Aplicaciones LLM

App RAG básica: QA de un documento

Problema: el límite de la context window

- Los Foundation LLMs están limitados por su context window.
- La versión gratuita de ChatGPT, por ejemplo, no puede asumir un texto de más de 4097 tokens.
- ¿Qué pasa si queremos hacer preguntas sobre un documento más largo de ese límite?

Solución

- Crearemos una Aplicación RAG básica:
  - Dividiremos el documento en pequeños fragmentos.
  - Convertiremos esos fragmentos en números (llamados “embeddings”).
  - Cargaremos los embeddings en una base de datos vectorial.
  - Crearemos un sistema de recuperación utilizando una chain predefinida de LangChain.

Proceso

- Cargar el documento de texto con un document loader.
- Dividir el documento en fragmentos con un text splitter.
- Convertir los fragmentos en embeddings con OpenAIEmbeddings.
- Cargar los embeddings en una base de datos vectorial FAISS.
- Crear una chain RetrievalQA para recuperar los datos.