Cerebralzip-Assignment

Workflow:

1. I extracted the text from all the pdfs by first converting them into image and take out the text using tesseract. By this both scanned and non-scanned pdfs can be covered. I put all the data in the csv for future reference.

2. I used foll. Modes:
    a. multilingual-MiniLM-L12 for embedding.
    b. Faiss for indexing the embeddings
    c. xlm-roberta for question answering.
    d. Mbart for summarization.

3. Code takes user query and provides relevant answer and summary.

Challenges:

1. Extracting text from pdf was the tricky one as we had scanned and non scanned pdfs and in non scanned pdfs, there was encoding issue, so if we check csv file, it looks problematic but in code, data looks good.
2. Extracted data is not clean enough and it affected the final summary.