

УДК 51-74

Создание интеллектуального ассистента-помощника на основе RAG-системыД.А. Редько, Е.М. Сухова

Научный руководитель: доцент, к.ф.-м.н., М.Е. Семенов

Национальный исследовательский Томский политехнический университет,

Россия, г. Томск, пр. Ленина, 30, 634050

E-mail: dar27@tpu.ru, ems16@tpu.ru**Creating an intelligent assistant based on the RAG system**D.A. Redko, E.M. Suhova

Scientific Supervisor: Ass. Prof., PhD, M.E. Semenov

Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050

E-mail: dar27@tpu.ru, ems16@tpu.ru

Abstract. *This paper presents the development of an intelligent assistance system aimed at users, utilizing the Retrieval-Augmented Generation (RAG) model. The proposed project aims to provide personalized assistance in a specific domain, based on a knowledge base. By integrating RAG, the model improves the quality of responses by incorporating relevant information from external sources. The software implementation of the project has been carried out using the DeepSeek AI API.*

Key words: *Retrieval Augmented Generation, Large Language Model, neural networks.*

Введение

В современном информационном мире объем данных становится огромным, что создает проблему информационного шума и перегрузки. Стремление к освоению всей доступной информации становится нереалистичным. Поэтому возникает потребность в инструментах, способных агрегировать и фильтровать этот поток информации, предоставляя только самое важное и релевантное. Создание таких интеллектуальных систем, способных адаптироваться к потребностям конкретного пользователя и предоставлять персонализированную информационную поддержку, становится крайне важным.

В этой работе представлена разработка интеллектуальной системы помощи, ориентированной на студентов, с использованием генеративной модели, использующей дополнительно найденную релевантную информацию (Retrieval Augmented Generation, RAG) [1]. Используя комбинацию больших языковых моделей (LLM) и модели RAG, создан чат-бот, способный проводить диалог с пользователем на естественном языке и предоставлять информацию, учитывая контекст запроса и актуальные данные. Предложенный подход направлен на значительное улучшение пользовательского опыта и обеспечение максимально удобного взаимодействия с системой.

Цель работы – создание специализированного чат-бота, способного предоставлять пользователям актуальную информацию из базы знаний и внешних источников.

Экспериментальная часть

Для работы модели RAG была подготовлена база знаний, состоящая из информации, собранной с официального сайта Томского политехнического университета. Данный набор данных включает в себя разнообразную информацию (адреса корпусов, локально-нормативную документацию), необходимую для создания интеллектуального ассистента-помощника для студентов.

Для сбора данных были использованы техники парсинга веб-страниц, позволяющие извлекать информацию из HTML-кода сайта университета. Полученные данные были обработаны и структурированы в формате, пригодном для использования модели RAG.

Выбор модели. При выборе модели мы учитывали особенности предметной области и требования проекта. Из литературы известно множество LLM: LLaMA, GigaChat, GPT, SAIGA, DeepSeek AI. Мы сделали выбор в пользу предобученной модели DeepSeek AI [2], которая снабжена API.

Архитектура модели RAG

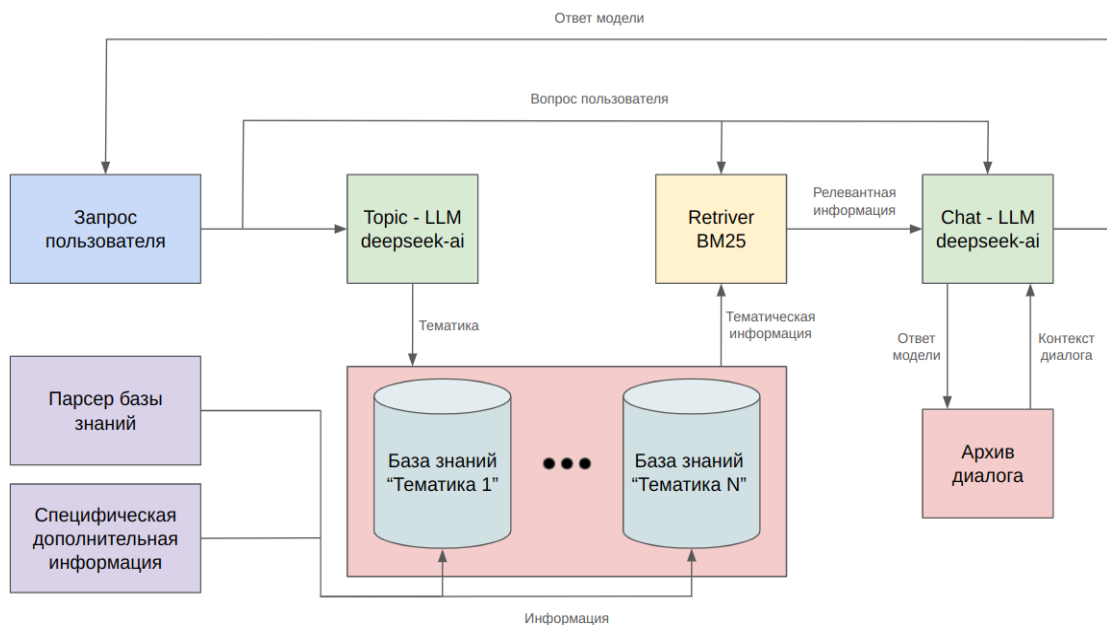


Рис. 1. Retrieval Augmented Generation

Предложенная модель (рис. 1) состоит из базы знаний, которая заранее пользователем разделена на тематики. Запрос пользователя попадает в LLM модель, которая возвращает из предложенного списка доступных тематик наиболее близкую. Далее полученная тематика попадает в ретривер, который при помощи метода BM25 [3] определяет наиболее близкие контекстуальные части из базы знаний. BM25 является вероятностной моделью информационного поиска, основанной на модели вероятности, которая оценивает релевантность документов по отношению к запросу пользователя.

Функция ранжирования BM25 вычисляется следующим образом:

$$BM25(D, Q) = \sum_{i=1}^n IDF(q_i) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b \frac{|d|}{avgdl})}$$

где $Q = (q_1, q_2, \dots, q_n)$ – вектор слов, $TF(q_i)$ – частотность слова q_i в документе D , $f(q_i, D)$ – частота употребления слова q_i в документе D , $|D|$ – количество слов в документе (длина документа), $avgdl$ – средняя длина документа в коллекции, k_1 и b – свободные коэффициенты (чаще всего 2.0 и 0.75 соответственно),

$$IDF(q_i) = \log \left(\frac{N - f(q_i, D) + 0.5}{f(q_i, D) + 0.5} \right) - \text{обратная частота документа.}$$

Далее полученный контекст из базы знаний попадает на LLM модель, которая ведет диалог с пользователем. LLM модель генерирует ответ на основе переданного контекста, вопроса и истории диалога и возвращает ответ пользователю.

Результаты

В результате проведенных исследований мы разработали интеллектуального ассистента-помощника, на рис. 2 приведен иллюстративный пример, где продемонстрированы типичные

вопросы. В результате взаимодействия с ассистентом, в каждом из случаев, были получены релевантные ответы, удовлетворяющие запросы пользователей.

Интеллектуальный ассистент-помощник студенту ТПУ

User: Где находится ГК ТПУ

Bot: Главный учебный корпус Томского политехнического университета находится по адресу г. Томск, пр-кт. Ленина, д.30.

User: Какая у студентов стипендия

Bot: У студентов существует несколько видов стипендий, и их размер зависит от типа образовательной программы, формы обучения и других факторов. Например:

1. Стипендия Президента РФ для студентов, обучающихся по направлению "Высшее образование" - 2 200 руб.
2. Стипендия Правительства РФ для студентов, обучающихся по направлению "Высшее образование" - 1 440 руб.
3. Государственная академическая стипендия для студентов, обучающихся по очной форме обучения за счет средств федерального бюджета, осваивающих образовательные программы бакалавриата, специалитета, магистратуры: Государственная академическая стипендия - 3 693 руб. (без РК).
4. Государственная социальная стипендия - 4 220 руб. (без РК).

Есть также стипендии Президента и Правительства РФ для студентов, обучающихся по направлениям подготовки, соответствующим приоритетным направлениям развития российской экономики. Эти стипендии составляют 7 000 и 5 000 рублей соответственно.

User: Где находится 10 корпус тпу?

Bot: Учебный корпус № 10 Томского политехнического университета находится по адресу г. Томск, пр-кт. Ленина, д.2.

User: А где посмотреть расписание занятий, можешь скинуть ссылку?

Bot: Расписание занятий ТПУ для студентов доступно на официальном сайте университета по ссылке: <https://rasp.tpu.ru/>.

Введите ваше сообщение Отправить

Рис. 2. Пример использования чат-бота

Заключение

В рамках исследования был создан интеллектуальный ассистент-помощник для студентов на основе RAG-системы с использованием современных методов обработки естественного языка и глубокого обучения. Из недостатков можно выделить требование к обязательной лемматизации, а также размер документа также может плохо сказываться на производительности, так как большое количество слов в рассматриваемом тексте приводит к долгим вычислениям. Данное решение легко адаптируемо и может быть применено для любых компаний и организаций, обеспечивая персонализированную информационную поддержку и улучшая взаимодействие с клиентами на основе их запросов и потребностей. Для адаптации под конкретную компанию достаточно заменить базу знаний на соответствующую предметную область.

Список литературы

1. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Computer Science. Computation and Language. – 2020. – URL: <https://arxiv.org/abs/2005.11401>
2. DeepSeek // URL: <https://www.deepseek.com/> (дата обращения: 22.02.2024).
3. Белова К.М., Судаков В.А. Исследование эффективности методов оценки релевантности текстов // Препринты ИПМ им. М.В.Келдыша. – 2020. – № 68. – 16 с. – URL: https://keldysh.ru/papers/2020/prep2020_68.pdf