Syed Abdul Rahim

sarpsl@gmail.com

Semantic Spotter 2.0

Project Report

Contents

[Project Goals 2](#_Toc172668175)

[Data Sources 2](#_Toc172668176)

[Document Source: 2](#_Toc172668177)

[Design Choices 2](#_Toc172668178)

[Framework Choice: 2](#_Toc172668179)

[Document Parsing: 2](#_Toc172668180)

[Vector Storage: 2](#_Toc172668181)

[Query Processing: 2](#_Toc172668182)

[Integration with LLM: 3](#_Toc172668183)

[Challenges Faced 3](#_Toc172668184)

[Complexity of PDF Documents: 3](#_Toc172668185)

[Indexing Efficiency: 3](#_Toc172668186)

[Query Accuracy: 3](#_Toc172668187)

[Flowchart 3](#_Toc172668188)

[Setup 4](#_Toc172668189)

[Usage 4](#_Toc172668190)

[Import Libraries: 4](#_Toc172668191)

[Run the Script: 4](#_Toc172668192)

[Query the System: 4](#_Toc172668193)

[Troubleshooting 4](#_Toc172668194)

[PDF Parsing Issues: 4](#_Toc172668195)

[Indexing Errors: 4](#_Toc172668196)

## Project Goals

This project aims to create a document processing and query system using LlamaIndex with an HDFC Policy PDF as the source document. The goal is to build a system that can parse complex documents, extract relevant information, and provide user-friendly responses to queries. By leveraging LlamaIndex, we aim to efficiently index and query document contents, offering a scalable solution for handling policy documents and similar texts.

## Data Sources

Document Source: HDFC-Life-Group-Term-Life-Policy PDF

Purpose: The HDFC Policy PDF serves as the primary data source for this project. The document contains detailed information about insurance policies, which will be parsed and indexed.

Format: PDF

Content: Policy terms, conditions, coverage details, and other relevant sections.

## Design Choices

Framework Choice: LlamaIndex is chosen for its capability to handle document parsing, indexing, and querying efficiently. It integrates well with various vector storage solutions and provides a robust querying engine.

Document Parsing: The PDF is parsed into structured data using LlamaIndex’s `SimpleNodeParser`. This choice allows us to handle complex documents with multiple sections and formats.

Vector Storage: For simplicity and to maintain local control, the vector storage is handled directly within the LlamaIndex framework without external databases.

Query Processing: The system uses LlamaIndex’s querying capabilities to search through the indexed document and retrieve relevant information based on user queries.

Integration with LLM: While not required in the current implementation, integration with a Language Model like OpenAI's GPT could be considered for enhanced interaction and natural language understanding.

## Challenges Faced

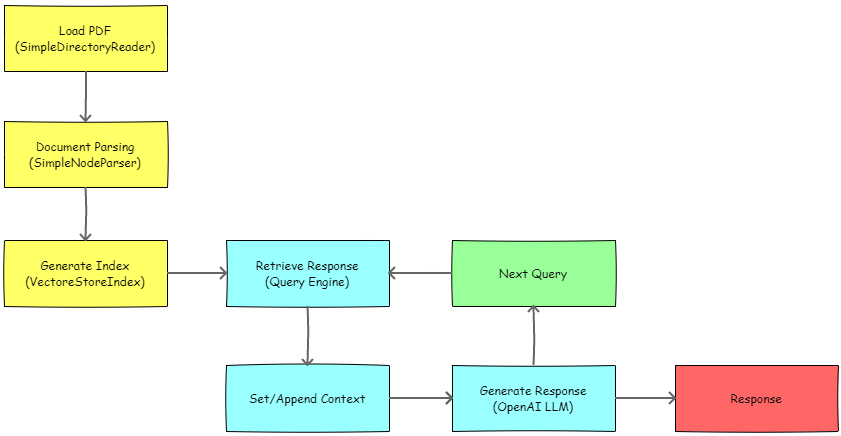
Complexity of PDF Documents: Parsing PDFs can be challenging due to the complex layout and varied formatting of documents. Ensuring that all relevant information is accurately extracted and indexed required careful handling.

Indexing Efficiency: Efficiently indexing large documents while maintaining performance and accuracy was a key challenge. Optimizations were needed to balance indexing speed and query performance.

Query Accuracy: Ensuring that the query results are accurate and relevant required fine-tuning of the parsing and indexing processes. Handling different types of queries and ensuring correct results involved iterative testing and adjustments.

## Flowchart

A flowchart illustrating the system design and various layers of the project:



## Setup

1. Place the HDFC Policy PDF file in the project directory.

2. Update the PDF path in the code if necessary.

## Usage

### Import Libraries:

If it’s a fresh environment or the libraries aren’t installed, then uncomment the first ‘pip install’ cell block and execute to get those libraries installed in your environment.

### Run the Script:

It’s a Jupyter notebook, feel free to open in your favourite editor, like VS Code or Google Collab and do Run All

### Query the System:

Enter queries related to the HDFC policy to receive relevant information from the document.

## Troubleshooting

PDF Parsing Issues: Ensure the PDF is correctly formatted and accessible.

Indexing Errors: Verify that all dependencies are correctly installed and configured.

[End of Document]