HelpMate AI Project

Building Effective Search Systems

Problem Statement

The goal of the project is to build a robust generative search system capable of effectively and accurately answering questions from a policy document.

Dataset

We will be using a single long life insurance policy document for this project. The PDF for the document can be downloaded below.

https://drive.google.com/file/d/1md-dHDxAIMHYCV0G4OqSK2thjLd02GkE/view?usp=drive_link

System Design

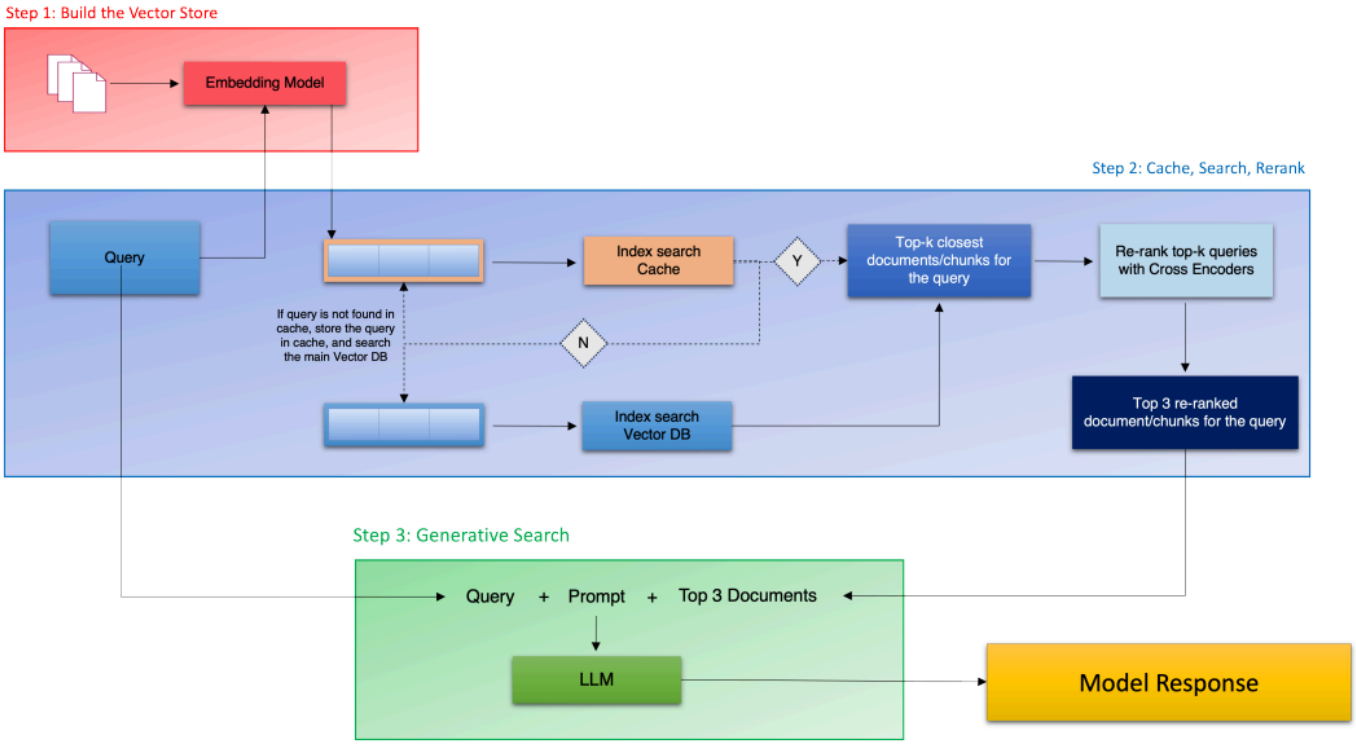
The project should implement all the three layers effectively. It will be key to try out various strategies and experiments in various layers in order to build an effective search system. Let's explore what we need to do in each of the layers.

1. **The Embedding Layer:** The PDF document needs to be effectively processed, cleaned, and chunked for the embeddings. Here, the choice of the chunking strategy will have a large impact on the final quality of the retrieved results. So, we need to make sure that we try out various stratgies and compare their performances.
Another important aspect in the embedding layer is the choice of the embedding model. we can choose to embed the chunks using the OpenAI embedding model or any model from the SentenceTransformers library on HuggingFace.

2. **The Search Layer:** Here, we need to design at least 3 queries against which you will test your system. You need to understand and skim through the document, and accordingly come up with some queries, the answers to which can be found in the policy document.
Next, we need to embed the queries and search your ChromaDB vector database against each of these queries. Implementing a cache mechanism is also mandatory.
Finally, we need to implement the re-ranking block, and for this we can choose from a range of cross-encoding models on HuggingFace.

3. **The Generation Layer:** In the generation layer, the final prompt that we design is the major component. We need to make sure that the prompt is exhaustive in its instructions, and the relevant information is correctly passedto the prompt. We may also choose to provide some few-shot examples inan attempt to improve the LLM output.

System Architecture



System Layers

• **Reading & Processing PDF File:** We will be using pdfplumber to read and process the PDF files. pdfplumber allows for better parsing of the PDF file as it can read various elements of the PDF apart from the plain text, such as, tables, images, etc. It also offers wide functionalities and visual debugging features to help with advanced preprocessing as well.
• **Document Chunking:** The document contains several pages and contains huge text, before generating the embeddings, we need to generate the chunks. Let's start with a basic chunking technique, and chunking the text with fixed size.
• **Generating Embeddings:** Generates embedding with SentenceTransformer with all-MiniLM-L6-v2 model.
• **Store Embeddings In ChromaDB:** In this section we will store embedding in ChromaDB.
• **Semantic Search with Cache:** In this section we will introduce cache collection layer for embeddings.
• **Re-Ranking with a Cross Encoder:** Re-ranking the results obtained from the semantic search will sometime significantly improve the relevance of the retrieved results. This is often done by passing the query paired with each of the retrieved responses into a cross-encoder to score the relevance of the response w.r.t. the query.
• **Retrieval Augmented Generation:** Now we have the final top search results, we can pass it to an GPT 3.5 along with the user query and a well-engineered prompt, to generate a direct answer to the query along with citations.

Queries and Results

```
[70] # First Query
     query1 = 'How many days of grace period is allowed'
     df1 = search(query1)
     df1 = apply_cross_encoder(query1, df1)
     df1 = get_topn(3, df1)
```

```
[71] # First query top 3 results
     print(df1)
```

```
                                          Documents                Metadatas
0  Section B - Premiums Article 1 - Payment Respo...  {'Page_No.': 'Page 20'}
1  Section C - Policy Termination Article 1 - Fai...  {'Page_No.': 'Page 23'}
2  Any individual policy issued will then be in f...  {'Page_No.': 'Page 43'}
```

```
[73] # Second Query
     query2 = 'what is the proof of ADL disability or total disability ?'
     df2 = search(query2)
     df2 = apply_cross_encoder(query2, df2)
     df2 = get_topn(3, df2)
```

```
[74] # Second query top 3 results
     print(df2)
```

```
                                          Documents                Metadatas
1  The Principal may require that a ADL Disabled ...  {'Page_No.': 'Page 50'}
0  Payment of benefits will be subject to the Ben...  {'Page_No.': 'Page 49'}
2  Coverage During Disability will cease on the e...  {'Page_No.': 'Page 51'}
```

```
] # Third Query
  query3 = 'When the proof of good health is not required ?'
  df3 = search(query3)
  df3 = apply_cross_encoder(query3, df3)
  df3 = get_topn(3, df3)
```

```
] # Third query top 3 results
  print(df3)
```

```
                                          Documents                Metadatas
0  Insurance for which Proof of Good Health is re...  {'Page_No.': 'Page 29'}
2  Scheduled Benefit in force for the Member befo...  {'Page_No.': 'Page 31'}
3  a. be actively engaged in business for profit ...  {'Page_No.': 'Page 17'}
```

```
[70] # First Query
     query1 = 'How many days of grace period is allowed'
     df1 = search(query1)
     df1 = apply_cross_encoder(query1, df1)
     df1 = get_topn(3, df1)
```

```
[71] # First query top 3 results
     print(df1)
```

```
                                          Documents                Metadatas
0  Section B - Premiums Article 1 - Payment Respo...  {'Page_No.': 'Page 20'}
1  Section C - Policy Termination Article 1 - Fai...  {'Page_No.': 'Page 23'}
2  Any individual policy issued will then be in f...  {'Page_No.': 'Page 43'}
```

```
[72] # First Query LLM Response
     response1 = generate_response(query1, df1)
     print("\n".join(response1))
```

```
The grace period allowed for premium payment is 30 days. This information is found in the 'Section B - Premiums Article 1 - Payment Responsibilities' document on Page
Citation:
- Policy Name: Section B - Premiums Article 1 - Payment Responsibilities
- Page Number: Page 20
```

```
     # Second Query
[73] query2 = 'what is the proof of ADL disability or total disability ?'
     df2 = search(query2)
     df2 = apply_cross_encoder(query2, df2)
     df2 = get_topn(3, df2)
```

```
[74] # Second query top 3 results
     print(df2)
```

```
                                          Documents                Metadatas
1  The Principal may require that a ADL Disabled ...  {'Page_No.': 'Page 50'}
0  Payment of benefits will be subject to the Ben...  {'Page_No.': 'Page 49'}
2  Coverage During Disability will cease on the e...  {'Page_No.': 'Page 51'}
```

```
[75] # Second Query LLM Response
     response2 = generate_response(query2, df2)
     print("\n".join(response2))
```

```
The proof of ADL (Activities of Daily Living) disability or total disability is typically demonstrated through a combination of medical records, functional assessment

In case you need more detailed guidelines or specific forms for submitting the proof of disability, it is recommended to review the policy document mentioned above fo

Here are the relevant details for the citation:
- Policy Name: Section B - Premiums Article 1 - Payment Responsibility
- Page Number: Page 20

In case you need additional information or have further questions, feel free to reach out to your insurance provider for personalized guidance.
```

```
# Third Query
query3 = 'When the proof of good health is not required ?'
df3 = search(query3)
df3 = apply_cross_encoder(query3, df3)
df3 = get_topn(3, df3)
```

```
# Third query top 3 results
print(df3)
```

Show hidden output

```
# Third Query LLM Response
response3 = generate_response(query3, df3)
print("\n".join(response3))
```

```
Proof of good health is typically not required when applying for a guaranteed issue life insurance policy. These types of policies do not require any medical exams or

Unfortunately, the provided documents do not seem to contain specific information regarding when proof of good health is not required. It's recommended to review the

Here is the relevant information from the data frame:

| Policy Name                                    | Page Number |
|------------------------------------------------|-------------|
| Section B - Premiums                           | Page 20     |
| Section C - Policy Termination                 | Page 23     |
| Individual Policy Issuance Details             | Page 43     |

I recommend reviewing the sections mentioned above in the respective policy documents to find detailed information about the requirements for proof of good health.
```