Symantic Spotter using LlamaIndex

Author: Rakesh Rangadal

Date: 01-10-2025

Project Objective

Build a project in the insurance domain, similar to the project you saw in the 'Retrieval Augmented Generation' session. The goal of the project will be to build a robust generative search system capable of effectively and accurately answering questions from various policy documents. You may use LangChain or LlamaIndex to build the generative search application

Solution Strategy

Build a solution which should solve the following requirements using LlamaIndex:

- Users would responses from insurance policy knowledge base.
- If user want to perform a query system must be able to response to query accurately.

Goal

Solving the above two requirements well in and would ensure that the accuracy of the overall model is good.

Data Used

HDFC various Insuracne policy documetns sotred in single folder

Tools used LlamaIndex, GhatGPT has been used due to its powerful query engine, fast data processing using data loaders and directory readers as well as easier and faster implementation using fewer lines of code.

Tools used

LlamaIndex, GhatGPT, disc cache has been used due to its powerful query engine, fast data processing using data loaders and directory readers as well as easier and faster implementation using fewer lines of code.

Why LlamaIndex ?

LlamaIndex is an innovative data framework specially designed to support LLM-based RAG framework application development. It offers an advanced framework that empowers developers to integrate diverse data sources with large language models.

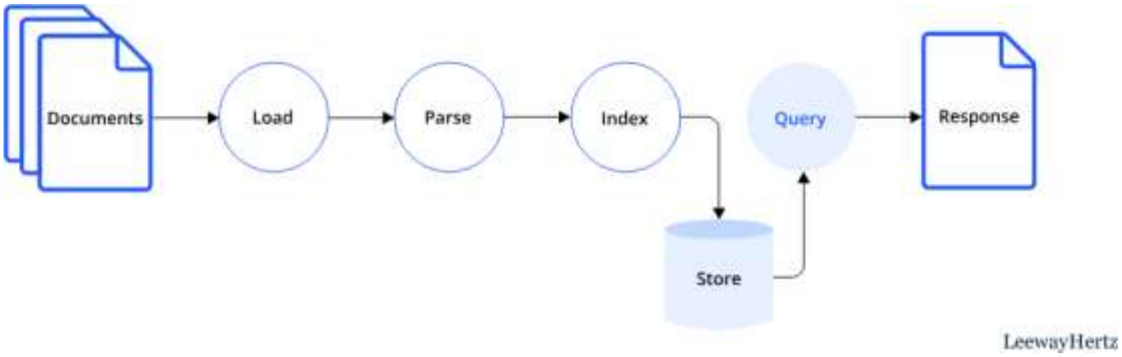
LlamaIndex includes a variety of file formats, such as PDFs and PowerPoints, as well as applications like Notion and Slack and even databases like Postgres and MongoDB.

The framework brings an array of connectors that assist in data ingestion, facilitating a seamless interaction with LLMs. Moreover, LlamaIndex boasts an efficient data retrieval and query interface.

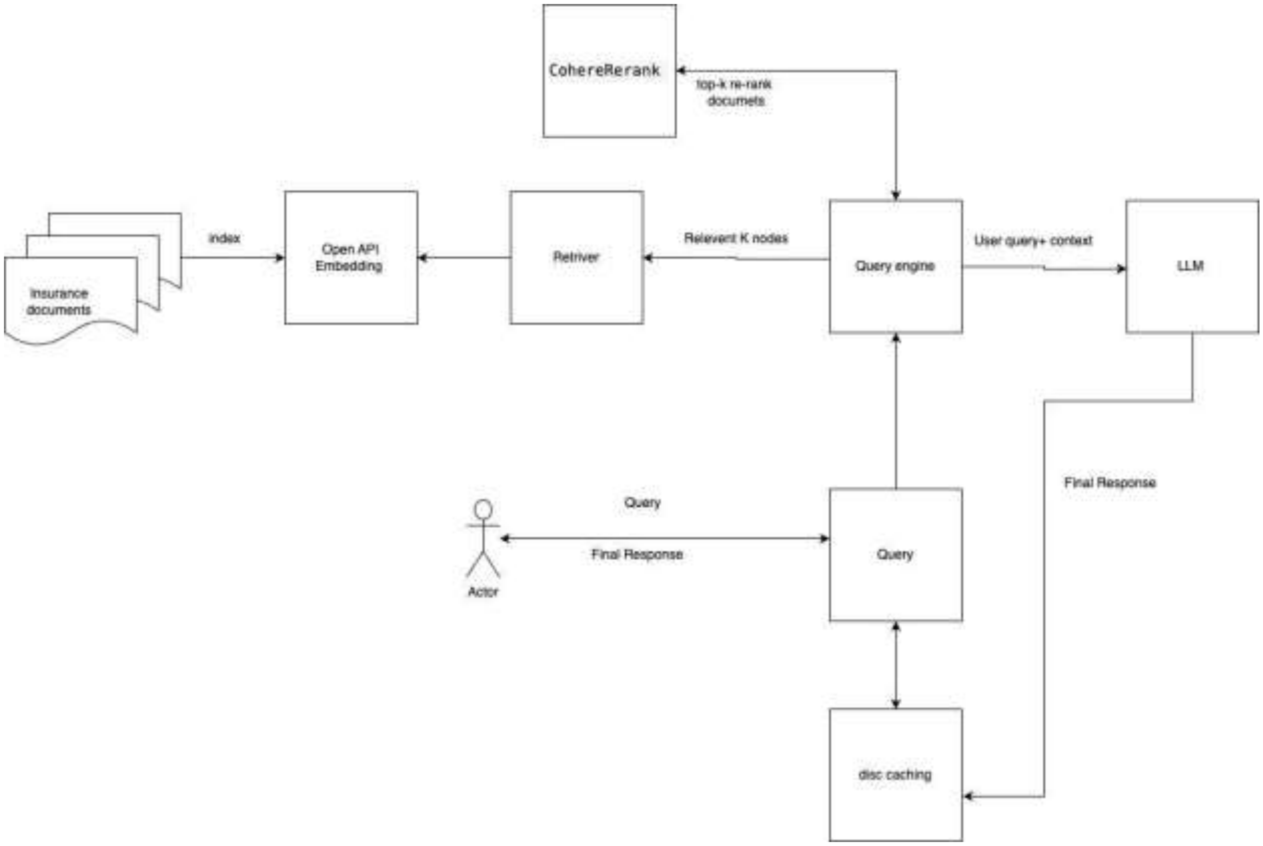
LlamaIndex enables developers to input any LLM prompt and, in return, receive an output that is both context-rich and knowledge-augmementation.

Key Feature of LlamaIndex:

- Data connectors allow ingestion from various data sources and formats.
- It can synthesize data from multiple documents or heterogeneous data sources.
- It provides numerous integrations with vector stores, ChatGPT plugins, tracing tools, LangChain, and more.



High Level Design of Semantic Spotter :



Architecture Descriptions:

1. ***Documents***: We will be using list of HDFC insurance documents provides inside a single folder.
2. ***Open API embedding***: We ars using OpenAPI embedding as Vector DB for indexing insureance documetns in the form of embedding.
3. ***Query Engine***: We are using Query Engine Module of Llammaindex for perfomring synantic Search. Query Engine will use internally Retriver and Cohere Rerank to retrrive top-k relvent nodes from embedding.
4. ***LLM*** : top k-documets along with usewr query will be passed to LLM to generate the accurate response.We are using chatGPT LLM.
5. ***Caching*:**" Caching is being used to improve the read operation. Recent similar search will be store in Caching and user query frist will be served from Cahcing. If user query not found in cahce then query will be forwared to query engine and then LLM to geneate

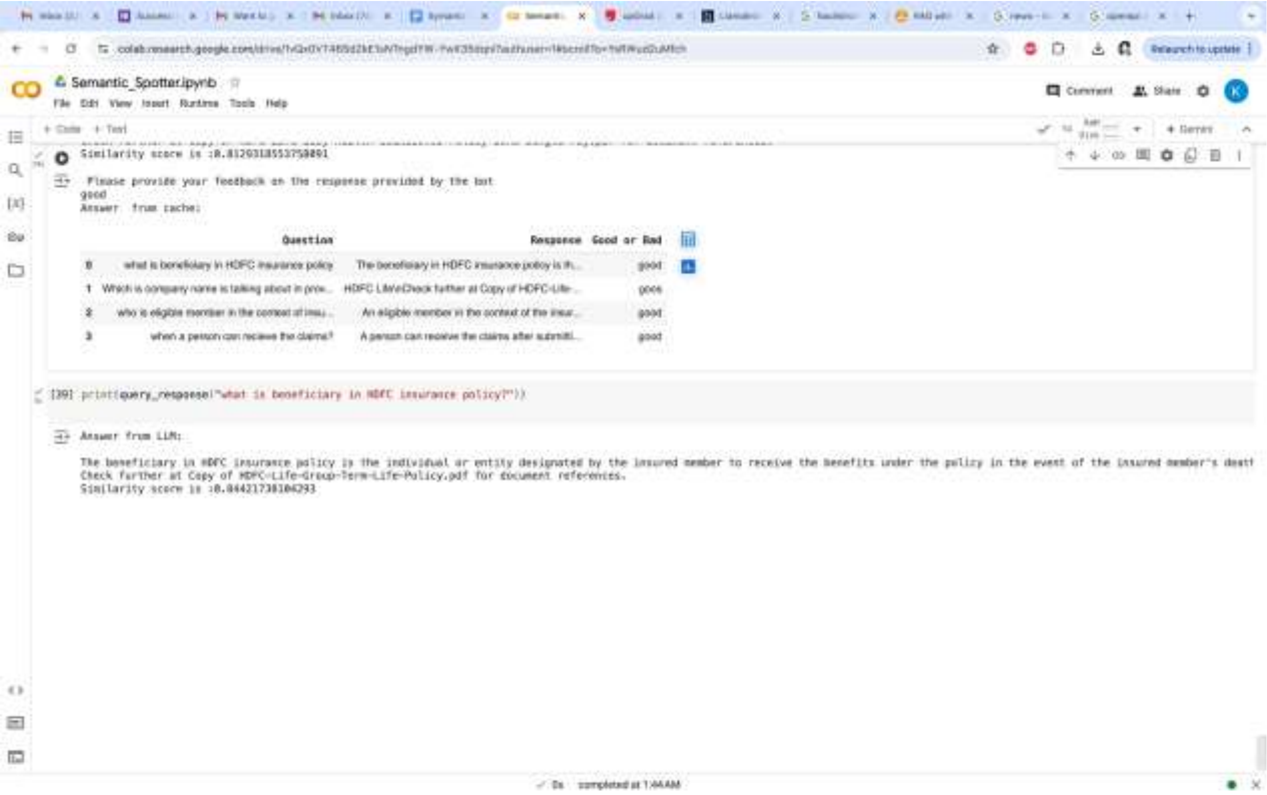
the resposne. user query and generated resposne will be cached in in cache and will be served from there based ttl.

6. ***Meta data***:Along with Resposne we are also returning docs referece and similary score to improve the user confidence towards the implemented RAG system.
7. ***Cohere-Rerank***: Is being used to rerank the query based on semantic score.

Generative Search Response from Insurance documents:

We have attached custom query generative search results.

1. Using a single Query Respone:

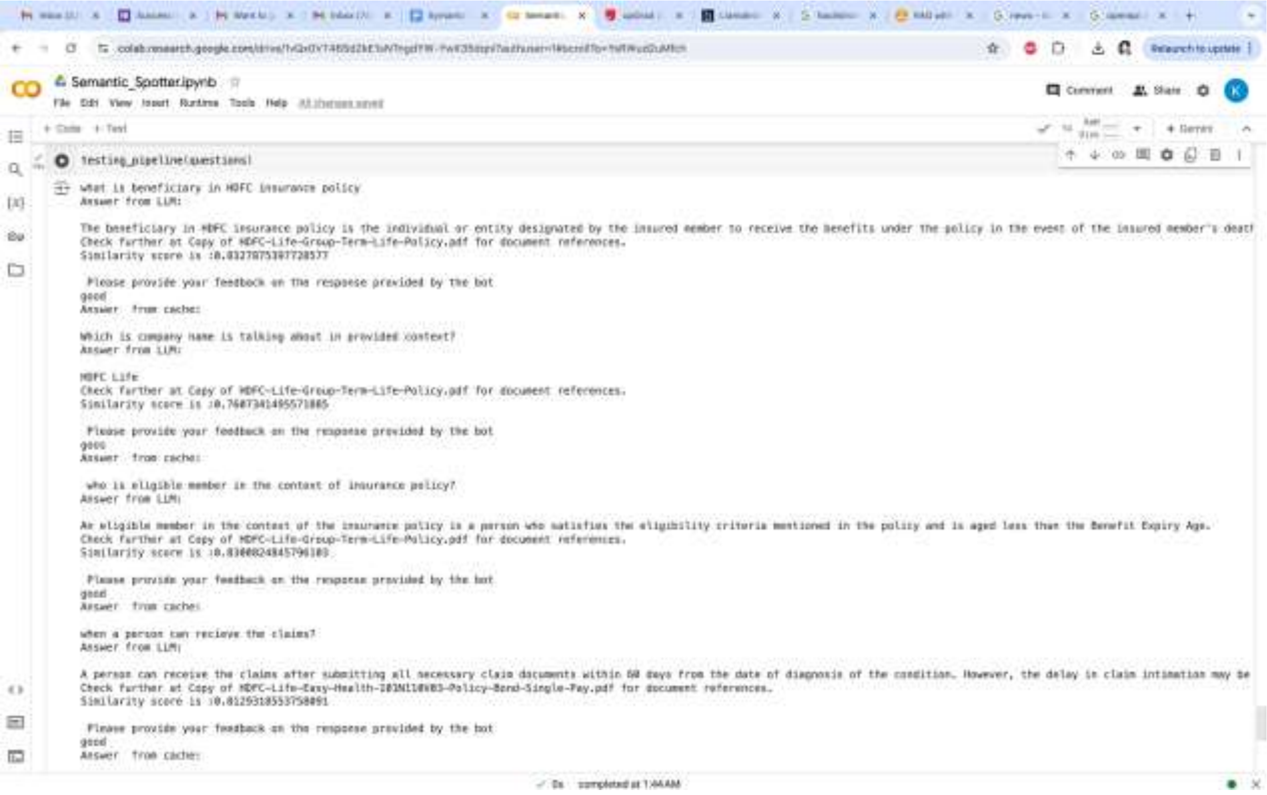


```
Similarity score is :0.8129318553758091

 Please provide your feedback on the response provided by the bot
good
Answer  from cache:

                                    Question                                        Response  Good or Bad
0        what is beneficiary in HDFC insurance policy   The beneficiary in HDFC insurance policy is th...        good
1  Which is company name is talking about in prov...   HDFC Life\nCheck further at Copy of HDFC-Life-...        good
2      who is eligible member in the context of insu...   An eligible member in the context of the insur...        good
3             when a person can recieve the claims?    A person can receive the claims after submitti...        good
```

```
[39] print(query_response("what is beneficiary in HDFC insurance policy?"))
```

```
Answer from LLM:

The beneficiary in HDFC insurance policy is the individual or entity designated by the insured member to receive the benefits under the policy in the event of the insured member's deatl
Check further at Copy of HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.8442173810429З
```

Multiple Query Resposne:



```
testing_pipeline(questions)
```

```
what is beneficiary in HDFC insurance policy
Answer from LLM:

The beneficiary in HDFC insurance policy is the individual or entity designated by the insured member to receive the benefits under the policy in the event of the insured member's deat
Check further at Copy of HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.8327875387728577

 Please provide your feedback on the response provided by the bot
good
Answer  from cache:

Which is company name is talking about in provided context?
Answer from LLM:

HDFC Life
Check further at Copy of HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.7607341495571885

 Please provide your feedback on the response provided by the bot
goos
Answer  from cache:

 who is eligible member in the context of insurance policy?
Answer from LLM:

An eligible member in the context of the insurance policy is a person who satisfies the eligibility criteria mentioned in the policy and is aged less than the Benefit Expiry Age.
Check further at Copy of HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.8300824845796103

 Please provide your feedback on the response provided by the bot
good
Answer  from cache:

when a person can recieve the claims?
Answer from LLM:

A person can receive the claims after submitting all necessary claim documents within 60 days from the date of diagnosis of the condition. However, the delay in claim intimation may be
Check further at Copy of HDFC-Life-Easy-Health-101N110V03-Policy-Bond-Single-Pay.pdf for document references.
Similarity score is :0.8129318553758091

 Please provide your feedback on the response provided by the bot
good
Answer  from cache:
```

Code:
Jupyter Notebook developed for Semantic Spotter.

Chalenges Faced:

We tried to use GPTCache with for caching system,but due compatiability issues, we couldn't interate it.

Alternative Solution:

We are usign disc caching for alternative to GPTCache.

Future of work

1. We can further improve the solution by using feedback mechanism usingc embedding the response along with feedback to vector database.
2. We can also integrate RAGAS/DeepEval Framework to improve the symnatic search and generated response.