Goals

I am executing the following project.

“Email Search AI: Develop a generative search system for emails that helps organisation find and validate past decisions, strategies, and data in a huge corpus of email threads”

Data Sources

Data Sources as provided on Kaggle on the following link.

https://www.kaggle.com/datasets/marawanxmamdouh/email-thread-summary-dataset

Design

The email thread details file contains the details of all the email threads while the email thread summary file contains the summary of the threads. In the preprocessing steps the goal is to create the subject as the metadata, summary as the document for vector store and email message as additional input for RAG.

As per that design decision, a single message needs to be created with all the threads. However, each thread contains repeating message due to forwarding. Hence all the forwarded text is removed from the message and then the threads are joined together along with the timestamp, from and to fields. Message column is joined with its summary column using the thread id. The subject column is also joined to the summary column using thread id.

This new table is now added to vector store with summary as the documents, thread id as the id and subject as the metadata. The message column is not added to vector store since the token count is too huge.

Once the documents are added to vector store, a sematic search is performed for various queries. The top 10 results generated are further reranked using the cross-encoder model and the message column for those are joined based on thread id. Finally, the top 3 are shortlisted.

The top 3 results are then passed to OpenAI to get an accurate response to the question. Appropriate prompts are provided to OpenAI to provide accurate results.

Challenges Faced

1. I was not able to run this code on local python notebook. The notebook crashed every time I tried to create a collection. I switched to Google Collab.

2. I was not able to add entire list to collection. It gave an input error. I am not sure why the error. However, I switched to adding each document separately to the vector database. That worked fine.
3. It was difficult getting the message and subject column joint with summary column. Typically joining using a for loop is trivial but joining using a dataframe required research.