Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.891.2, 004.42

Гоглев

Иван Валентинович

МЕТОДЫ ПРИКЛАДНОГО АНАЛИЗА И ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ

АВТОРЕФЕРАТ

на соискание степени магистра инженерных наук

по специальности 1-40 80 02 Системный анализ, управление и обработка информации

_____

Научный руководитель

Герман Олег Витольдович

кандидат технических

наук, доцент

_____

Минск 2022 г.

КРАТКОЕ ВВЕДЕНИЕ

В наше время большую популярность приобретают различные системы обработки естественного языка. В первую очередь это связано с естественным развитием рыночных отношений в современном глобальном мире. Постепенно системы со сложными интерфейсами управления уступают место системам с интуитивно понятным управлением.

В связи с бурным ростом сферы информационных технологий появилось такое отдельное направление как обработка естественного языка (*Natural Language Processing, NLP*). Одним из самых популярных методов обработки текстовой информации является векторизация. Данный метод заключается в преобразовании текста в формат числового вектора и дальнейшее обучение модели. Таким образом, обработка теста сводится к сравнению векторов.

Все больше компаний и юридических департаментов задумываются о том, как избавиться от рутинной работы, которая отбирает массу времени. На сегодняшний день действует много программ акселерации, как в корпоративном, так и в государственном секторе, которые в той или иной степени затрагивают тему автоматизации юридической функции.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования. Целью данной магистерской диссертации является разработка прототипа программного обеспечения для использования метода векторизации для анализа текстовых данных на естественном языке. Для достижения поставленной цели необходимо решить следующие задачи:

1. Обзор существующих методов анализа текстовых данных;
2. Обзор методов обработки текстовых данных;
3. Разработка юридической системы на основе метода векторизации;
4. Анализ разработанной системы.

Новизна полученных результатов. В ходе магистерской работы был разработан прототип автоматической юридической системы с использованием алгоритма векторизации *TF-IDF*. Также был разработан метод лемматизации текстовых данных с помощью функционала библиотеки *Pymorphy2*.

Положения, выносимые на защиту. В результате работы были получены следующие результаты:

1. Был разработан метод лемматизации текстовых данных на русском языке с использованием методов библиотеки обработки тексовых данных *Pymorphy2*;
2. Был представлен способ анализа текстовых данных через попарное сравнение векторов;
3. Был создан прототип автоматической юридической системы с использованием векторного представления текстовых данных *TF-IDF*.

Апробация результатов диссертации. Промежуточные результаты данной магистерской работы были представлены на конференции «Информационные технологии и системы 2021» в секции «Автоматизированные системы обработки информации».

Опубликованность результатов исследования. По теме диссертации опубликована статья: Гоглев, И. Применение метода векторизации для анализа русскоязычной текстовой информации / И. В. Гоглев // Секция «Автоматизированные системы обработки информации»: программа конференции «Информационные технологии и системы 2021».

Структура и объём диссертации. Данная магистерская работа обладает следующей структурой:

1. Анализ зарубежных и белорусских юридических систем;
2. Описание и анализ существующих библиотек и алгоритмов, необходимых для реализации прототипа юридической системы;
3. Программная реализация прототипа юридической системы и графического интерфейса пользователя;
4. Анализ функциональности прототипа. Повышение эффективности работы системы путём использования предварительно обработанных данных.

Полный объём диссертации: 55 страниц.

Количество изображений: 27

Количество таблиц: 4

Количество использованных библиографических источников: 14

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Анализ проблемы и существующих средств её решения. В данной главе производится анализ систем автоматизации юридической деятельности. Анализируются как зарубежные системы, так и белорусские. Зарубежные системы представлены продуктами *Ravel Law* и *StoryBuilder*. Белорусские системы представлены системами "Эксперт", "Нормативка.by" и "Законодательство СНГ Онлайн".

Глава 2. Алгоритмическая реализация. В этой главе представлен краткий обзор методов предварительной обработки и векторизации текстовых данных. Проведён обзор различных библиотек для реализации системы. Также был представлен способ сравнения векторов между собой.

Глава 3. Программная реализация. В главе описана программная реализация юридической системы с использованием библиотек и алгоритмов, описанных в предыдущей главе. Проведено тестирование основных частей системы. Также была показана разработка графического интерфейса пользователя системы.

Глава 4. Анализ работы системы. В главе произведён анализ эффективности работы системы и приведён способ увеличения скорости работы системы за счёт использования предварительно обработанных данных. Также было проведено тестирование изменённой системы.

ЗАКЛЮЧЕНИЕ

В рамках данной работы были получены следующие практические и научные результаты:

1. Проведён обзор современных средств автоматизации юридической деятельности и существующих технологий в данной сфере.
2. Проведён обзор различных методов переработки текстовой информации на естественном языке в векторный вид.
3. Разработан способ нормализации текстовой информации на русском языке с использованием библиотеки *Pymorphy2*
4. Разработан алгоритм поиска ответа на вопрос через попарное сравнение векторов предложений.
5. Оптимизирован алгоритм поиска ответа при помощи использования предварительно нормализованных текстовых данных.

Разработан прототип юридической вопросно-ответной системы на основе векторизации с использованием метода *TF-IDF* и нормализации на основе методов библиотеки *Pymorphy2*.

Научная новизна диссертационной работы:

- Создан прототип вопросно-ответной системы на основе метода векторизации;
- Создана система для предварительной обработки текстов (приведение всех слов к нормальной форме и удаление стоп-слов);

Разработан метод предварительной обработки текстовых данных с использованием нормальной формы слов.

Результатом данной диссертации является алгоритм вопросно-ответной системы. Разработанный алгоритм может применяться для создания вопросно-ответных систем на разных естественных языках в любых сферах человеческой деятельности.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

Гоглев, И. Применение метода векторизации для анализа русскоязычной текстовой информации / И. В. Гоглев // Секция «Автоматизированные системы обработки информации»: программа конференции «Информационные технологии и системы 2021».