Отчёт по заданию №5

Выполнила: Учар Айгуль, 324 группа

Выбранный ресурс для построения RAG системы: учебник по ML с сайта Яндекс.Практикума https://education.yandex.ru/handbook/ml

Программная реализация

Парсинг производится в отдельной программе parser.py. Используется библиотека BeautifulSoup: извлекался контент из блоков “article-content”, удаляется служебная информация.

Остальной код реализован в ноутбуке LLM_task5.ipynb. Текст разбивается на чанки по 512 символов с сохранением структуры.

Векторная БД реализована на Qdrant с моделью эмбеддингов “sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2”. Для генерации ответов используется модель TinyLlama/TinyLlama-1.1B-Chat-v1.0.

Анализ ответов

С RAG ответы точнее и специализированнее. Например, на вопрос «Что такое ROC-кривая?» модель с RAG дала определение с упоминанием TPR/FPR и AUC, а без RAG — общее описание «графика для оценки классификаторов».

Без RAG часто возникали одинаковые проблемы:

- Ответы на испанском/английском (например: «El coeficiente de fuerza...»)
- Рекурсивные повторы («функция — это функция, которая...»)
- Уход в другие области знаний (энтропия рассматривалась как физический термин, а не термин из области ML).

С RAG таких ошибок не было. большинство ответов содержали термины из учебника (например, «опорные векторы», «градиентный бустинг»), даже если ответ генерировалсся на английском он был более релевантным, чем без использования RAG (вопросы 11, 12, 16). Также ответы с RAG были более полными (13 вопрос). Ответы были связаны именно с машинным обучением, а не с другими областями (вопросы 17, 19).

Выводы:
Реализация RAG-системы значительно улучшила качество ответов: ответы с RAG демонстрируют точность, узкую специализацию в ML и отсутствие критических ошибок, рекурсивных повторов и ухода в смежные области.
Модель TinyLlama, имеющая всего 1.1B параметров, без использования RAG давала очень несвязные ответы, часто встречались рекурсивные повторы или отклонения от темы, добавление же векторной БД позволило выдавать более релевантные ответы.