Инженерный вестник Дона, №6 (2022)
ivdon.ru/ru/magazine/archive/n6y2022/7694

Разработка сервиса для генерации словоформ в корпусной лингвистике

М.Р. Сибгатуллин[1], Р.Ш. Минязев[1], И.И. Сафиулин[2], А.Ш. Бикташева[2], Н.П. Пашин[1]

[1]Казанский национальный исследовательский технический университет имени А. Н. Туполева – КАИ, Казань
[2]Казанский (Приволжский) федеральный университет, Казань

Аннотация: Предмет исследований – разработка сервиса для генерации различных форм заданного слова, исходя из анализа слов, найденных в словаре. Были изучены имеющиеся подходы к решению такой задачи и выбран наиболее релевантный. Сервис осуществляет поиск внутри файла словаря с текстовым содержимым с целью автоматизации процесса выделения нужных слов среди всего множества. Выполняется поиск основы слова, учитывающий морфологию. С выполнением морфологического разбора слова, находится общая для всех его грамматических форм основа, отсекаются суффиксы и окончания. В результате алгоритм работы сервиса позволяет искать все формы слова по заданному ключевому слову, учитывая словоформы. При этом также анализируется, к какой части речи относится слово, это позволяет задавать разные методики определения словоформ. Для каждого типа слова: глагол, существительное, прилагательное, наречие, используется свой алгоритм для выделения словоформ. Особенность сервиса в том, что он позволяет не только искать словоформы по словарю, но и позволяет генерировать наборы словоформ, исходя из типа заданного слова. Сервис функционирует на платформе Linux под управлением веб-сервера Apache. Для разработки использованы бесплатные программные инструменты. Разработка велась на языках JavaScript, HTML и CSS, также использовался серверный язык программирования PHP7.
Ключевые слова: поисковая система, анализ документов, лингвистика, словоформы, морфология, генерация слов, веб-сервис.

Введение

Задача написания качественных учебных материалов для школьников и студентов является актуальной всегда. На сегодняшний день представлено большое количество текстового контента, используемого для решения тех или иных задач, к примеру, изучение иностранных языков, обучение программированию, методические указания для выполнения разного рода задач. Этот контент не всегда является понятными для всех кругов общества и разных складов ума. В последние годы появилось большое число программных инструментов, которые позволяют качественно решить такую задачу [1]. В представленной работе описывается разработка информационной системы для анализа словоформ турецкого языка. Конечная

цель использования такой системы – выделение наиболее часто встречающихся слов выбранного языка для подготовки качественного контента для учебных материалов, в частности, книг для изучающих турецкий язык [2].

Целью работы являлась разработка информационной системы для генерирования словоформ в корпусной лингвистике [3]. В рамках этого проекта решались 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 [4]. Данная функция позволяет нам определить те самые слова, которые являются более понятными для осознания и понимания текста. То есть, с помощью данного ресурса мы можем выявлять слова, чаще всего встречающиеся в том или ином тексте, для дальнейшего пользования ими [5]. Так как, чем чаще встречается слово, тем проще текст для понимания, образованный с помощью этих слов [6-8].

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, заканчивая авторами книг, статей и писателями. Основные функции, доступные каждому из пользователей системы, можно продемонстрировать в виде UML- диаграммы деятельности (рис. 1):

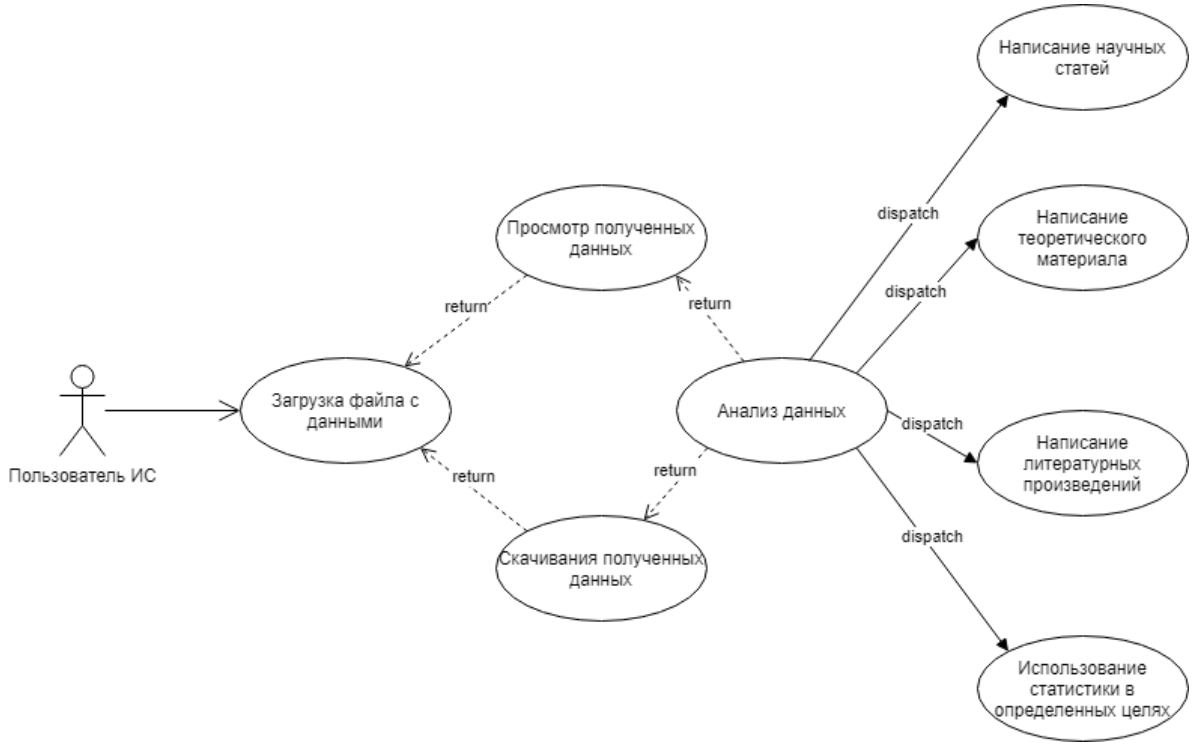

Рис. 1. Диаграмма деятельности пользователя

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.

В качестве операционной системы (ОС) для сервера используется Windows 10, на этой системе устанавливается Open Server, который является локальной серверной платформой и программной средой. Данное приложение создано специально для веб-разработчиков. Для хранения

данных используется реляционная СУБД MySQL, на нее приходится основная нагрузка по обработке запросов пользователей к ИС.

Для отладки веб-приложения в процессе разработки используется браузер Google Chrome.

Постановка задачи

Главная цель проекта – создание инструмента, который поможет лингвистам находить в турецких текстах предложения и более крупные фрагменты, удовлетворяющие определенные поисковые критерии [9]. На поздних этапах разработки приложения, оно будет позволять использовать в качестве критериев поиска следующие типы информации:

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

Планируется, что приложение в будущем будет позволять также осуществлять контекстные запросы для поиска сочетаний нескольких слов [10].

Архитектура и описание разрабатываемого веб-приложения

Было принято решение о разработке информационной системы без использования какого-либо фреймворков для языка php. Архитектуру данного веб-приложения можно увидеть ниже (рис. 2).

После полной загрузки сайта вам остаётся лишь загрузить ваш файл для работы в табличном формате XLS либо XLSX. Данный формат являются частью программы Microsoft Excel, которая входит в состав пакета приложений Microsoft Office 2007. Этот формат представляет собой мощный инструмент, позволяющий создавать и форматировать электронные таблицы, графики, а также выполнять математические и др. операции. Пользователь может создавать различные электронные таблицы с несколькими листами,

формулами, а также источниками данных. Полученные файлы можно сохранять в формате XLSX.

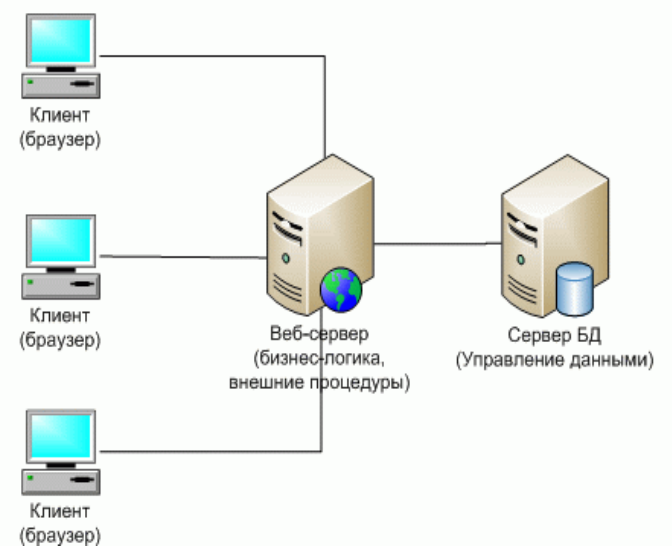

Рис. 2. Архитектура разрабатываемого веб-приложения

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).

Для реализации клиентской части мы выбрали CSS и HTML. Язык HTML позволяет нам создавать определенные текстовые и графические объекты, размещая их на экране. Язык CSS предоставляет нам возможность позиционирования и изменения внешнего облик веб-страниц большим количеством разных способов.

Далее был выбран веб-сервер Apache. Данный сервер является кроссплатформенным, он был выбран ввиду того, что является гибким и очень прост в настройке. К тому же он отличается тем, что на нем будет работать любое веб-приложение или сайт без дополнительных настроек и доработок.

Для реализации серверной части мы остановились на PHP. Нами была написано простое приложение с понятным интерфейсом, при этом мы не нагружали наш ресурс дополнительными фреймворками для языка PHP.

Основная часть приложения содержится в файле index.php. В нём происходит работа с базой данных и формирование словоформ, а также подсчёт количества вхождений производных слов.

Для образования результирующего файла и дальнейшей возможности его скачать был создан специальный скрипт DownloadFile.php.

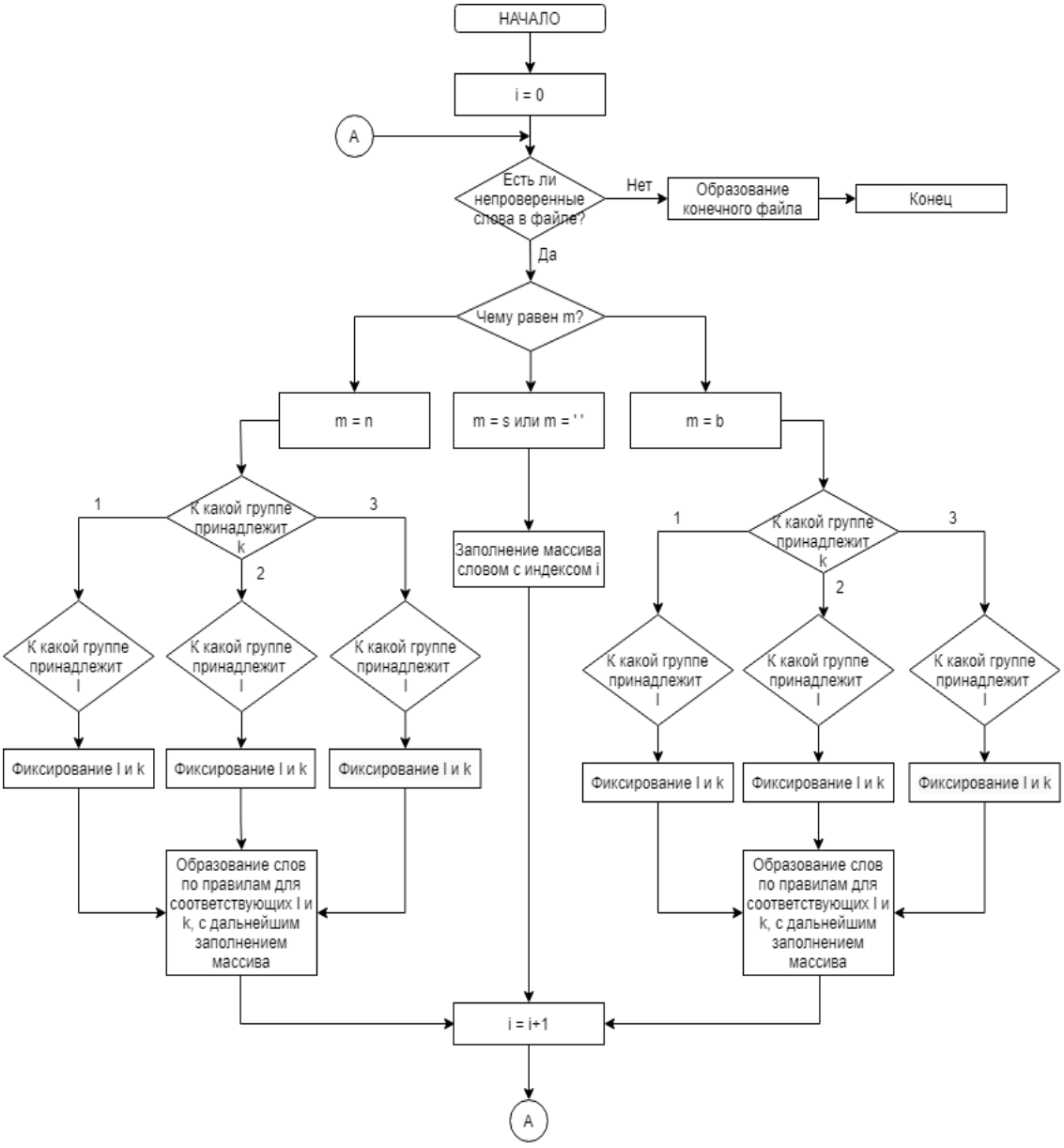

Рис. 3. Блок-схема логики работы приложения

Было принято решение выдавать итоговый файл в текстовом формате txt. Но для работы с начальным файлом, возникла необходимость в дополнительных инструментах и bash-скриптах. Таким образом, из

начального файла в формате xlsx с помощью конвертера для Linux xlsx2csv мы получаем файл в формате csv. Данный формат можно открыть с помощью любого текстового редактора.

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).

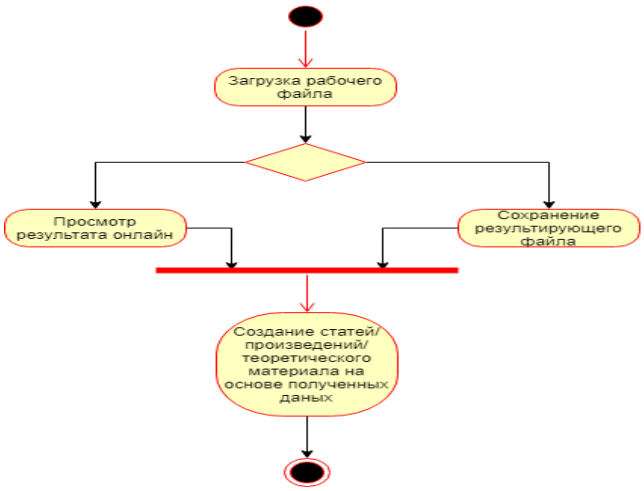

Рис. 4. Диаграмма активности авторов

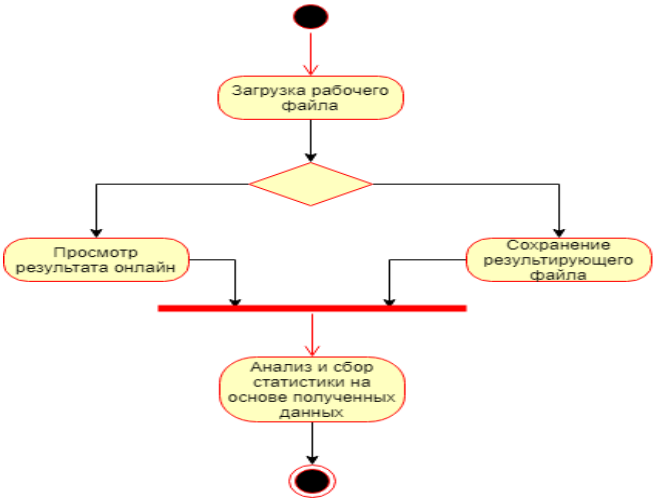

Рис. 5. Диаграмма активности заинтересованных в статистике

2.3. 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. Структура таблиц для разрабатываемой ИС показаны ниже (рис. 6).

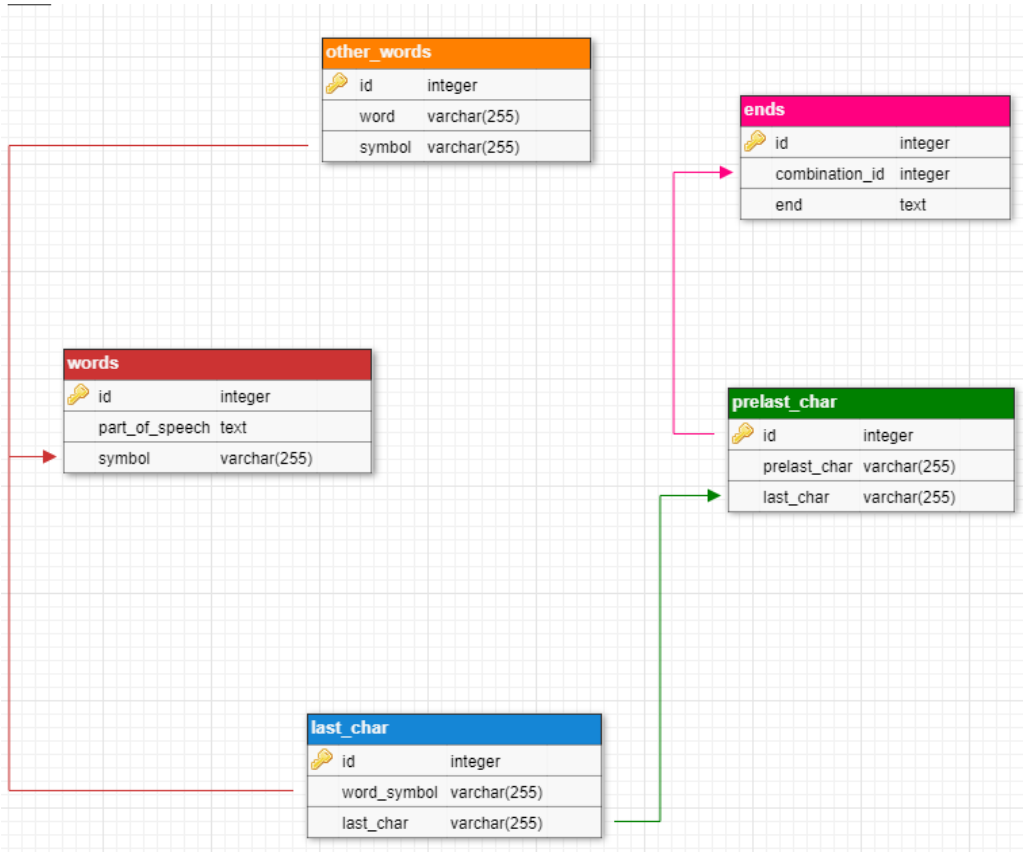

Рис. 6. Структура таблиц базы данных

Заключение

Результатом проделанной работы можно считать выполнение поставленных задач. В конечном итоге мы получили веб-приложение в сети Иинтернет, 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. Были решены следующие задачи:

– проанализирована функциональность аналогов разработанной системы;

– разработаны диаграммы активности и спроектирована база данных для информационной системы;

– разработан алгоритм работы информационной системы;

– разработана блок-схема для описания логики работы приложения;

– реализовано веб-приложение.

Литература

1. Hobson L., Cole H., Hannes H. Natural Language Processing in Action - Munning, 2019. 544 p.

2. Строцев В.А.. Информативность частотных характеристик N-грамм текстовых фрагментов // Инженерный вестник Дона, 2013, №1. URL: ivdon.ru/ru/magazine/archive/n1y2013/1492.

3. Белоногов Г.Г. Языковые средства автоматизированных информационных систем / Белоногов Г.Г., Кузнецов Б.А.. – М.: Наука, 1983. 288 с.

4. Oliveira R.A., Junior C. M. Experimental Analysis of Stemming on Jurisprudential Documents Retrieval // Information, 2018, 9, 28. URL: mdpi.com/2078-2489/9/2/28.

5. Иванова Д.Н., Яровая Л.Е.Модели анализа словообразования в современном английском языке // Инженерный вестник Дона, 2020, №8. URL: ivdon.ru/ru/magazine/archive/n8y2020/6584.

6. Nuriev S.I., Gazizova A.I., Minyazev R.Sh. Searching inside binary and text files Материалы конференций ГНИИ «НАЦРАЗВИТИЕ». 2019. С. 271-274.

7. Минязев Р.Ш., Дыганов С.А., Гумеров И.Р., Перухин М.Ю. Разработка сервиса для идентификации полей сканированного документа с

использованием библиотеки машинного распознавания tesseract-ocr // Вестник технологического университета. 2018. Т. 21. № 9. Сс. 132-135.

8. Porter M.F. An algorithm for suffix stripping. Program. 14(3). 1980. Pp. 130−137.

9. Лойко В.И. Структуры и алгоритмы обработки данных. Учебное пособие для вузов.– Краснодар: КубГАУ. 2004. 261 с.

10. Линник, Ю.В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений / Линник Ю.В. – М.: Физматгиз, 1958. – 336 с.

References

1. Hobson L., Cole H., Hannes H. Natural Language Processing in Action Munning, 2019. 544p.

2. Strocev V.A.. Inzhenernyj vestnik Dona, 2013, №1. URL: ivdon.ru/ru/magazine/archive/n1y2013/1492.

3. Belonogov, G.G. Yazy`kovy`e sredstva avtomatizirovanny`x informacionny`x system [Language tools of automated information systems] Belonogov G.G., Kuzneczov B.A. M.: Nauka, 1983. 288 p.

4. Oliveira R.A., Junior C. M. Experimental Analysis of Stemming on Jurisprudential Documents Retrieval Information, 2018, 9, 28. URL: mdpi.com/2078-2489/9/2/28.

5. Ivanova D.N., Yarovaya L.E. Inzhenernyj vestnik Dona, 2020, №8. URL: ivdon.ru/ru/magazine/archive/n8y2020/6584.

6. Nuriev S.I., Gazizova A.I., Minyazev R.Sh. Materialy` konferencij GNII «NACzRAZVITIE». 2019. P. 271-274.

7. Minyazev R.Sh., Dy`ganov S.A., Gumerov I.R., Peruxin M.Yu. Vestnik texnologicheskogo universiteta. 2018. V. 21. № 9. pp. 132-135.

8. Porter M.F. An algorithm for suffix stripping, Program, 14(3), 1980, pp. 130−137.

9. Lojko V.I. Struktury` i algoritmy` obrabotki danny`x. Uchebnoe posobie dlya vuzov [Data processing structures and algorithms. Textbook for universities]. Krasnodar: KubGAU. 2004. 261 p.

10. Linnik Yu.V. Metod naimen`shix kvadratov i osnovy` matematiko-statisticheskoj teorii obrabotki nablyudenij [The method of least squares and the foundations of the mathematical and statistical theory of observation processing]. M.: Fizmatgiz, 1958. 336 p.