LLM

Monday, February 24, 2025 9:43 AM

Retrival Augmented Generation (RAG)

Class notes --> generate embeddings
    Chunk {FileName, PDF page, Vector}

    --> Redis Stack

User asks "What is SQL?"

Generate an embedding (vector) for this question
Send it back to Redis and retrieve the n most similar chunks
Send 10 most similar chunks as context to Mistral model
    "Act like a smart human and use this context to answer the question. If the context doesn't include an answer say 'I don't know'"
We then get an answer that is contextual to our class notes