На правах рукописи

Забавин Алексей Сергеевич

Создание системы «База Знаний ТОГУ» с полнотекстовым поиском на основе лексем естественного языка

09.04.04 – Программная инженерия

Автореферат  
магистерской диссертации

Хабаровск – 2025

Общая характеристика работы

Диссертация посвящена вопросам обработки естественного языка и методиках поиска и индексации текстовых данных.

Актуальность темы исследования

Актуальность продиктована необходимостью разработки автоматического ассистента службы поддержки ТОГУ, который бы с большей релевантностью предоставлял пользователю ответы на поставленные вопросы. Существующие сторонние разработки менее ориентированы на контекст работы и информационную специфику вуза и требуют больших усилий по интеграции.

Цель и задачи исследования

Цель данного исследования состоит в разработке программного средства, использующего RNN искусственную нейронную сеть семейств Word2vec для синтаксического и семантического анализа и оптимизации поисковых запросов.

Научная новизна:

* разработанная программа и модель Word2Vec осуществляет хранение и полнотекстовый поиск ответов на вопросы на естественном языке;
* предложен метод оптимизации поискового запроса улучшает качество поиска в базе знаний и может в дальнейшем может быть использован для генерации RAG контекста к современным LLM для создания виртуального ассистента для студентов ТОГУ.

Теоретическая и практическая значимость

Предложенные модели позволяют осуществлять синтаксический, морфологический и семантический анализ предложений

Методология и методы исследования

В работе использована теория искусственных нейронных сетей, методы обработки естественного языка, в том числе методы векторизации слов в пространстве языка с помощью моделей со скользящим окном.

Основные положения, выносимые на защиту:

1. Использование полнотекстового PostgreSQL поиска позволяет эффективнее индексировать и искать информацию в документах текстового типа;
2. Применение синтаксического и семантического анализа пользовательского запроса позволило улучшить поисковую выдачу путем оптимизации исходного запроса;
3. Разработанный метод оптимизации запроса позволяет поднять качество релевантных ответов в два раза перед полнотекстовым поиском без модификаций.

Апробация работы и публикации

Основные результаты докладывались на конференции «VI региональной научно-практической конференции» (г. Хабаровск, 2025). По материалам работы опубликованы статьи в сборнике СНПК 2025.

Объем и структура диссертации

Диссертация состоит из введения, трех глав, заключения, списка использованных источников и двух приложений. Полный объем диссертации составляет 84 страницы, включая 22 рисунка и 5 таблиц. Список использованных источников содержит 33 наименований.

Содержание работы

**Во введении** описывается актуальность исследований, определяются цель и задачи, обосновывается научная новизна, теоретическая и практическая значимость работы.

**Первая глава** посвящена изучению предметной области. Приведены основный понятия, развернутое описание аспектов полнотекстового поиска, метрик релевантности ответов, и методов обработки естественного языка. Дается определения векторных представлений слов или эмбеддингов, способов их получения что является теоретической основой исследования и используется в разработанных алгоритмах.

**Во второй главе** приведено описание и схемы работы для разработанных алгоритмов синтаксического анализа запроса, выявление основной части запроса и оптимизации по семантической близости и TF-IDF. Данные алгоритмы используются в практической части исследования (в третьей и четвертой главах).

**Третья глава** наглядно демонстрирует пример работы программы, разработанной для хранения базы знаний и поиска в ней. Программа демонстрирует разложение хранимых вопросов и ответов на лексемы, расчет важности каждой лексемы в базе. В программе представляется граф синтаксического разбора, вводимого пользователем запроса и анализ семантическим близких запросу лексем. Рассмотрены результаты тестирования разработанных алгоритмов применительно к измерению качества (релевантности) найденных в базе ответов на заданные запросы. Приведено подробное описание технической реализации программы.

**В заключении** приведены основные результаты, полученные в диссертационном исследовании.

**Приложение А** включает исходный код реализованных классов и функций разработанных программ.

**Приложение Б** содержит проверку на заимствования.

Публикации автора по теме диссертации

1. Забавин А.С. Исследование технологий морфологического и семантического анализа и их применение в задаче информационного поиска / Забавин А.С., Э. М. Вихтенко // СНПК: III региональная студенческая научно-практическая конференция : Материалы III Региональной студенческой научно-практическая конференции, Хабаровск, 28–29 апреля 2025 года. – Хабаровск: Тихоокеанский государственный университет, 2025. – (готовится к публикации).