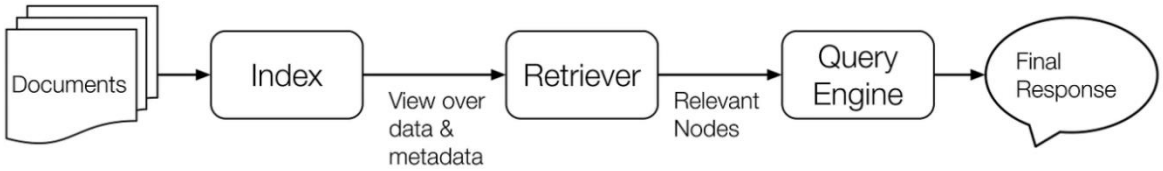
Creating AI Assistant for HDFC Policy Documents Using RAG Pipeline

1. The basic RAG pipeline in Llama Index is illustrated below.



2.We started by importing the necessary Libraries for implementing Llama Index.

3.Then we mounted our google drive to retrieve the data from that location and we set the API Key for authentication at OpenAI Platform.

4.We checked whether the API was working properly and we asked some general query to the Open AI and got a very generic response. Our aim is to improve this response using RAG Pipeline.

5.Again we imported some important libraries and loaded the seven HDFC policy documents, then we counted the number of documents which was 217,to confirm that we all loaded all the pdf files successfully we also checked the number of PDF files and found that to 7,which was correct.

6.Next we built the query engine. In building so, we imported the necessary libraries, we built the parser and parse the documents into nodes, we built the index and finally built the query engine

7.Then we checked the response using query engine and found it to be much more accurate as compared to our earlier response (without RAG pipeline).

8.We checked the source node and meta data of the response.

9.We extracted File name and page no of the response

9.We also extracted the response score and found it to be 0.90,so the query vector is similar to the document vector.

10.Then we created a query response pipeline, which takes a user input and returns a response.

11.Then we created a function for the user to initialize conversation and get response based on the query, the function also has an option to exit to terminate the conversation.

12.Then we created a function to test our pipeline, we created a list of five queries. We provided feedback to each of the responses to the five queries and based on the feedback, we created a customized template to fine tune the response. (reducing detailing in responses and creating more concise responses)

13.We imported the necessary libraries and fine tune few parameters such as temperature, max token, chunk size, chunk overlap, context window, similarily_top_k to get a less detailed and concise response

14.Finally, our Customized Prompt Template is ready to take query and provide accurate response.

Submitted by:

Nairit Dutta

Sanjay Fartyal

Vinita Yadav

DS 64