УДК 004.4

Сибгатуллин Марсель Рауфович

Sibgatullin Marsel Raufovich

Аспирант

Казанский национальный исследовательский технический университет им. А.Н.Туполева

Kazan national research technical university named after A.N.Tupolev

РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ГЕНЕРИРОВАНИЯ СЛОВОФОРМ В КОРПУСНОЙ ЛИНГВИСТИКЕ

DEVELOPMENT OF AN INFORMATION SYSTEM FOR GENERATING WORD FORMS IN CORPUS LINGUISTICS

Аннотация: Данная статья посвящена разработке информационной системы для генерирования словоформ в корпусной лингвистике. Целью работы является создание легко масштабируемой системы, способной генерировать словоформы и выдавать по ним статистические данные.

Abstract: This thesis is devoted to the development of an information system for generating word forms in corpus linguistics. The aim of the work is to create an easily scalable system capable of generating word forms and producing statistical data on them.

Ключевые слова: база данных (БД), информационная система (ИС), веб-приложение, архитектура, клиент, сервер.

Keywords: database (DB), information system (is), web application, architecture, client, server.

Введение. Задача написания качественных учебных материалов для школьников и студентов является актуальной всегда. В наше время представлено очень большое количество контента, для выполнения тех или иных задач. К примеру, изучения иностранных языков, обучения программированию, методические указания для выполнения разного рода задач. Но они не всегда являются понятными для всех кругов общества и разного склада ума. На сегодняшний день появилось большое число программных инструментов,

которые позволяют качественно решать такую задачу. В представленной работе описывается разработка информационной системы для анализа словоформ тюркских языков. Конечная цель использования такой системы — выделение наиболее часто встречающихся слов выбранного языка для подготовки качественного контента для учебных материалов, книг для изучающих тюркские языки.

Целью работы являлась разработка информационной системы для генерирования словоформ в корпусной лингвистике. В рамках этого проекта будут решатся следующие задачи:

1. Проанализировать функциональность аналогов.
2. Разработать архитектуру создаваемой информационной системы.
3. Разработать диаграммы активности пользователя и спроектировать базу данных для информационной системы.
4. Разработать алгоритм работы веб-приложения.
5. Реализовать проект.

Информационная система для генерирования словоформ в корпусной лингвистике является ресурсом для образования форм слов по заданным критериям и получения статистики по частоте вхождения указанных слов и их производных. Данная функция даёт нам возможность определить те самые слова, которые являются более понятными для осознания и понимания текста.

Для реализации нашего проекта решено разработать информационную систему (ИС), представляющую собой веб-приложение. В качестве клиента для доступа к веб-порталу выступает интернет-браузер пользователя. Разработка осуществляется с использованием языков программирования PHP и JavaScript.

Пользователями ИС могут являться люди разного рода деятельности. Начиная от студентов и преподавателей лингвистических и литературных направлений, вплоть до авторов книг, статей и писателей. Основные функции доступные каждому из пользователей системы можно продемонстрировать в виде UML диаграммы деятельности (рис. 1) [1]:

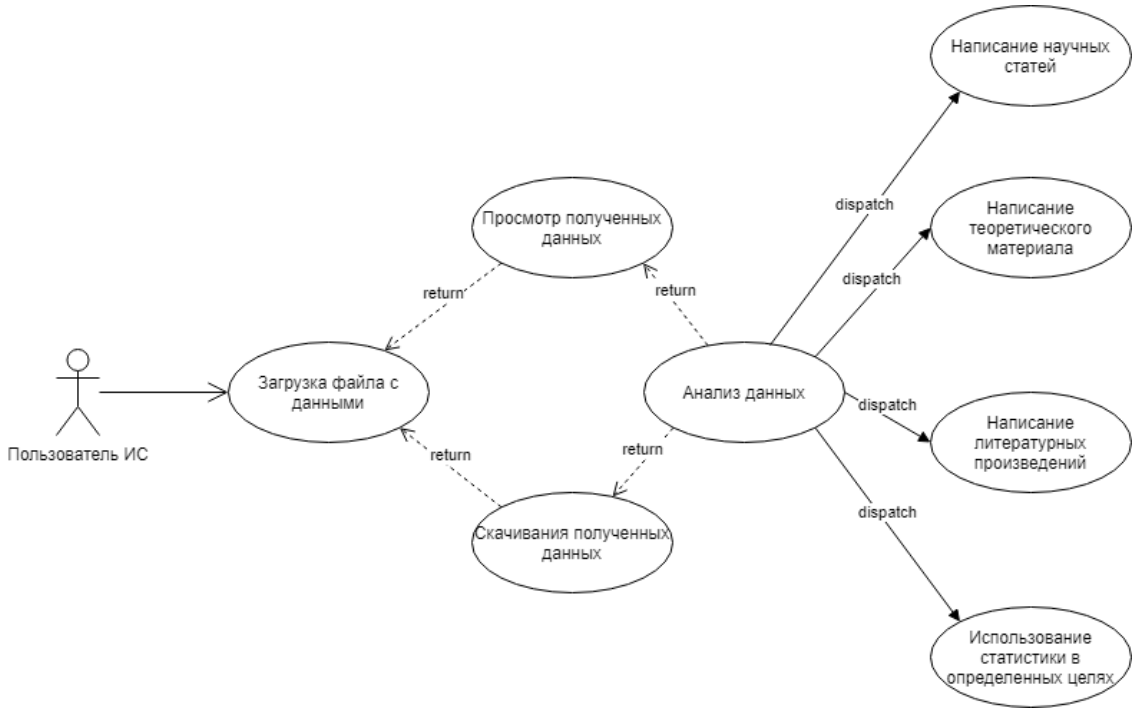

Рис. 1. Диаграмма деятельности пользователя

Основной функционал ИС, который необходимо разработать:

- удобный и лёгкий для понимания интерфейс;
- форма загрузки рабочего файла;
- выдача результата на экран;
- выдача результата для скачивания файлом;
- возможность работать с файлами с кодировкой турецкого языка;

Постановка задачи

Главная цель проекта – создание инструмента, который поможет лингвистам находить в турецких текстах предложения и более крупные фрагменты, удовлетворяющие определенным поисковым критериям. На поздних этапах разработки приложения, оно будет позволяет использовать в качестве критериев поиска следующие типы информации:

- словоформы и лексемы
- лексические и грамматические категории, словоизменительные типы
- пунктуация и регистр

Планируется, что наше приложение в будущем будет позволяет также осуществлять контекстные запросы для поиска сочетаний нескольких слов.

Архитектура и описание разрабатываемого веб-приложения. Было принято решение о разработке информационной системы без использования какого-либо фреймворка для языка php [2]. Архитектуру данного веб-приложения можно увидеть ниже (рис. 2).

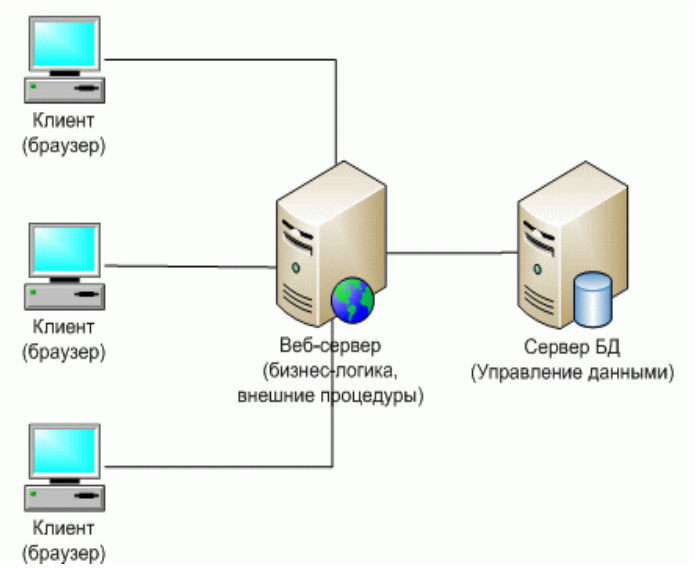

Рис. 2. Архитектура разрабатываемого веб-приложения

Пользователи получают доступ к веб-приложению через сеть Интернет. Далее получают доступ к функционалу сайта без регистрации или авторизации. В качестве клиента доступа выступает интернет-браузер. Для получения полного функционала веб-приложения нет необходимости проходить регистрацию или авторизацию. Данным ресурсом могут пользоваться все желающие без дополнительных действий.

Перед началом разработки мы решили представить логику работы приложения в виде блок-схемы, где i – индекс слова, m – часть речи, k – последняя буква слова, l – предпоследняя буква слова (рис. 3).

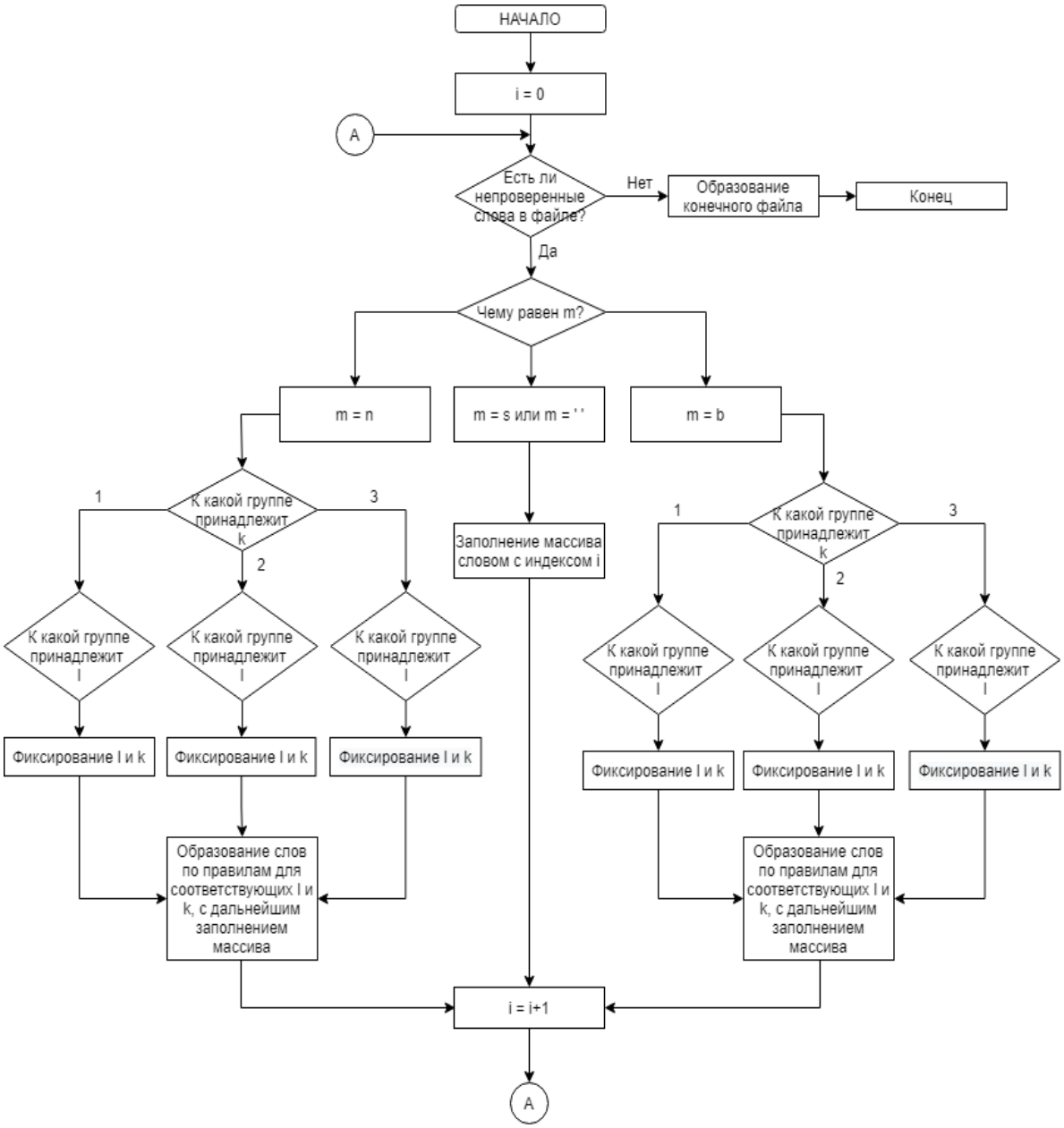

Рис. 3. Блок-схема логики работы приложения

Диаграммы активности пользователя. Были выделены две основные группы (роли) пользователей нашей системы: авторы статей и литературных произведений и люди, заинтересованные в статистике. У каждой из них существуют свои функции. Взаимодействие происходит через общую БД при помощи СУБД MySQL. Алгоритмы работы представлены ниже в виде UML диаграмм активности при выполнении обеих групп пользователей своих функций (рис. 4 – 5) [1].

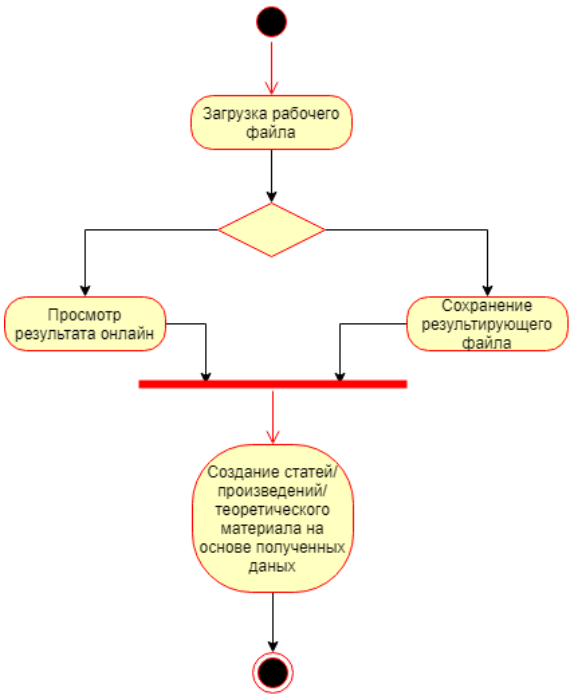

Рис. 4. Диаграмма активности авторов

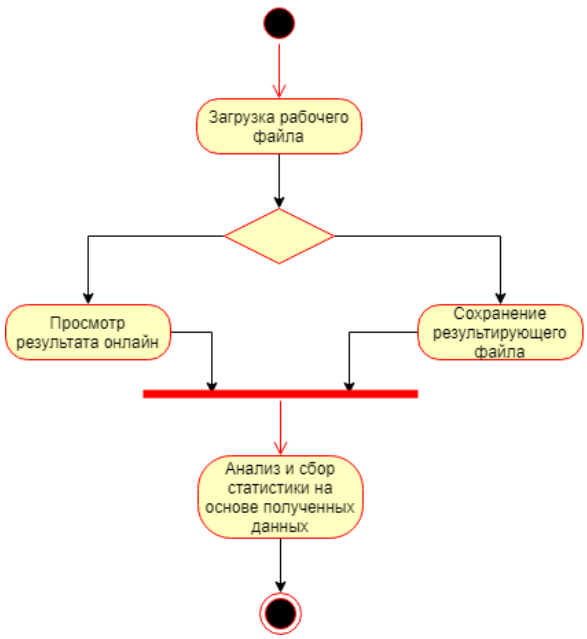

Рис. 5. Диаграмма активности заинтересованных в статистике

Проектирование базы данных. Для работы, разрабатываемой ИС (веб-приложения), была спроектирована простая база данных, содержащая в себе 5 таблиц. Взаимодействие нашего веб-приложения происходит через общую БД при помощи СУБД MySQL через передачу SQL запросов [4]. Структура таблиц для разрабатываемой ИС показаны ниже (рис. 6).

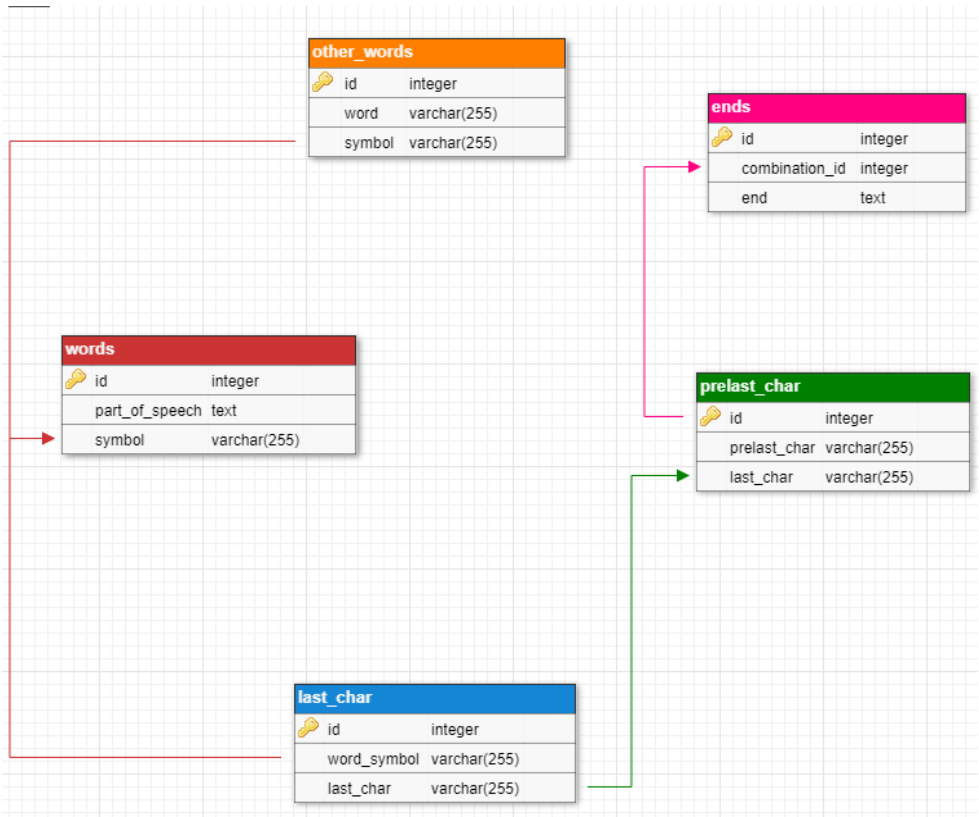

Рис. 6. Структура таблиц базы данных

Выводы

Результатом проделанной работы можно считать выполнение поставленных задач. В конечном итоге мы получили веб-приложение в сети интернет, которое позволяет нам генерировать словоформы в корпусной лингвистике, а также предоставляет данные о статистике и количестве производных слов. Функционалом является образование словоформ по изначально заданным частицам и словам. В состав выполненных задач входит:

– Проанализирована функциональность аналогов разработанной системы.

– Разработана диаграмма активности и спроектирована база данных для информационной системы.

– Разработан алгоритм работы системы.
– Разработана блок-схема для логики работы приложения.
– Реализовано веб-приложение.

Библиографический список:

1. Мартин Фаулер. UML. Основы, 3 е издание. – Пер. с англ. – СПб: Символ Плюс, 2004. – 192 с.
2. Сибгатуллин М.Р. «Разработка информационной системы поддержки для управления выпускными работами студента» // Сборник конференции «Передовые инновационные разработки. Перспективы и опыт использования, проблемы внедрения в производство» (В печати).
3. Простые MVC-приложения. [Электронный ресурс] // Разработка веб-сайтов, PHP. URL: https://habr.com/ru/post/320480/(дата обращения: 15.11.2020).
4. Основные команды SQL, которые должен знать каждый программист. [Электронный ресурс] // Переводы для программистов. URL: https://tproger.ru/translations/sql-recap/ (дата обращения: 20.12.2020).
5. Мэтт Зандстра PHP. Объекты, шаблоны и методики программирования. 4-е издание. // Пер. с англ. - М.: ООО И.Д. Вильямс, 2015. – 576 с.
6. Игорь Симдянов, Дмитрий Котеров. PHP 7. В подлиннике // БХВ-Петербург, 2016. – 1069 с.