

Semantic Spotter using LlamaIndex

1. Project Objective:

The objective of this project is to build a **generative search system** in the insurance domain, similar to the project from the "Retrieval-Augmented Generation" session. The system will effectively and accurately answer user queries from various insurance policy documents. This solution is built using **LlamaIndex** and **OpenAI's GPT-3.5-turbo**.

2. Overview

This project builds a robust **Q&A system** for insurance policy queries using **LlamaIndex**, **OpenAI's GPT-3.5-turbo**. The system applies a **Retrieval-Augmented Generation (RAG)** approach to retrieve relevant sections from insurance documents and generate accurate, context-aware answers. It also incorporates caching to optimize performance for frequently asked questions (FAQs).

Solution Strategy

The project aims to solve two primary requirements using **LlamaIndex**:

- Provide users with accurate responses from an insurance policy knowledge base.
- Ensure the system accurately responds to user queries by retrieving relevant sections from the documents.

Achieving these goals will ensure the system's overall accuracy and usefulness.

Data Used

The data used includes **HDFC insurance policy documents**, all stored in a single folder. These documents form the knowledge base for the system.

Tools Used

- **LlamaIndex**: Used as the main framework for document retrieval and search.
- **OpenAI GPT-3.5-turbo**: Utilized for generating accurate and context-rich responses.
- **Diskcache**: Implemented to store and retrieve frequently asked questions and their responses.

The reason for using **LlamaIndex** is due to its powerful query engine, fast data processing with data loaders and directory readers, and the ability to build the system with fewer lines of code.

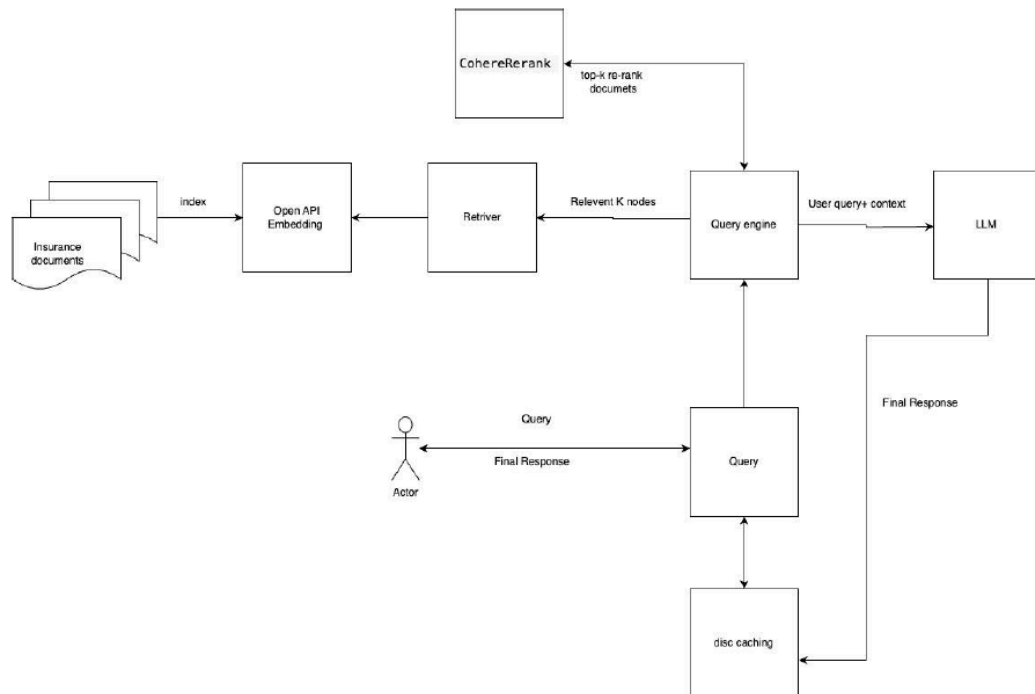
Why LlamaIndex?

LlamaIndex is an innovative data framework designed to support the development of **LLM-based Retrieval-Augmented Generation (RAG) systems**. It provides efficient data ingestion, retrieval, and query processing capabilities with large language models.

Key Features of LlamaIndex:

- **Data Connectors**: It supports a wide range of data formats and sources, including PDFs, PowerPoints, databases, and more.
- **Data Synthesis**: Combines information from multiple documents and heterogeneous data sources.
- **Integrations**: Integrates with vector stores, ChatGPT plugins, LangChain, and tracing tools.
- **Query Interface**: Provides a seamless interaction for querying large datasets and generating context-rich, knowledge-augmented outputs.

3. Design:



Architecture Overview:

1. **Documents:** A collection of **HDFC insurance documents** stored in a single folder is used as the knowledge base.
2. **OpenAI Embeddings:** **OpenAI embeddings** are used as a **Vector DB** for indexing the insurance documents by embedding their content.
3. **Query Engine:**
The **Query Engine** from LlamaIndex is used for **semantic search**. It retrieves top-K relevant document chunks based on the user query, and these are passed to the LLM for response generation.
4. **LLM Integration:**
The **top-K retrieved documents** along with the user query are passed to the **ChatGPT** model (GPT-3.5-turbo) to generate a precise and context-aware response.
5. **Caching Layer:**
Diskcache is implemented to improve system efficiency. When a similar query is made, the system first checks the cache. If the query isn't found in the cache, it's forwarded to the query engine and LLM, after which the generated response is cached for future use.

4. Implementation Steps:

Step 1: Data Loading and Preprocessing

- **File Formats:** The system handles document loading using *SimpleDirectoryReader* which processes and reads text from documents.
- **Document Parsing:** The documents are parsed, and large documents are split into chunks for efficient retrieval.

Step 2: Indexing and Query Engine Setup

- **LlamaIndex:**
 - Documents are indexed using **VectorStoreIndex** for fast retrieval of relevant sections.

Step 3: LLM Integration

- **OpenAI GPT-3.5-turbo:**
 - The retrieved documents are passed to **GPT-3.5-turbo**, which generates accurate answers based on a **custom prompt** that ensures context and precision.

Step 4: Caching

- **Diskcache:**
 - Implemented to store responses to frequently asked questions. This reduces retrieval time and enhances the efficiency of the system by caching similar queries.

Step 5: Interactive Testing Pipeline

- **User Interaction:**
 - The system allows users to interact and query the insurance policy database. Feedback on the accuracy of the responses is collected for further refinement of the model.

Step 6 :Metadata Handling Along with the response, the system returns metadata such as the document reference and similarity score to improve user confidence in the generated results.

Generative Search Response from Insurance documents :

We have attached custom query generative search results.

1. Using a single Query Response:

```
1 print(query_response("What are the conditions under which the insurance
2 coverage for a Scheme Member will terminate?"))
```

Answer from cache:

The insurance coverage for a Scheme Member will terminate under the following conditions:


- Master Policy being terminated
- End of Coverage Term
- Surrender of Certificate of Insurance
- Free Look Cancellation
- Payment of Plan Benefit
- Refund of premium under Suicide Clause

Check further at [HDFC-Life-Group-Poorna-Suraksha-101N137V02-Policy-Document.pdf](#) for document references. Similarity score is :0.8503726392043668

2. Multiple Query Response:

1 to 4 of 4 entries Filter ?			
index	Question	Response	Good or Bad
0	What are the conditions under which the insurance coverage for a Scheme Member will terminate?	The insurance coverage for a Scheme Member will terminate under the following conditions: - Master Policy being terminated - End of Coverage Term - Surrender of Certificate of Insurance - Free Look Cancellation - Payment of Plan Benefit - Refund of premium under Suicide Clause Check further at HDFC-Life-Group-Poorna-Suraksha-101N137V02-Policy-Document.pdf for document references. Similarity score is :0.8503726392043668	Good
1	what is beneficiary in HDFC insurance policy?	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 death. If no designated nominee is filed or if the nominee predeceases the insured member, the benefits will be payable to the legal heir of the insured member. The insured member can change the nominee during their lifetime by providing written notice to the policyholder. Check further at HDFC-Life-Group-Term-Life-Policy.pdf for document references. Similarity score is :0.8442442153031304	Good
2	What are Accidental Death Benefits	Accidental Death Benefits include the payment of a benefit if the death of a Scheme Member occurs within 180 days from the date of an accident. However, there are specific exclusions for this benefit, such as instances like intentionally self-inflicted injury, engaging in hazardous pursuits, involvement in criminal acts, or participating in war or civil unrest. Check further at HDFC-Life-Group-Poorna-Suraksha-101N137V02-Policy-Document.pdf for document references. Similarity score is :0.8170240007215589	Good
3	who is eligible member in the context of insurance policy?	An eligible member in the context of the insurance policy is a person who satisfies the eligibility criteria mentioned in the policy document and is aged less than the Benefit Expiry Age. Check further at HDFC-Life-Group-Term-Life-Policy.pdf for document references. Similarity score is :0.8316478216990661	Good

Show 25 per page



TESTING PIPELINE QUESTIONS AND FEEDBACK :

```
1 testing_pipeline(questions)

What are the conditions under which the insurance coverage for a Scheme Member will terminate?
Answer from cache:

The insurance coverage for a Scheme Member will terminate under the following conditions:
- Master Policy being terminated
- End of Coverage Term
- Surrender of Certificate of Insurance
- Free Look Cancellation
- Payment of Plan Benefit
- Refund of premium under Suicide Clause
Check further at HDFC-Life-Group-Poorna-Suraksha-101M137V02-Policy-Documents.pdf for document references.
Similarity score is :0.8503726392043668

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

what is beneficiary in HDFC insurance policy?
Answer from cache:

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 death. If no designated nominee is filed
Check further at HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.8442442153031304

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

What are Accidental Death Benefits
Answer from cache:

Accidental Death Benefits include the payment of a benefit if the death of a Scheme Member occurs within 180 days from the date of an accident. However, there are specific exclusions for this benefit, such as instances 1:
Check further at HDFC-Life-Group-Poorna-Suraksha-101M137V02-Policy-Documents.pdf for document references.
Similarity score is :0.8170240007215589

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

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

An eligible member in the context of the insurance policy is a person who satisfies the eligibility criteria mentioned in the policy document and is aged less than the Benefit Expiry Age.
Check further at HDFC-Life-Group-Term-Life-Policy.pdf for document references.
Similarity score is :0.8316478216990661

Please provide your feedback on the response provided by the bot
Good
```

The **testing pipeline** plays a crucial role in ensuring the accuracy and efficiency of the **Semantic Spotter** system. It allows users to interact with the system by posing queries related to insurance policies and ensures that the generated responses are correct, contextually relevant, and timely. The testing pipeline includes mechanisms for user feedback and provides a platform to evaluate the performance of the Retrieval-Augmented Generation (RAG) system in real-world scenarios.

5. Challenges Faced:

1. **GPTCache Compatibility Issues:** Due to compatibility issues with **GPTCache**, an alternative caching solution using **Diskcache** was implemented.

6. Lessons Learned

1. **Preprocessing is Key:** Proper preprocessing of documents is critical for accurate document indexing and retrieval.
2. **Custom Prompts Enhance Accuracy:** Tailoring the prompt for GPT-3.5-turbo helped in generating more accurate and relevant answers.
3. **Caching is Essential for Performance:** Efficient cache management greatly enhances response times for frequently asked queries.
4. **Document Metadata:** Providing additional metadata such as document references and similarity scores boosts user trust in the system.

THE END