Reading: Summary and Highlights

Congratulations! You have completed this lesson. At this point in the course, you know that:

- RAG is an AI framework that helps optimize the output of large language models or LLMs.
- RAG combines retrieved information and generates natural language to create responses.
- RAG consists of two main components: the retriever, the core of RAG, and the generator, which functions as a chatbot.
- In RAG process:
    - The retriever encodes user-provided prompts and relevant documents into vectors, stores them in a vector database, and retrieves relevant context vectors based on the distance between the encoded prompt and documents.
    - The generator then combines the retrieved context with the original prompt to produce a response.
- The Dense Passage Retrieval (or DPR) Context Encoder and its tokenizer focus on encoding potential answer passages or documents. This encoder creates embeddings from extensive texts, allowing the system to compare these with question embeddings to find the best match.
- Facebook AI Similarity Search, also known as Faiss, is a library developed by Facebook AI Research that offers efficient algorithms for searching through large collections of high-dimensional vectors.
- Faiss is essentially a tool to calculate the distance between the question embedding and the vector database of context vector embeddings.
- The DPR question encoder and its tokenizer focus on encoding the input questions into fixed-dimensional vector representations, grasping their meaning and context to facilitate answering them.