

УДК 004.8:004.056.55

В. Готович, к. т. н., доц; В. Попович

(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ДОСЛІДЖЕННЯ ТА РОЗРОБКА AI-АСИСТЕНТА НА ОСНОВІ МОДЕЛІ MISTRAL ДЛЯ СЕРЕДОВИЩА УНІВЕРСИТЕТУ

UDC 004.8:004.056.55

V. Hotovych, PhD., Assoc. Prof.; V. Popovych

RESEARCH AND DEVELOPMENT OF AN AI ASSISTANT BASED ON THE MISTRAL MODEL FOR THE UNIVERSITY ENVIRONMENT

Сучасні університети оперують великим обсягом інформації, включно з навчальними планами, розкладами, документами та новинами. Часто ці дані розпорошені по різних розділах вебсайту та неструктуровані, що ускладнює швидкий доступ до потрібної інформації. Створення AI-консультанта дозволяє централізувати знання, автоматизувати пошук та надавати студентам і викладачам точні відповіді на запитання. Моделі сімейства Mistral мають відкритий код, високу продуктивність та можливість інтеграції з векторними базами знань, що робить їх оптимальним вибором для розробки системи на базі Retrieval-Augmented Generation (RAG).

Метою роботи є розробка AI-асистента, який забезпечує ефективний пошук і надання релевантної інформації з внутрішньої бази університету, сформованої на основі даних сайту, з використанням можливостей моделі Mistral та механізму RAG. Робота передбачає створення системи збору даних із вебсайту університету, розробку власної бази знань, інтеграцію її з моделлю Mistral через RAG та реалізацію інтерфейсу для взаємодії користувачів із системою.

Для реалізації проекту застосовується комбінація сучасних методів машинного навчання та технологій обробки природної мови. Перший етап включає парсинг вебсайту університету із автоматичним вилученням тексту та HTML-структур, а також конвертацією документів у формат, придатний для подальшого аналізу. Отримані дані проходять очищення, нормалізацію та розбиття на семантичні блоки (chunking). Для створення індексованої бази знань застосовується векторне представлення тексту (embeddings), що дозволяє виконувати семантичний пошук за допомогою інструментів на кшталт FAISS або ChromaDB. Використання RAG забезпечує поєднання реального пошуку по базі даних із генеративними можливостями моделі Mistral. Це дозволяє AI-консультанту формувати точні та контекстно-залежні відповіді на запитання користувачів, базуючись на актуальній інформації університету. Крім того, система інтегрує API для взаємодії з користувачами, що може бути реалізовано через веб-інтерфейс, чат або інші канали комунікації.

Розроблений AI-асистент підтвердив на практиці ефективність поєднання моделей Mistral з технологією RAG для обробки внутрішніх даних університету.

Література

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, 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems (NeurIPS 33), pp. 9459–9474.
2. Nils Reimers, Iryna Gurevych, 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992.
3. Yucheng Hu, Yuxing Lu, 2024. *RAG and RAU: A Survey on Retrieval-Augmented Language Models in Natural Language Processing*.
4. Akari Asai, Matt Gardner, Hannaneh Hajishirzi, 2021. *Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks*