Context-Aware Question-Answering System

Data Collection:

I have selected two PDFs from the internet that encompass a diverse range of information pertaining to the field of environmental studies.

Preprocessing:

To perform data extraction and preprocessing, I employed PyPDFLoader from the Lang Chain Framework.

Creation of Chunks:

I utilized the CharacterTextSplitter function from the Lang Chain Framework to segment the extracted text into chunks.

Contextual Embeddings:

I used open Ai for the creation of the embeddings and Chroma db for storing the embeddings. Vector stores and embeddings come after text splitting as we need to store our documents in an easily accessible format. Embeddings take a piece of text and create a numerical representation of the text. Thus, text with semantically similar content will have similar vectors in embedding space. Thus, we can compare embeddings(vectors) and find texts that are similar.

Q-A Model:

We need to compress the relevant splits to fit into the LLM context. Finally, we send these splits along with a system prompt and human question to the language model to get the answer.

In the process of retrieving relevant answers from the Chroma database, I employed both the OpenAI LLM (Language Model) and the LangChain QA model.