Alex Schell

COMP SCI 5588

Week 4 HandsOn Reflection

This week's hands-on assignment helped me understand a few concepts by introducing them in the notebooks. These include Vector database, Embedding, Chunking, RAG, Chroma and LangChain.

A vector database is a database that will store manage and search for vector embeddings. Embeddings are numerical representations of data (text in our case) that capture semantic meaning of the data. The database uses fast similarity searches to find the data that are close to meaning in a query. The vector database used in the notebook was Chroma.

Chunking experiments in the notebook have shown me how it works. The tests with 300 chunks and 500 chunks were similar in this case but that may be due to the similarity of the datasets I used. One thing I was impressed with was that none of the tests tried to hallucinate an answer for my last question about fixing a broken chair leg. In this case, I could see that using smaller chunks ran much faster. If the entire dataset is more similar than not, this will likely be a good strategy for our project.

The RAG is the culmination of all these moving parts of using the vector database. LangChain is the python framework that basically runs RAG.