**Белорусский государственный университет**

**Факультет прикладной математики и информатики**

**Кафедра информационных систем управления**

**Программное обеспечение для работы**

**с лингвистической базой знаний**

Справочное руководство

**Исполнитель**

Студент 3 курса 12 группы

Козунова Алексея Леонидовича

Минск, 2021

**Оглавление**

[1. Назначение и условия применения программы 3](#_Toc90925685)

[1.1. Цель, назначение решаемой задачи 3](#_Toc90925686)

[1.2. Требования к структуре и функционированию 3](#_Toc90925687)

[1.3. Требования к программному обеспечению 4](#_Toc90925688)

[1.4. Требования к техническому обеспечению 4](#_Toc90925689)

[2. Характеристика программы 5](#_Toc90925690)

[2.1. Общее описание программы 5](#_Toc90925691)

[2.2. Возможности программы 5](#_Toc90925692)

[2.3. Запуск приложения 6](#_Toc90925693)

[2.4. Хранение и обработка данных 6](#_Toc90925694)

[2.5. Интерфейс приложения 6](#_Toc90925695)

[2.6. Автоматическая раскраска словаря 7](#_Toc90925696)

[3. Обращение к программе 9](#_Toc90925697)

[3.1. Создание словаря на основе текста 9](#_Toc90925698)

[3.2. Работа со словарем 10](#_Toc90925699)

[3.3. Раскраска текста и снятие статистики 12](#_Toc90925700)

[4. Входные данные 18](#_Toc90925701)

[4.1. Перечень о описание структурных единиц 18](#_Toc90925702)

[4.2. Требования к входным данным 18](#_Toc90925703)

[5. Выходные данные 19](#_Toc90925704)

# 1. Назначение и условия применения программы

## 1.1. Цель, назначение решаемой задачи

Среди систем, составляющих основу перспективных информационных технологий, особое место занимают системы автоматической обработки естественного языка. Это определяется тем, что естественный язык является не только инструментом мышления, но и универсальным средством общения – средством восприятия, накопления, хранения, обработки и передачи информации.

Естественный язык не требует специального изучения, предоставляет широкие возможности для выражения любого содержания и, следовательно, является тем средством общения с компьютером, которое может удовлетворить пользователей любой специальности и квалификации. В наше время, когда пользователем может оказаться практически каждый, проблема взаимодействия человека с ЭВМ на естественном языке стала важной практической задачей.

Прогресс в развитии и использовании вычислительной техники привел к осознанию роли естественного языка как универсального средства описания действительности и коммуникации с вычислительной системой. Причем наблюдается четкая тенденция перехода как к «развитым» моделям представления знаний и разработке лингвистических процессоров извлечения классических элементов знаний непосредственно из текста, так и к языкам программирования более высокого уровня, максимально приближенных к естественному языку, для профессионального пользователя-программиста и к развитому интерфейсу для необученного пользователя.

Цель: разработать образец лингвистической базы знаний (ЛБЗ); соответствующий инструментарий (программное обеспечение) для работы с ЛБЗ.

## 1.2. Требования к структуре и функционированию

Программное обеспечение (ПО) работы с ЛБЗ должно поддерживать выполнение следующих действий:

* создание базового эталонного словаря и корпуса текстов с возможностью их просмотра;
* получение аннотированного корпуса текстов - "раскрасить" все подобранные тексты с возможностью постредактирования;
* аннотирование базового словаря;
* снятие всей необходимой статистики с корпуса текстов, продумав их хранение в ЛБЗ:
* абсолютная частота встречаемости слова (без кода) – выполенено на этапе создания словаря;
* абсолютная частота встречаемости каждого кода;
* абсолютная частота встречаемости пары слово\_код;
* абсолютная частота встречаемости пар кодов (рядом стоящих);
* пополнение словаря и корпуса новыми текстами и словами – весь процесс со снятием статистики. ПО должно обладать дружественным интерфейсом пользователя;
* разработка простейшую ИПС на основе созданной ЛБЗ.

## 1.3. Требования к программному обеспечению

Приложение представляет собой многофункциональную программу с дружественным графическим интерфейсом пользователя. Для работы программы компьютер должен иметь установленную реализацию Java Runtime Environment (JRE), необходимую для исполнения Java-приложений.

## 1.4. Требования к техническому обеспечению

Основные требования для корректной работы приложения:

* Операционная система Windows или Linux;
* 2 Гб оперативной памяти;
* Устройство вывода графической информации (монитор);
* Устройство ввода (клавиатура);
* 500 Мб свободного места на диске.

# 2. Характеристика программы

## 2.1. Общее описание программы

Данное приложение представляет собой настольное оконное приложение с графическим интерфейсом, написанное на языке Java с использованием технологии JavaFX, предназначенное для удобной обработки текстов естественного языка.

## 2.2. Возможности программы

Программа предоставляет пользователю следующие возможности:

* создание базового эталонного словаря и корпуса текстов с возможностью их просмотра;
* получение аннотированного корпуса текстов - "раскрасить" все подобранные тексты с возможностью постредактирования;
* аннотирование базового словаря;
* снятие всей необходимой статистики с корпуса текстов, продумав их хранение в ЛБЗ:
* абсолютная частота встречаемости слова (без кода) – выполенено на этапе создания словаря;
* абсолютная частота встречаемости каждого кода;
* абсолютная частота встречаемости пары слово\_код;
* абсолютная частота встречаемости пар кодов (рядом стоящих);
* пополнение словаря и корпуса новыми текстами и словами – весь процесс со снятием статистики. ПО должно обладать дружественным интерфейсом пользователя;
* разработка простейшую ИПС на основе созданной ЛБЗ.

Таким образом, требования, поставленные к набору функций, предоставляемых приложением, выполнены.

## 2.3. Запуск приложения

Программа собрана в jar-архив, поэтому для её запуска достаточно дважды кликнуть по архиву. Пример запуска изображён на рисунке 2.1.



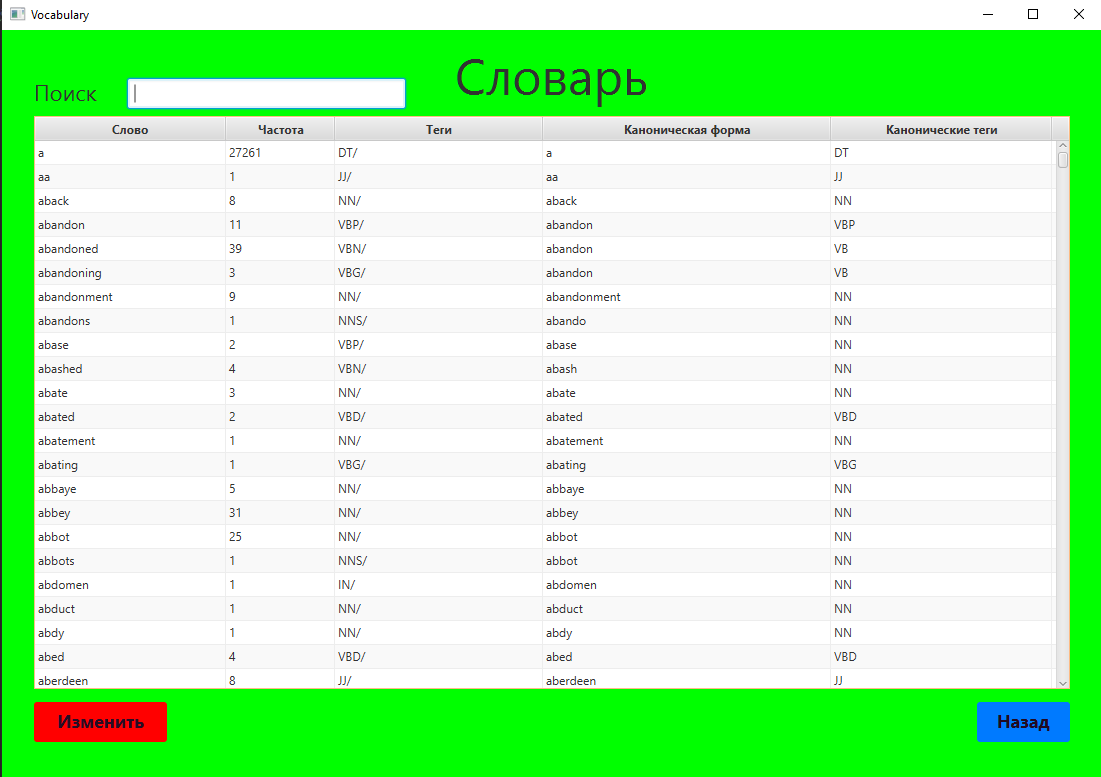
**Рисунок 2.1 – Запуск приложения**

## 2.4. Хранение и обработка данных

Приложение по обработке естественного языка должно хранить всю необходимую информацию о тексте, поэтому нужно каким-то образом организовать хранение всех данных. Весь процесс хранения и обработки данных осуществляется с помощью графического интерфейса. При необходимости пользователь может сохранить обработанные словари как по-отдельности, так и все вместе. От запуска к запуску приложения данные требуется новая загрузка словаря или текста. Помимо этого, пользователь также имеет возможность сохранения раскрашенных текстов, чтобы при необходимости вернуться к раскраске и внести исправления, а также для возможности снимать с них статистику.

## 2.5. Интерфейс приложения

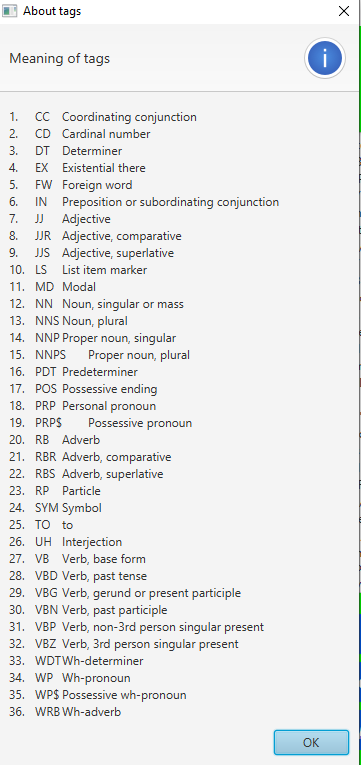
Пользователь работает со словарём посредством оконного приложения, которое визуализирует в виде таблиц всю необходимую информацию. Таблицы чётко структурированы и предоставляют всю важную с точки зрения лингвиста информацию. Внешний вид приложения с загруженным текстом представлен на рисунке 2.2.

   
**Рисунок 2.2 – Приложение с загруженным словарем**

## 2.6. Автоматическая раскраска словаря

Одна из базовых функций электронного словаря – создание словаря непосредственно из входящего текста. Ручное составление словаря очень затратно, сложно и неэффективно. Автоматическая раскраска (определение части речи каждого слова) словаря в программном комплексе «Электронный словарь» осуществляется благодаря использованию библиотеки Stanford Log-linear Part-Of-Speech Tagger.

Список частей речи, которые определяет библиотека *stanford-postagger*, можно увидеть на рисунке 2.3.

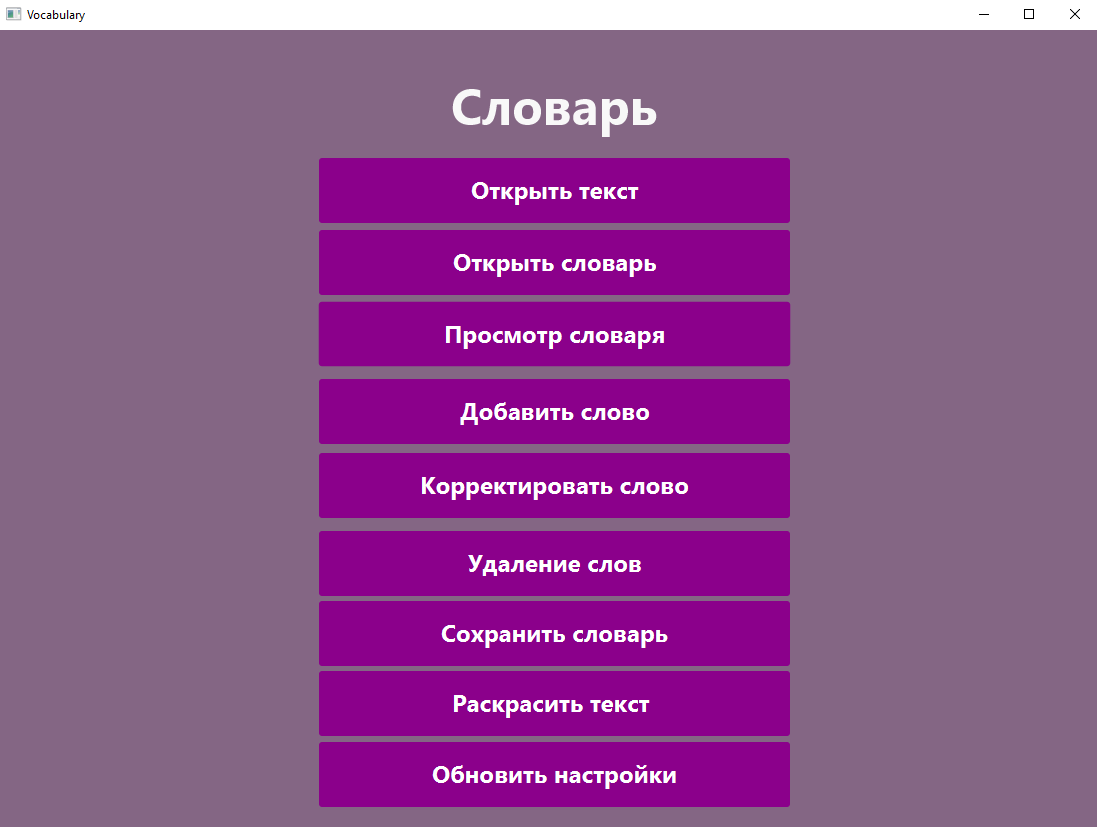


**Рисунок 2.3 – Справка о частях речи и соответствующих им кодах**

# 3. Обращение к программе

## 3.1. Создание словаря на основе текста

Для начала работы пользователю необходимо запустить приложение. Внешний вид приложения после запуска представлен на рисунке 3.1.

  
**Рисунок 3.1 – Главный экран приложения**

Для выбора исходного текста нужно выбрать пункт меню Open file, после чего откроется окно с выбором файла для создания словаря. Далее требуется нажать на кнопку Open file и в диалоговом окне выбрать необходимый текстовый файл. После выбора файла в окне отобразиться его полное название. После чего нажать кнопку ОК. Окно с выбранным файлом представлено на рисунке 3.2.

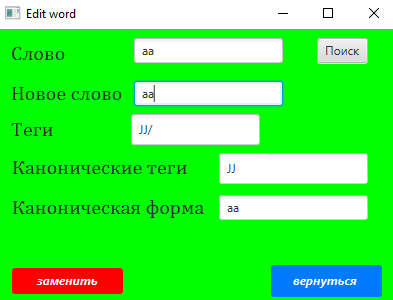
  
**Рисунок 3.2 – Выбор файлов для создания словаря**

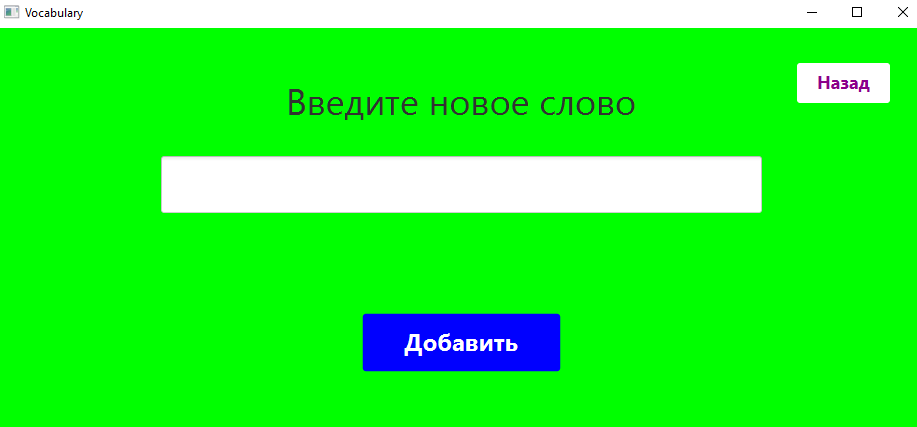
После загрузки файла словарь создается автоматически. Чтобы увидеть результат необходимо из главного меню перейти по кнопке Show vocabulary. В новом окне можно увидеть таблицу со словарем, что на рисунке 2.3.

## 3.2. Работа со словарем

Приложение предоставляет пользователю большое разнообразие возможных операций с созданным словарем. Для удобства отображения слова в словаре можно отсортировать как по возрастанию, так и по убыванию: по частоте, по алфавиту, по тегам, по канонической форме.

Также приложение предоставляет возможность редактирования, добавления и удаления слов. Для этого требуется в меню приложения выбрать соответствующие пункты: Correct word, Add word, Delete word. На рисунке 3.3 показана окно редактирования слова.

  
**Рисунок 3.3 – Редактирование слова в словаре**



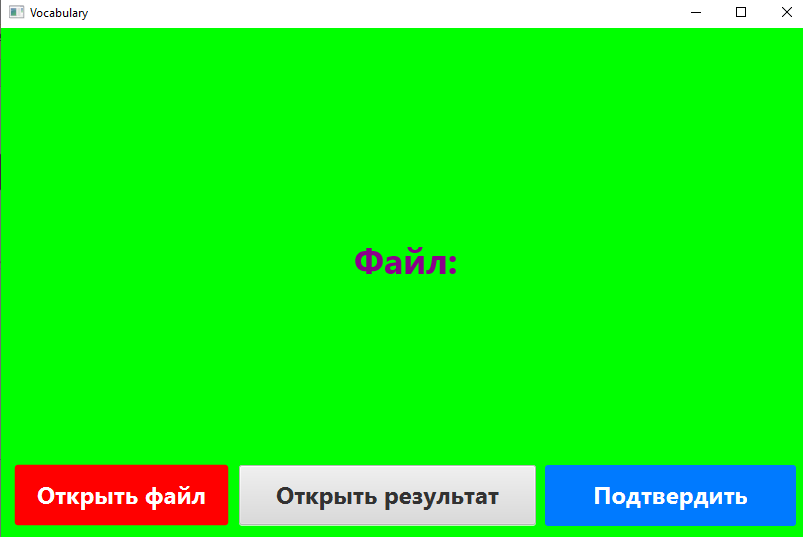
**Рисунок 3.4 – Добавление нового слова в словарь**

****

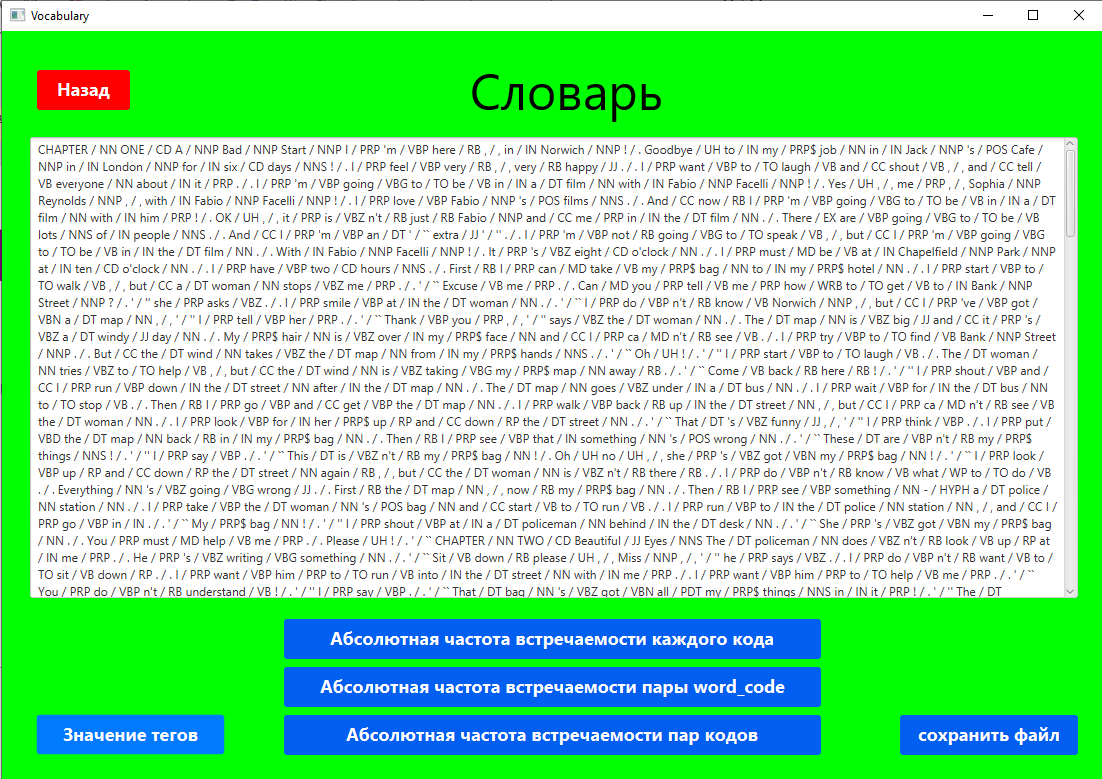
**Рисунок 3.5 – Удаление слова из словаря**

## 3.3. Раскраска текста и снятие статистики

С помощью созданного полного словаря можно раскрашивать тексты и снимать по ним статистику. Раскраска текста заключается в определении кодов слов исходного текста на основе созданного словаря. Для того, чтобы раскрасить текст, в меню нужно выбрать Color the text. В новом окне можно выбрать текст, который будем раскрашивать, или текст, который раскрасили ранее, и править его.

****

**Рисунок 3.6 – Выбор текста**

****

**Рисунок 3.7 – Раскрашенный текст**

В данном представлении пользователь может редактировать слова, а также каждому слову ставить в соответствие код (если автоматическая раскраска содержит неточности или ошибки). После завершения редактирования или в процессе работы имеется возможность сохранять аннотированный текст, выбрав Save taging. В дальнейшем, к этой раскраске можно вернуться, выбрав пункт меню Color the text – Open colorized file.

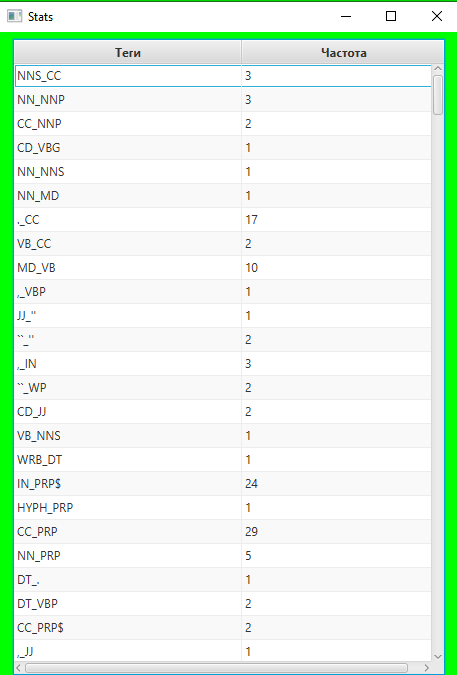
С раскрашенного текста можем снять следующую статистику: абсолютная частота встречаемости каждого кода, абсолютная частота встречаемости пары слово\_код, абсолютная частота встречаемости пар кодов (рядом стоящих).

****

**Рисунок 3.8 – Абсолютная частота встречаемости каждого кода**

****

**Рисунок 3.9 – Абсолютная частота встречаемости пары слово\_код**

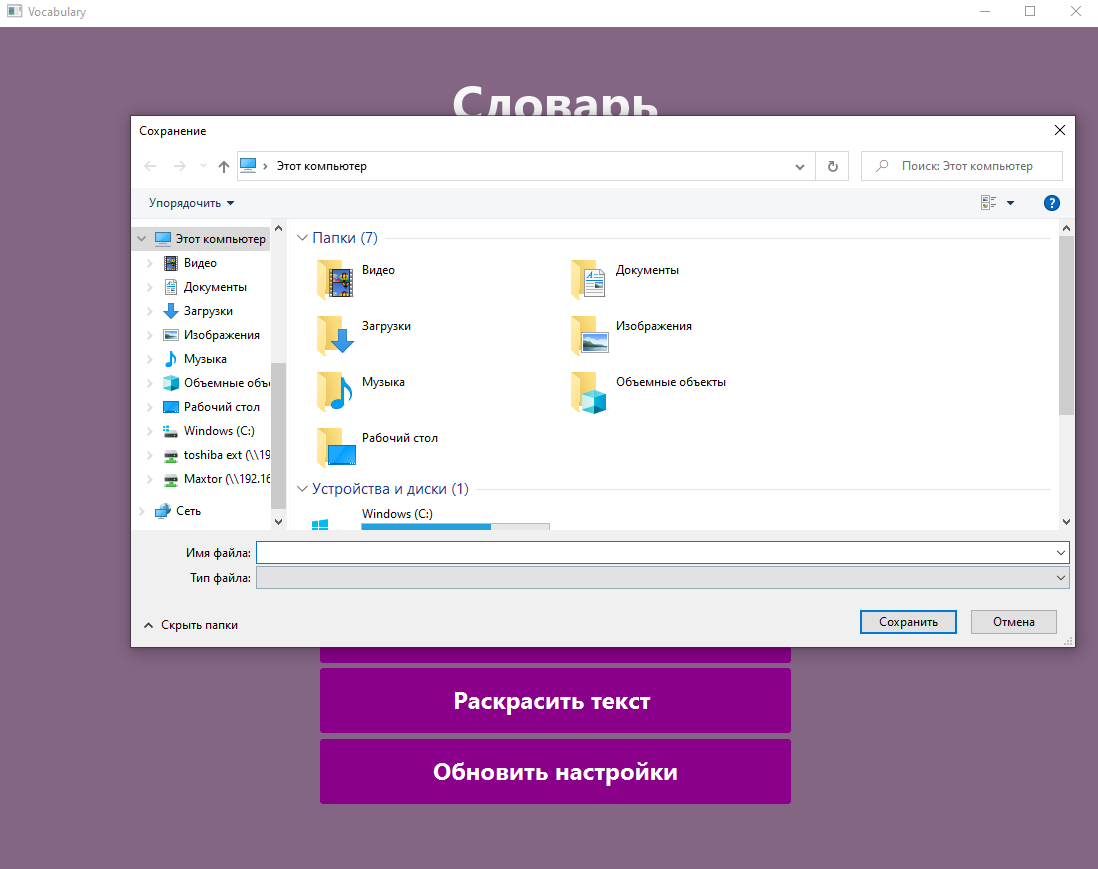
****

**Рисунок 3.10 – Абсолютная частота встречаемости пар кодов (рядом стоящих)**

Также в главном меню можно открыть существующий словарь для дальнейшей работы с ним, сохранить текущий словарь и очистить текущий словарь и корпус текстов. Для этого требуется выбрать соответствующие кнопки: Open vocabulary, Save vocabulary, Reset.

**Рисунок 3.11 – Открытие словаря**

При нажатии на Save vocabulary появится диалоговое окно для выбора директории и имени файла.

**Рисунок 3.12 – Сохранение словаря**

# 4. Входные данные

## 4.1. Перечень о описание структурных единиц

Входными данными для программы являются текстовые файлы. Тексты имеют свободную форму и не требуют никакой предварительной обработки.

## 4.2. Требования к входным данным

Ключевым требованием является расширение входных файлов, которое должно иметь формат txt.

# 5. Выходные данные

Выходными данными программного комплекса являются словари и раскрашенные тексты. Раскрашенные словари хранятся в формате .txt, который включает слово, его частоту и тег. Раскрашенные тексты хранятся в формате .txt, в связи с чем их можно просмотреть с помощью любого текстового редактора, однако изменение слова или его тега гораздо удобнее осуществлять, открывая раскрашенные тексты внутри разработанного приложения.