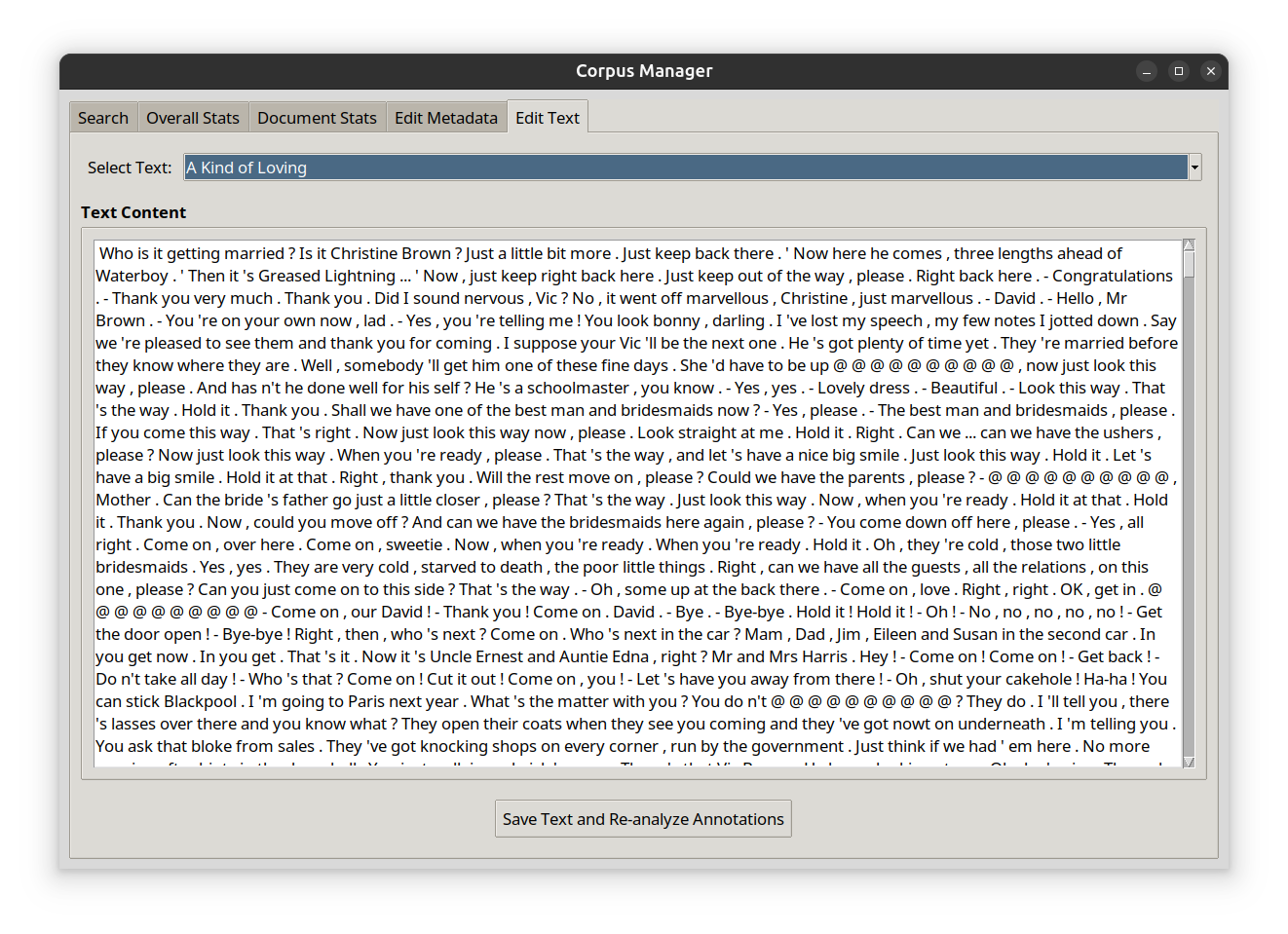
*Рис. 5 - общая статистика*

**

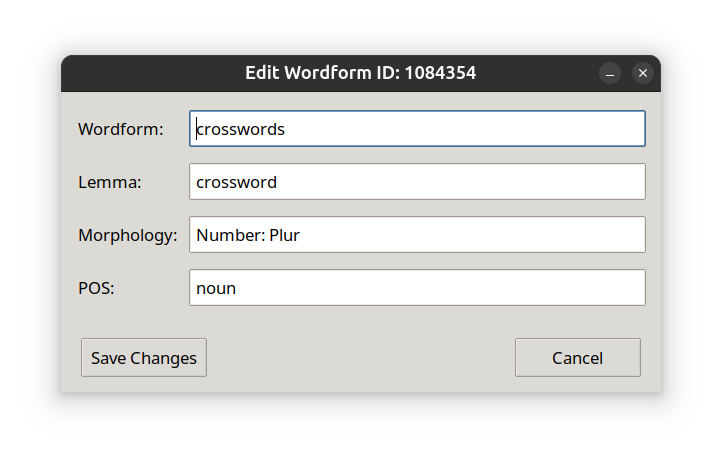
*Рис. 6 - статистика по фильму “A Kind of Loving”*

**

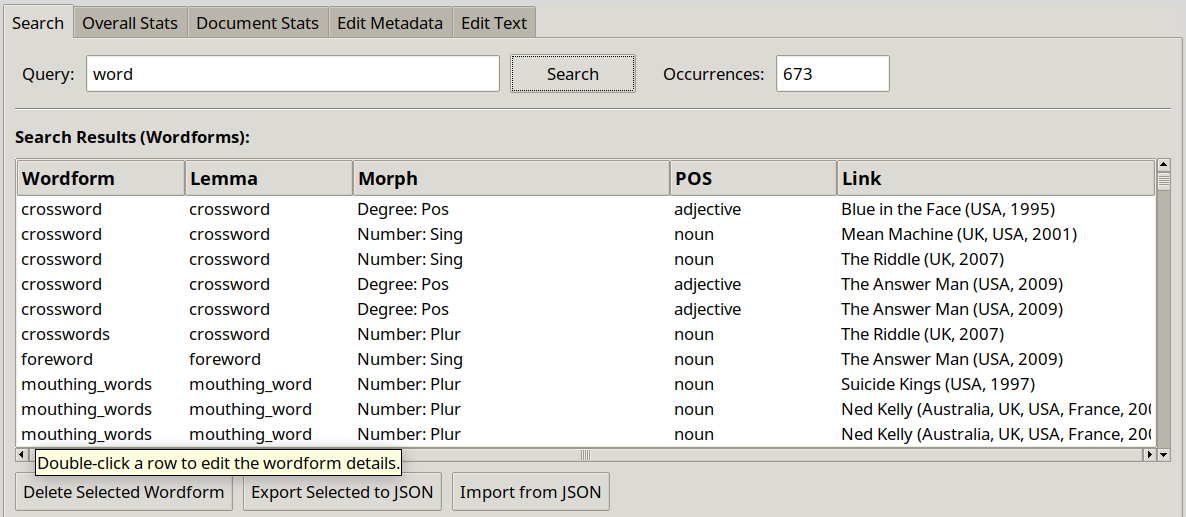
*Рис. 7 - Окно редактирования/просмотра метаданных о фильме “A Kind of Loving”*



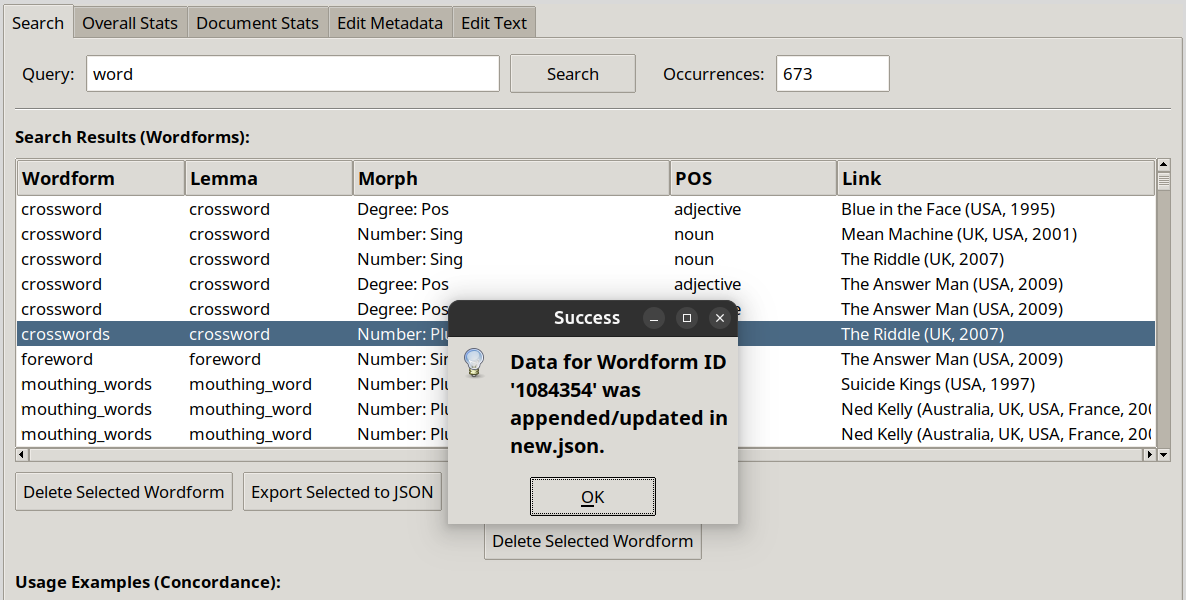
*Рис. 8 – Окно редактирования/просмотра текста фильма “A Kind of Loving”*

****

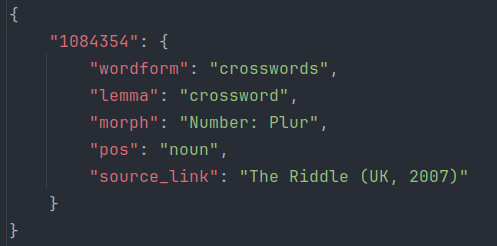
*Рис. 9 – Всплывающее окно редактирования информации о слове “crosswords”*

**

*Рис. 10 – Пример окна с помощью к таблице*

****

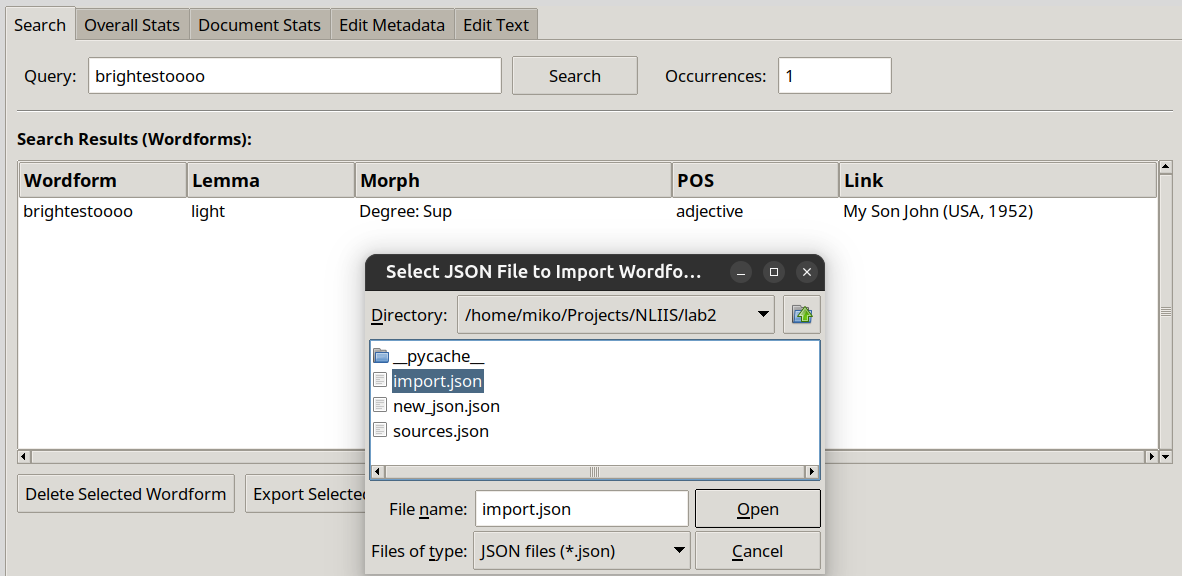
*Рис. 11 – Успешный экспорт слова в файл new.json*

**

*Рис. 12 – Содержимое файла new.json с экспортированным словом “crosswords”*

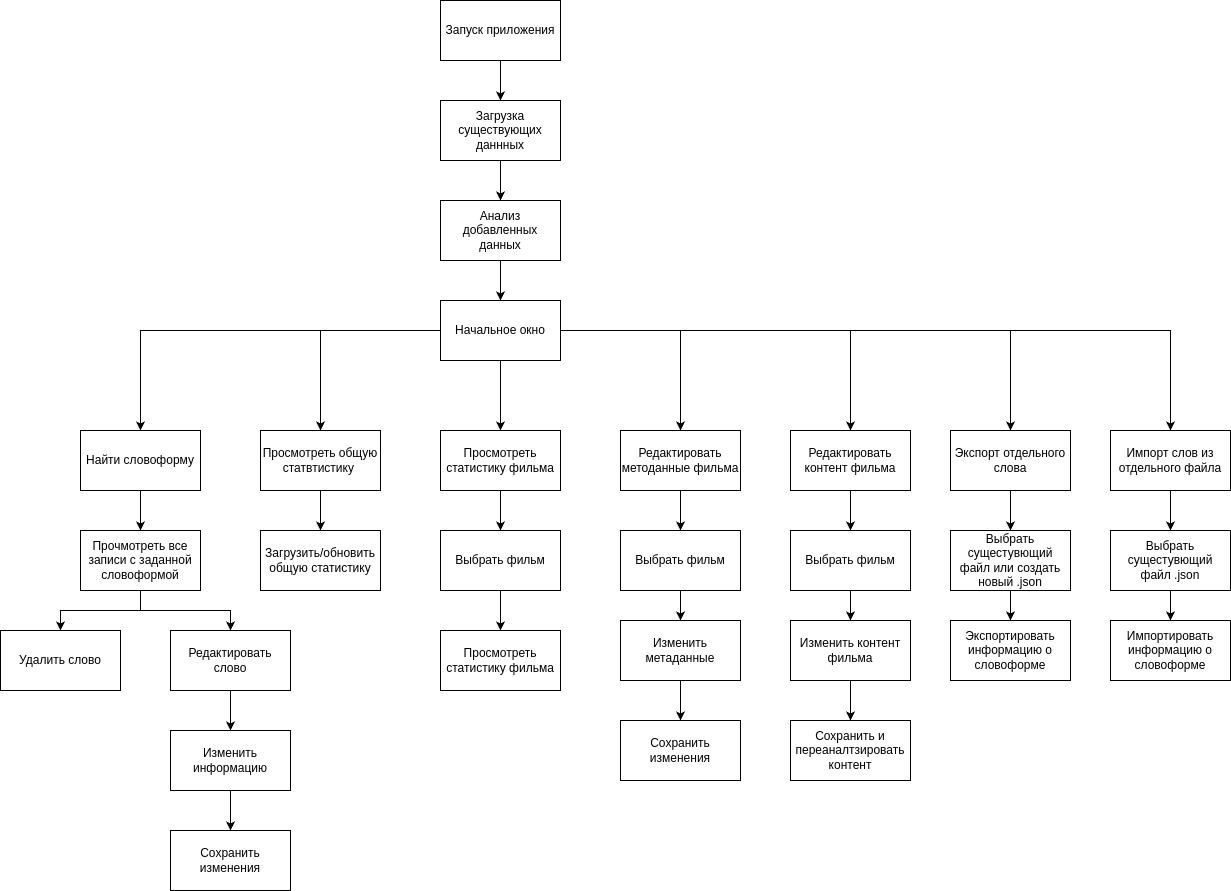
**

*Рис. 13 – Содержимое файла import.json с импортированным словом “brightestoooo”*

**

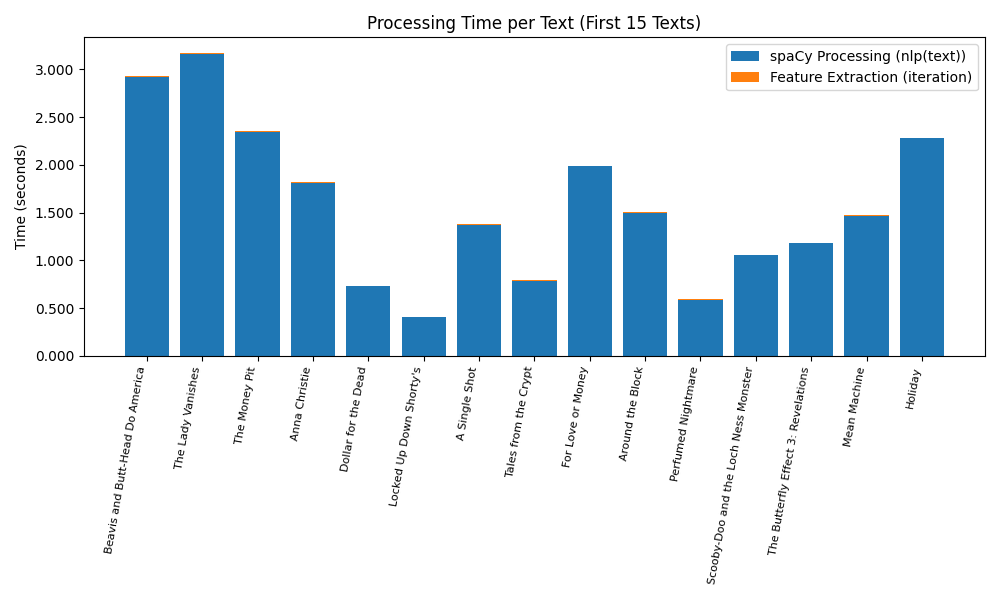
*Рис. 14 – Окно импорта и импортированное слово “brightestoooo” из файла import.json*

**Структурно-функциональная схема разрабатываемого приложения:**

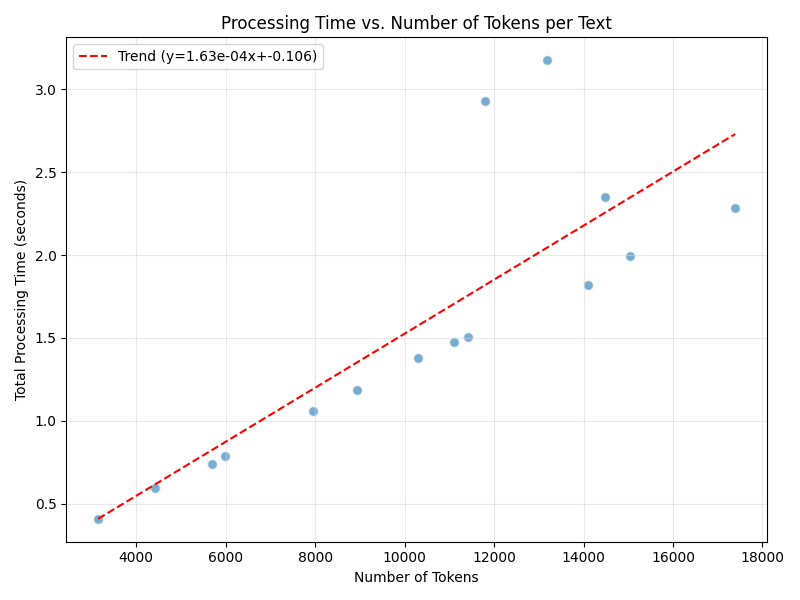


*Рис. 15 - структурно-функциональная схема приложения*

**График оценки быстродействия приложения:**



*Рис. 16 - график оценки быстродействия приложения. Зависимость времени от размера корпуса*

**

*Рис. 17 - график быстродействия обработки токенов. Зависимость времени от количества токенов*

**Выводы по работе и по перспективам развития приложения:**

Разработанное приложение представляет собой корпусный менеджер, который позволяет выполнять различные операции над корпусом текстов. Также был спроектирован корпус текстов, предметной областью которого является кинематограф. В рамках перспективы развития приложения можно выделить возможность сделать приложение многопользовательским.