УДК 004.891.2

РАЗРАБОТКА ЭКСПЕРТНОЙ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ И ВЕКТОРНОЙ БАЗЫ ДАННЫХ

Кучер М. (ИТМО),
Научный руководитель – кандидат технических наук, Доцент Кугаевских А. В.
(ИТМО)

Введение. В современной эпохе информационных технологий разработка экспертной вопросно-ответной системы, основанной на больших языковых моделях и векторной базе данных, представляет собой значимый прорыв в области прикладного искусственного интеллекта и сферы поиска информации. Целью проекта является создание высокоэффективной системы, способной обрабатывать узкоспециализированные запросы и формировать точные ответы, опираясь на обширные массивы текстовых данных из конкретных предметных областей. Это позволяет значительно повысить качество и скорость предоставления информации пользователям.

Основная часть. Разработка системы включает несколько ключевых этапов, каждый из которых играет решающую роль в создании эффективного и надежного продукта.

Анализ и подготовка данных. На первом этапе осуществляется сбор и предварительная обработка текстовых данных из выбранных предметных областей. Это включает в себя очистку данных от не релевантной информации, структурирование и классификацию данных для обеспечения их эффективного использования в дальнейшем.

Интеграция больших языковых моделей. Центральным элементом системы является использование продвинутых языковых моделей, таких как GPT (Generative Pre-trained Transformer), которые способны анализировать запросы пользователей, понимать контекст и генерировать соответствующие ответы. Для достижения высокой степени экспертности ответов модели обучаются на специализированных данных, что позволяет им выявлять сложные зависимости и интерпретировать запросы с учетом специфики предметной области.

Разработка и интеграция векторной базы данных. Эффективное хранение и быстрое извлечение информации обеспечивается за счет использования векторной базы данных. Это позволяет системе мгновенно находить необходимые данные среди огромных объемов информации, опираясь на векторные представления текста. Механизмы поиска и ранжирования оптимизируются для обеспечения высокой скорости ответов и их максимальной релевантности запросам пользователей.

Завершающий этап включает в себя интеграцию системы в целевую информационную среду - телеграм. Для удобного поиска с компьютера либо мобильного устройства.

Выводы. Результатом проекта является экспертная вопросно-ответная система, которая обладает способностью к глубокому пониманию запросов и выдаче обоснованных ответов, опираясь на специфические знания из предметной области.

Список использованных источников:

1. Гудфеллоу И., Бенджио Ю., Курвилль А. Глубокое обучение. – MIT Press, 2016.
2. Радфорд А., Ву Дж., Чайлд Р., Луан Д., Амодеи Д., Сатскевер И. Языковые модели — это неконтролируемые многозадачные ученики. – OpenAI Blog, 2019.
3. Рассел С., Норвиг П. Искусственный интеллект: Современный подход. 4-е изд. – Pearson, 2021.