Project Report

Semantic Spotter - RAG Insurance Assistant

Upgrad-IIIT Bangalore

Compiled By: **Dipak Sah**

Cohort 60 AI/ML

Date:2024-12-25

1. Project Goal

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. We have to choose LangChain or LlamaIndex to build the generative search application.

Easily find the information you need in your insurance documents with our simplified query system, built using RAG, Llama-Index, and OpenAI's GPT models

2. About the Assignment

Deciphering insurance documents can be a frustrating and time-consuming process. The RAG Insurance Assistant offers a solution by using **Retrieval-Augmented Generation (RAG)** powered by **LlamaIndex**. This technology efficiently retrieves relevant information from policy documents, claim guidelines, and legal texts, then uses advanced AI models like GPT-4 or Gemini to provide clear, concise answers to user queries.

2.1 Key Benefits of the project

Here are some the key benefits of executing this project

- **Streamlined Information Access**: Forget endless searches—ask specific questions and receive precise, concise answers instantly.
- **Enhanced Contextual Understanding**: Breaks down complex legal language into user-friendly explanations.
- **Powerful Scalability**: Handles large datasets effortlessly, making it suitable for both personal and enterprise-level applications.

2.2 Use Case Scope

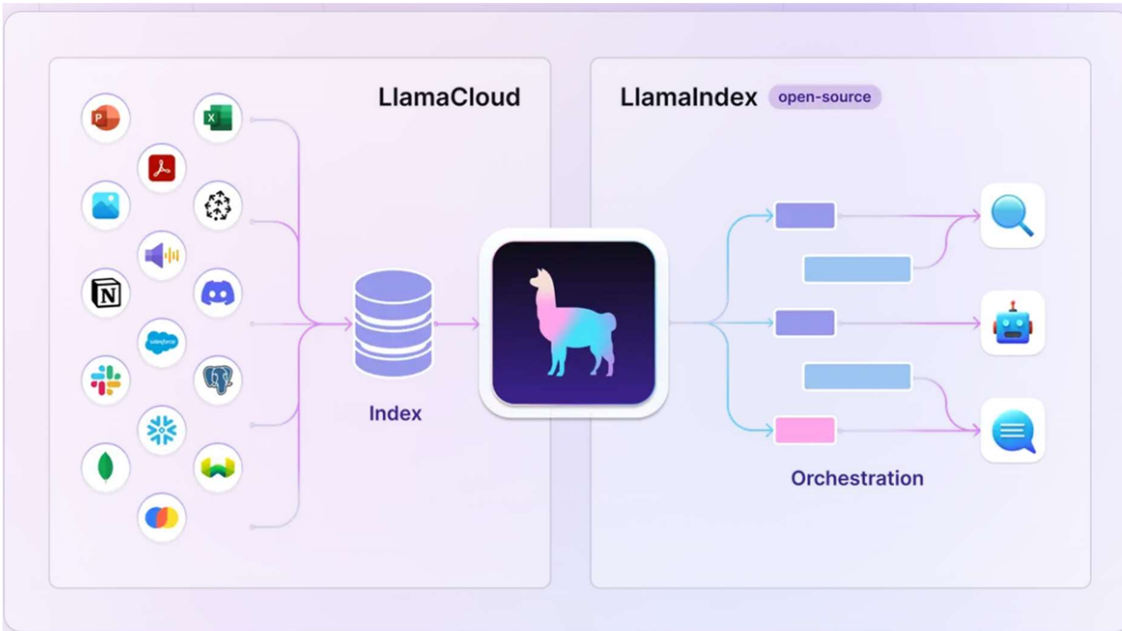
Whether you're a policyholder seeking clarity on coverage or an insurance agent streamlining customer service, the RAG Insurance Assistant transforms the way users interact with complex insurance documents.

Example Use Cases:

- "What are the benefits of HDFC Sampoorna-Jeevan insurance?"
- "What will happen if premium not paid on time for HDFC Surgicare Plan ?"
- "Is there any age limit for HDFC Life Smart pension Plan ?"
- "Is blindness covered in HDFC Life Easy Health Policy?"
- "What is the timeline to cancel HDFC Life Sanchay Plus Life Long Income Option plan, in case of any dissatisfaction?"

3. About LlamaIndex

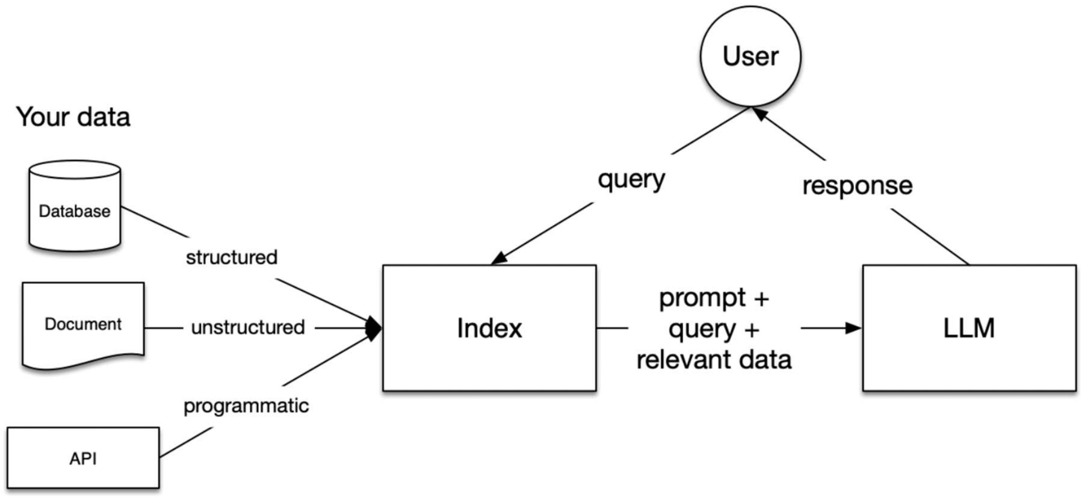
LlamaIndex is a framework for building context-augmented LLM applications. Context augmentation refers to any use case that applies LLMs on top of your private or domain-specific data.



Some popular use cases include the following:

- Question-Answering Chatbots (commonly referred to as RAG, or "Retrieval-Augmented Generation")
- Document Understanding and Extraction
- Autonomous Agents that can perform research and take actions

LlamaIndex provides the tools to build any of these above use cases from prototype to production. The tools allow you to ingest and process this data and implement complex queries. It combines data access with LLM prompting as shown in the below flow diagram.



3.1 Key Features of LlamaIndex

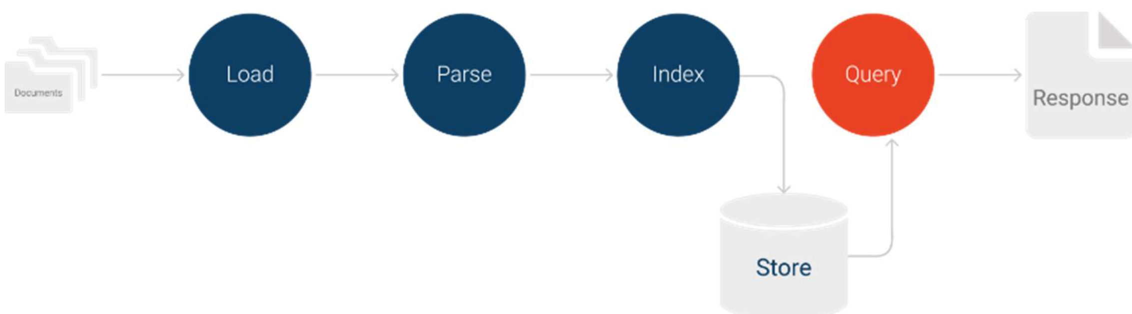
- **Fast, Relevant Information Retrieval**: LlamaIndex ensures quick access to the most relevant sections of insurance documents, providing precise answers tailored to user queries.
- **Contextually Accurate Answers**: Advanced retrievers combined with AI language models (GPT-4 or Gemini) generate natural-language responses that understand the context of complex insurance policies.
- **Scalable Performance with Vector Stores**: It enables fast and scalable performance, even with large datasets, by efficiently storing and querying embeddings.
- **Document Agnostic**: Processes various document types, including PDFs, Word files, and text files, making it adaptable to any insurance material.
- **Flexible Chunking Strategies**: Employs customizable document chunking methods, including overlapping techniques, to optimize retrieval accuracy and avoid information gaps.
- **Cloud or Local Deployment**: Fully adaptable to your infrastructure needs—deploy locally for personal use or on the cloud for enterprise applications.
- **Secure Access**: Ensures data privacy by handling sensitive insurance data with secure API key management and controlled access.
- **Analytics-Ready**: Optionally integrates with analytics tools to track query patterns and optimize document updates or user experience.
- **Easy-to-Use Conversational Interface**: Provides a simple, conversational way to ask policy questions, eliminating the need to read through lengthy documents.
- **Adaptable to Future Needs**: Designed for extensibility, easily supporting additional domains beyond insurance, such as legal or healthcare.

3.2 Why Llama-Index for this project?

LlamaIndex excels for PDF-heavy RAG due to its specialized data connectors optimized for diverse document formats (including PDFs), robust data indexing and structuring capabilities (crucial for efficient PDF retrieval), simpler abstraction for building document-centric applications, and strong focus on data management and querying which is often more complex in PDF datasets than simpler text sources. While LangChain is more general-purpose, LlamaIndex is purpose-built for handling complex document structures, making it often more effective for PDF-based RAG. Hence, for our project which is having PDF as information base, Llama-Index is excellent choice.

4. System Design

Below system design is used in the solution.



- **Data Loading**: SimpleDirectoryReader loads documents (including PDFs) from a directory. No explicit PyPDFLoader is needed, as Llama Index handles it internally.
- **Parsing**: SimpleNodeParser creates chunks or nodes for given documents. Here, chunk size = 1024, chunk overlap = 128 are used in this solution development.
- **Indexing**: VectorStoreIndex creates the vector index. The service context is initialized with the LLM and embedding model.
- **Storage**: StorageContext manages index persistence. The example shows how to load and save the index to disk, effectively providing caching.
- **Querying**: RetrieverQueryEngine() creates a query engine.
- **Citations**: Llama Index directly provides source nodes with metadata (filename, page number if available). The code formats the response to include these citations.
- **LLM and Embeddings**: The code explicitly sets the LLM (gpt-3.5-turbo) and embedding model (HuggingFaceEmbedding(model_name="BAAI/bge-small-en-v1.5")).
- **Service Context**: The ServiceContext bundles the LLM and embedding model for use throughout the Llama Index pipeline.

- **Caching**: The create_or_load_index function handles loading the index from disk if it exists, implementing the caching mechanism and improving the performance.

4.1 Tech Stack Used

- **Language**: Python-In Jupyter Notebook
- **Frameworks/Libraries**: Transformers, HuggingFace, Llama-Index, Disk Cache
- **APIs/Models**: OpenAI's GPT-3.5
- **Tools used**: Jupyter Notebook(Collab)

4.2 Data Source

Below are the pdf files upload in the GitHub, used as data sources for this project. [Semantic-Spotter-Using-Llama-Index/insurance_docs at main · dipaksah20/Semantic-Spotter-Using-Llama-Index]()

Semantic-Spotter-Using-Llama-Index / insurance_docs / Add file ···

dipaksah20 Delete insurance_docs/ss d0d50e9 · 2 days ago History

Name	Last commit message	Last commit date
..		
HDFC-Life-Easy-Health-101N110V03-Policy-Bond-Single-Pay.pdf	Add files via upload	2 days ago
HDFC-Life-Group-Poorna-Suraksha-101N137V02-Policy-Docum...	Add files via upload	2 days ago
HDFC-Life-Group-Term-Life-Policy.pdf	Add files via upload	2 days ago
HDFC-Life-Sampoorna-Jeevan-101N158V04-Policy-Document (...	Add files via upload	2 days ago
HDFC-Life-Sanchay-Plus-Life-Long-Income-Option-101N134V1...	Add files via upload	2 days ago
HDFC-Life-Smart-Pension-Plan-Policy-Document-Online.pdf	Add files via upload	2 days ago
HDFC-Surgicare-Plan-101N043V01.pdf	Add files via upload	2 days ago

5. Getting Started

5.1 Prerequisites

Ensure you have the following installed:

- Python Latest version
- Llama-Index libraries as mentioned in ipynb file
- HuggingFaceEmbedding(model_name="BAAI/bge-small-en-v1.5")

5.2 Installation & Execution

1. Clone the repo: git clone https://github.com/dipaksah20/Semantic-Spotter-Using-Llama-Index/tree/main
2. Run the Jupyter: "Semantic_Spotter_LlamaIndex_V6.ipynb"

3. Ensure the input files are copied from insurance_docs folder from GitHub.

   Please note: OpenAI API keys are required for the project to function. You can obtain them from the OpenAI website and change the same in the code. We have updated the code and added more models to make it more dynamic.

6. Documentation

Please refer to the following links for more information:

- OpenAI
- Llama-Index

7. Challenges/Issues Faced with fixes

- Cache layer was added to improve the performance and efficiency of systems by storing frequently accessed data in a readily available location. This was done to make the retrieval process more efficient.
- Vector index generation was stored in the local storage to avoid regeneration and stored vector index is used every time we ran the solution.

8. Lesson Learned During Project

Key lessons learned from developing a PDF-based RAG system for insurance documents using LlamaIndex:

- **Chunking is critical**: Experiment with different chunking strategies (size, overlap, methods) to optimize retrieval accuracy, as fixed sizes may not handle complex PDF layouts well.
- **Embedding model choice matters**: Evaluate different embedding models, possibly domain-specific ones, as embedding quality directly impacts retrieval relevance.
- **Caching needs management**: While disk caching improves performance, implement a mechanism to invalidate and rebuild the cache when source documents change.
- **Metadata is crucial for citations**: Ensure robust metadata extraction (especially page numbers from PDFs) for transparent and trustworthy responses.
- **LLM and prompts impact quality**: Consider more powerful LLMs or fine-tuning, and use careful prompt engineering to guide the LLM towards accurate and relevant answers.

9. Search response from solution

We used our 5 use case mentioned in the use case scope. Below is the response in blue from the search system on insurance documents.

```
[ ] # 1st search
    initialize_conv()
```

```
Feel free to ask Questions regarding Uber financial document. Press exit once you are done
what are the benefits of HDFC Sampoorna-Jeevan insurance?
```

The benefits of HDFC Sampoorna Jeevan insurance include participation in the surplus profits of the company, guaranteed benefits such as Sum Assured on Death, Survival Benefits, and bonuses like Simple Reversionary Bonus, Simple Reversionary Income Bonus, and Terminal Bonus. Additionally, the policy offers options for Guaranteed Income Benefit, various Bonus Options, and Special Surrender Value. Check further at HDFC-Life-Sampoorna-Jeevan-101N158V04-Policy-Document (1).pdf document page no 18

```
exit
Exiting the search... bye bye
```

```
[ ] # 2nd search
    initialize_conv()
```

```
Feel free to ask Questions regarding Uber financial document. Press exit once you are done
What will happen if premium not paid on time for HDFC Surgicare Plan ?
```

If the premium for the HDFC SurgiCare Plan is not paid on time after the grace period of fifteen days following the Due Date, the policy may lapse with effect from the Due Date of the first unpaid premium. Check further at HDFC-Surgicare-Plan-101N043V01.pdf document page no 12

```
exit
Exiting the search... bye bye
```

```
] # 3rd search
  initialize_conv()
```

```
Feel free to ask Questions regarding Uber financial document. Press exit once you are done
Is there any age limit for HDFC Life Smart pension Plan ?
```

Yes, there are age limits specified for the HDFC Life Smart Pension Plan. The minimum and maximum entry ages are considered for the Life Assured at the Policy Commencement Date. If the Age of the Life Assured is found to be higher than the maximum or lower than the minimum entry Age permissible under the Policy at the time of issue, adjustments may be made to the Premiums and Benefits under the Policy. Check further at HDFC-Life-Smart-Pension-Plan-Policy-Document-Online.pdf document page no 21

```
exit
Exiting the search... bye bye
```

```
] # 4th search
  initialize_conv()
```

```
Feel free to ask Questions regarding Uber financial document. Press exit once you are done
Is blindness covered in HDFC Life Easy Health Policy ?
```

Blindness is covered in the HDFC Life Easy Health Policy. Check further at HDFC-Life-Easy-Health-101N110V03-Policy-Bond-Single-Pay.pdf document page no 7

```
exit
Exiting the search... bye bye
```

```
# 5th search
initialize_conv()
```

```
Feel free to ask Questions regarding Uber financial document. Press exit once you are done
What is the timeline to cancel HDFC Life Sanchay Plus Life Long Income Option plan, in case of any dissatisfaction ?
```

The timeline to cancel the HDFC Life Sanchay Plus Life Long Income Option plan in case of any dissatisfaction is within 15 days from the date of receipt of the policy. Check further at HDFC-Life-Sanchay-Plus-Life-Long-Income-Option-101N134V19-Policy-Document.pdf document page no 1

```
exit
Exiting the search... bye bye
```

10. Future Improvements

- Add more selectable GPT models to the project(Gemini, Claude AI, Huggingface models etc).
- Add more features to the project.
- Add more selectable Vector Store to the project(Pinecone, Weaviate etc).

11. Contact

For any queries or feedback, feel free to reach out:

- **Email**: dipaksah@gmail.com
- **GitHub**: https://github.com/dipaksah20