УДК 004.8

РАЗРАБОТКА ВОПРОСНО-ОТВЕТНОГО МОДУЛЯ ДЛЯ АГРЕГАТОРА ПЕРСОНАЛИЗИРОВАННЫХ НОВОСТЕЙ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Терёшкина Н.А.[1], Федоров Д.А.[1]
Научный руководитель – канд. техн. наук, доцент Федоров Д.А.[1]
[1]Университет ИТМО
na.teryoshkina@mail.ru

Введение

С развитием тренда на оптимизацию и автоматизацию рутинных процессов у людей возникла потребность в быстром доступе к структурированной информации. Эта потребность особенно остра при работе с динамичными новостными потоками. Прорывной технологией, предлагающей решение этой проблемы, является архитектура Retrieval-Augmented Generation (RAG), которая объединяет поиск по актуальным данным и генерацию ответов на естественном языке [1]. Применение подобных решений стремительно набирает популярность по всему миру, в том числе и на российском рынке из-за наиболее эффективной работы с данными [2]. Вопросно-ответные системы, построенные на RAG архитектуре, становятся неотъемлемой частью последних разработок и находят наибольшее применение в медиапространстве для анализа потоковых новостей. Однако существующие коммерческие решения, включая зарубежные аналоги, часто являются закрытыми для широкого использования, не предоставляют персонализации источников и не гарантируют прозрачности предоставляемой информации, что создает нишу для разработки доступного и адаптируемого решения.

Основная часть

Суть предлагаемого исследования заключается в разработке открытого вопросно-ответного модуля, интегрированного в агрегатор персонализированных новостей. Модуль основан на RAG-архитектуре и реализует следующий пайплайн:

– Автоматизированный сбор новостей из заданных пользователем источников;

– Семантическая индексация контента с использованием моделей векторных представлений (эмбеддингов);

– Интеллектуальная обработка запроса, предполагающая поиск релевантных новостей и генерацию краткого обоснованного ответа большой языковой моделью;

– Предоставление результата с обязательным цитированием источника.

Научная и практическая новизна заключается в адаптации RAG-подхода для потоковых новостных данных с обеспечением высокой актуальности информации [3], внедрении модели доверенных источников и создании гибридного интерфейса. Практическая значимость заключается в решении проблемы отсутствия подобных доступных и персонализированных инструментов на отечественном рынке. Технология автоматизирует длительный процесс ручного поиска и анализа, значительно сокращая время на получение ответа и снижая когнитивную нагрузку на пользователя.

Выводы

В результате будет разработан и протестирован функциональный прототип системы, представляющий ценность, как для конечных пользователей, так и для медиа индустрии.

Литература

1. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv. — 2020. — URL: https://arxiv.org/abs/2005.11401 (дата обращения: 19.11.2025)
2. Мельников А.В., Николаев И.Е., Русанов М.А., Аббазов В.Р. Сравнительный анализ методов RAG для построения русскоязычных интеллектуальных сервисов // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. - 2025. - №2 - С. 5-18.
3. Герасимов А. С., Бобков С. П. Разработка самообучающейся системы обработки естественного языка с динамической реорганизацией знаний // Известия высших учебных заведений. Серия: Экономика, финансы и управление производством - 2025. - №3 (65) - С. 77-84.