Informe RAG:

Tecnicatura Universitaria en Inteligencia Artificial

Fecha: 27/02/2024

Integrante:

- Mateo Rovere

Profesor:

- Juan Pablo Manson
- Alan Geary
- Andrea Carolina Leon Cavallo
- Ariel D'Alessandro

Se puede encontrar todo el repositorio en mi github:
https://github.com/Mateorovere/TP2_NLP

RAG (Retrieval-Augmented Generation):

Yo elegí implementar RAG que sea experto sobre la anatomía humana a partir de libros de fuentes confiables, a partir de un archivo csv (que contiene información de los sistemas del cuerpo) y de datos de wikidata.

Para hacer split en los textos use RecursiveCharacterTextSplitter con un chunk_size de 500.

Teniendo en cuenta que el entorno de Colab tiene como límite 13 mil millones de parámetros para los modelos, elegí a la versión de 13b de LLAMA 2, dado que cumplía ese requisito y que tenía buen performance en el lenguaje natural.

El modelo de embedding que elegí es el “intfloat/multilingual-e5-base”, dado que la versión “large” del mismo me daba error por CUDA debido a un OOM por el colab.

Luego a las bases de datos vectoriales estaban con chromaDB