Simple RAG Project

Przemysław Pietrzak

April 9, 2024

1 System Design Technologies Selected

Vector Database

I utilized MongoDB Atlas database, configuring the connection to be accessible only by the admin user. The code facilitates the use of the built-in vector search capability of the database. While most of the database setup can be replicated using the provided code, setting the search index requires a separate step, with the search index definition located in the project files under data/SemanticSearchIndex.json.

Retrieval

The retrieval mechanism is implemented within MongoDB Atlas. For simplicity, I utilized embeddings from the SentenceTransformers package with a hidden dimension size of 384. Following embedding, a dot product is computed between the query and prospective chunk embeddings (with the embedding vectors normalized beforehand). Prior to embedding, articles are split into chunks, with an aim to minimize the number of chunks.

2 Challenges Encountered Potential Areas for Future Development

2.1 Challenges

A significant portion of my time (60%) was consumed by database-related tasks, leaving 20% for refactoring and 20% for the primary task. I aim to better allocate my time in the future by implementing custom search methods or refining chunking strategies.

2.2 Improving Chunking

Hierarchical chunking may prevent splitting sentences in the middle, offering a potential improvement. Additionally, incorporating semantic knowledge into text splitting could enhance results, albeit at an increased computational cost.

2.3 Improving Embeddings

Utilizing a more robust model and fine-tuning chunking hyperparameters could enhance results. Currently, I employed only small, locally-run open-source models and minimized the number of chunks by splitting articles into chunks of the maximum number of tokens.

2.4 Improving Retrieval

While my retrieval method relied on a pre-existing solution, implementing more sophisticated search methods, such as hierarchical searching using clusters, could yield better outcomes.