Lazăr Cătălin-Mihai

Retrieval Augmented Generation with Vector Database

This aim of this project is to implement a better searching method from a large source of text data by using a type of database called a vector database (VDB).

The text data is split into chunks and inserted into the database. Each entry contains the text chunk, a random string id but also an embedding vector. This embedding vector is the key element of the whole process.

Embeddings are representations of words as real valued vectors that are made in such a way that the words that are closer in the vector space are similar in meaning. The embeddings are done by an open-source AI model that was trained on a large text corpus.

When the user wants to retrieve some information from the VDB, the question is also encoded and the question vector is compared to each encoding in the VDB. The similarity search returns a number and the larger the number is, the greater the relevance of that particular chunk is to the user question.

Thus, this type of search is better than a simple keyword search because the meaning of the question is considered.

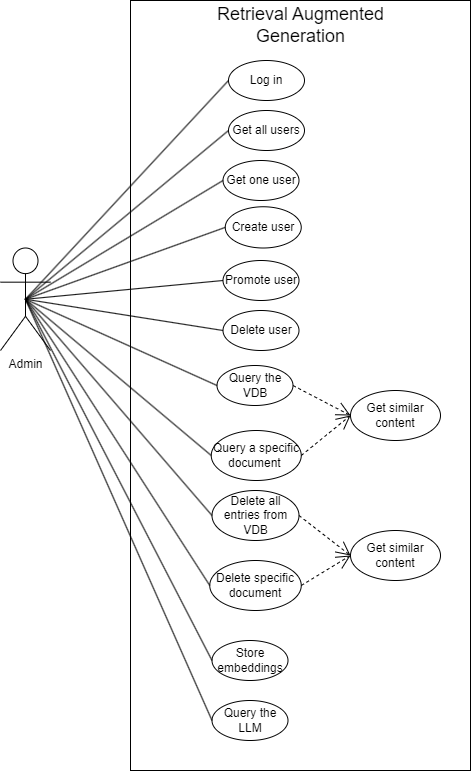
The project is a web application that takes the question of the user from a text field and returns the results from the VDB. A Large Language Model that responds to the query, based on the results returned from the VDB is also considered.

The application calls a local Flask endpoint that takes the query. For this project, the VDB is Azure AI Search (but might be changed for an open source DB in the future).

The data added in the VDB has to be processed beforehand and it is stored in an Azure Blob Storage.

For the login part of the project, the SQL Alchemy toolkit is used in order to simplify the connection between Python and an SQLite database where the users are stored. There are 2 types of users: admin and non admin. Only the admin user has access to the VDB operations. The admin can also add or remove other users, or promote a user to be an admin.

Use case diagram



Database diagram

A black and white rectangle

Description automatically generated