Development of an Article Retrieval System

25 kwietnia 2024

1 Introduction

A system was created to index articles from the ”1300 Towards Data Science Medium Articles” dataset, which uses Retrieval Augmented Generation for the effective retrieval of article fragments in response to queries. This report covers the system’s design, selected technologies, encountered challenges and areas for future development.

2 System’s Design

The system was implemented in two files: *main.ipynb* and *app.py*. In the first one the system was created including loading and chunking the data, generating embedding and creating Chroma database, while the *app.py* contains an application that allows you to enter prompts and returns relevant fragments of articles based on the similarity scores. The application is designed to facilitate the use of the system and the reading of received fragments.

3 Selected Technologies

While working on this project, I chose to primarily use the following Python libraries: LangChain becouse of its ease of use and becouse it offers a number of text processing functions and Streamlit because it is a library specifically designed for machine learning engineers and requires no web development knowledge, making it easier to use.

Regarding the vector store, I chose ChromaDB for its ability to handle semantic search, which allows to quickly and accurately find the most relevant information.

4 Encountered Challenges

The biggest challenges I encountered were related to fees for using the APIs, e.g. I was interested in using OpenAI embeddings but due to the length of the documents I was working with, I had to stay with the free ones.

5 Areas for Future Development

In the future, I plan to expand the current system with a Q&A application, so that in response I do not receive only fragments, but a coherent answer based on documents.

I would also like to work on an application designed for entering queries and reading answers, so that previous questions and answers can be seen there, which will improve the comfort of using it.