Intelligent Document Finder with LlamaIndex 🦙

Installation guide:

Prerequisites Before running this project:

Python 3.6 or higher (I have used python==3.10.0) A virtual environment (recommended) Create a virtual environment:

python -m venv venv Activate the virtual environment:

On Windows: .\venv\Scripts\activate

Set up your environment variables by creating a .env file in the root directory of the project. Add your OpenAI API key to the .env file:

Create .streamlit folder and create secrets.toml and add open ai api key [openai_key=”sk-..”]

To launch the Streamlit application, run the following command in terminal: streamlit run main.py

Task 1: User Data Upload

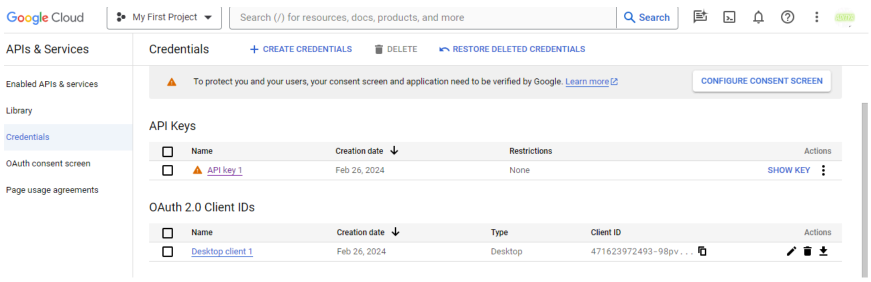
For uploading user data I have created a folder in google drive and shared folder in editor mode where anyone can upload documents of various formats (PDF, PPT, Word, etc.)

Then i have gone through this Documentation:
https://developers.google.com/drive/api/guides/about-sdk

Where it is given how can i get access of that designated folder to read file and access data inside the file by the help of Google Drive API.

In the Google Cloud console,
Created project -> API & Services -> Credentials -> Create Credential
Also created OAuth 2.0 Client IDs

Here is the snap shot of credential.json file created on google cloud console:



To protect you and your users, your consent screen and application need to be verified by Google. Learn more

API Keys

□	Name	Creation date	Restrictions	Actions
□	API key 1	Feb 26, 2024	None	SHOW KEY

OAuth 2.0 Client IDs

□	Name	Creation date	Type	Client ID	Actions
□	Desktop client 1	Feb 26, 2024	Desktop	471623972493-98pv...	

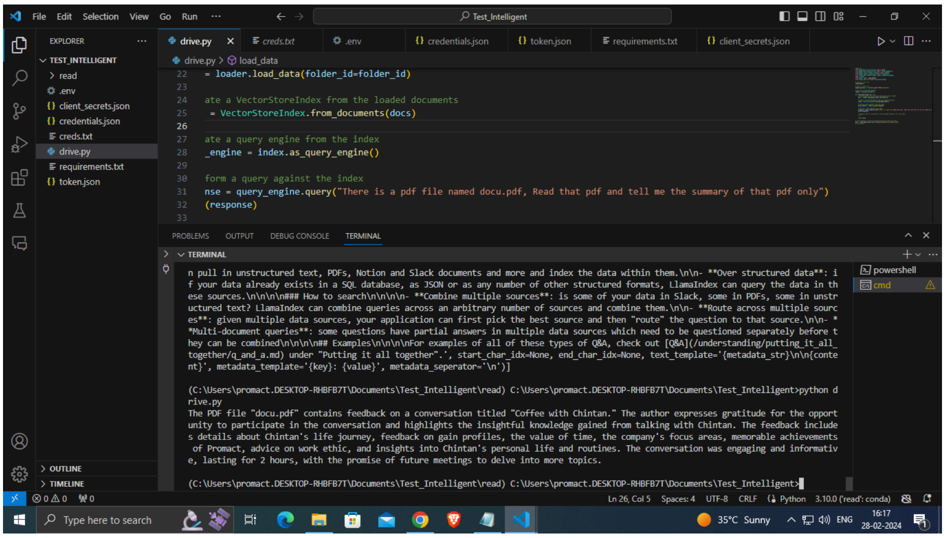
Downloaded Credentials.json file and stored in main project folder.

Authenticated with google drive and a message " The authentication flow has completed" displayed on the screen indicates that authentication is successfully implemented.

Roadblock(while implementing this feature): File name and File ID was successfully fetching from google drive but when i was trying to fetch the data inside the file, I was getting Authentication error, Then i researched how can i solve this roadblock but i was not getting the desired solution of this problem so then i started to again checked my OAuth Consent Screen and i have checkmarked scopes of drive.metadata.readonly.

SCOPES = ['https://www.googleapis.com/auth/drive.metadata.readonly']

Then i have written code to reauthenticate with google drive API. Then i was getting the result and roadblock solved successfully here is the snapshot of content along with meta data of files located at the given folder ID



```
 = loader.load_data(folder_id=folder_id)

ate a VectorStoreIndex from the loaded documents
 = VectorStoreIndex.from_documents(docs)

ate a query engine from the index
_engine = index.as_query_engine()

form a query against the index
nse = query_engine.query("There is a pdf file named docu.pdf, Read that pdf and tell me the summary of that pdf only")
(response)
```

```
n pull in unstructured text, PDFs, Notion and Slack documents and more and index the data within them.\n\n- **Over structured data**: i
f your data already exists in a SQL database, as JSON or as any number of other structured formats, LlamaIndex can query the data in th
ese sources.\n\n\n\n### How to search\n\n\n\n- **Combine multiple sources**: is some of your data in Slack, some in PDFs, some in unstr
uctured text? LlamaIndex can combine queries across an arbitrary number of sources and combine them.\n\n- **Route across multiple sourc
es**: given multiple data sources, your application can first pick the best source and then "route" the question to that source.\n\n- *
*Multi-document queries**: some questions have partial answers in multiple data sources which need to be questioned separately before t
hey can be combined\n\n\n\n## Examples\n\n\n\nFor examples of all of these types of Q&A, check out [Q&A](/understanding/putting_it_all_
together/q_and_a.md) under "Putting it all together".', start_char_idx=None, end_char_idx=None, text_template='{metadata_str}\n\n{conte
nt}', metadata_template='{key}: {value}', metadata_seperator='\n')]

(C:\Users\promact.DESKTOP-RHBFB7T\Documents\Test_Intelligent\read) C:\Users\promact.DESKTOP-RHBFB7T\Documents\Test_Intelligent>python d
rive.py
The PDF file "docu.pdf" contains feedback on a conversation titled "Coffee with Chintan." The author expresses gratitude for the opport
unity to participate in the conversation and highlights the insightful knowledge gained from talking with Chintan. The feedback include
s details about Chintan's life journey, feedback on gain profiles, the value of time, the company's focus areas, memorable achievements
 of Promact, advice on work ethic, and insights into Chintan's personal life and routines. The conversation was engaging and informativ
e, lasting for 2 hours, with the promise of future meetings to delve into more topics.

(C:\Users\promact.DESKTOP-RHBFB7T\Documents\Test_Intelligent\read) C:\Users\promact.DESKTOP-RHBFB7T\Documents\Test_Intelligent>
```

Now google drive reader is working fine and we can read data from any given folder ID using google drive API.

Task 2: Automated Data Storage and Indexing

As we know there are five stage in Retrieval Augumented Generation:
Loading -> Indexing -> Storing -> Querying -> Evaluating

Loading Stage:
Nodes and Documents: A document is a container around any data source - for instance, a PDF, an API output, or retrieve data from a database. A node is the atomic unit of data in LlamaIndex and represents a "chunk" of a source document. Nodes have metadata that relate them to the document they are in and to other nodes,.

Connectors: A data connector like here we are using Google Drive Reader to ingests data from data source that is Folder ID into documents and nodes.

Indexing Stage:

Once we have ingested our private data from google drive, Llama index help to index the data into a structure that's easy to retrieve,. This usually generating vector embeddings which are stored in a specialized database called vector store (eg Chroma DB, Vector Store Index). Indexes can also store a variety of metadata about our private data.

Embeddings: LLMs generate numerical representation of data called embeddings. When filtering data for relevance, LlamaIndex will convert queries into embeddings, and vector store will find data that is numerically similar to the embedding of our query.

Querying Stage:
Retrievers: A retriever defines how to efficiently retrieve relevant context from an index when given a query.

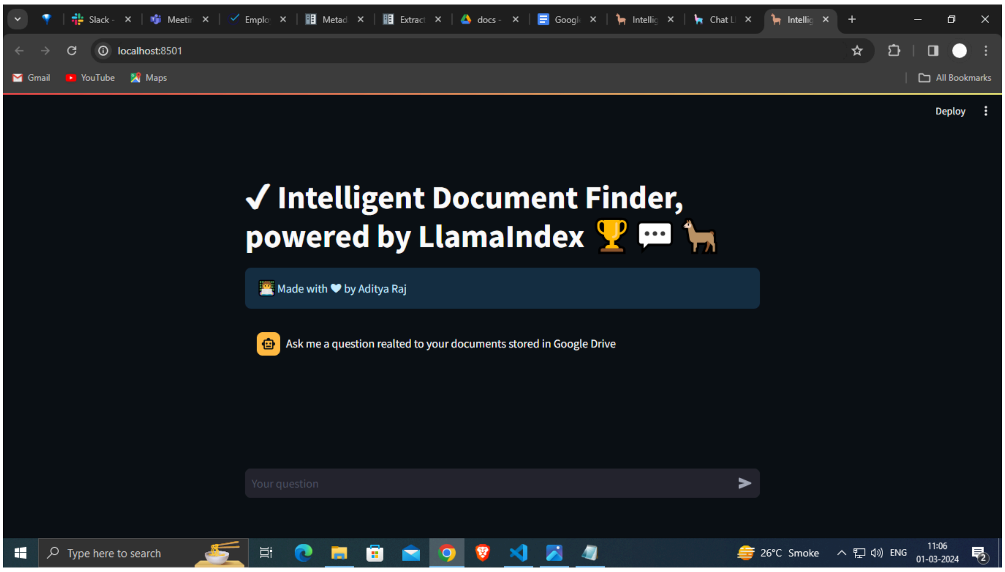
Routers: A routers determines which retriever will be used to retrieve relevant context from the knowledge base. More specifically , the Router Retriever class, is responsible for selecting one or multiple candidate retrievers to execute a query. They use a selector to choose the best option based on each candidate's metadata and the query

Node Postprocessors: A node postprocessor takes in a set of retrieved nodes and applies transformations, filtering, or re-ranking logic to them.

Response Synthesizers: A response synthesizer generates a response from an LLM, using a user query and a given set of retrieved text chunks.

Task 3: Development of a query interface

Developed an UI by using Streamlit UI to build light weight UI for our project. Here is the snap shot of the Streamlit UI



✓ Intelligent Document Finder, powered by LlamaIndex 🏆 💬 🦙

Made with ❤ by Aditya Raj

Ask me a question realted to your documents stored in Google Drive

Your question

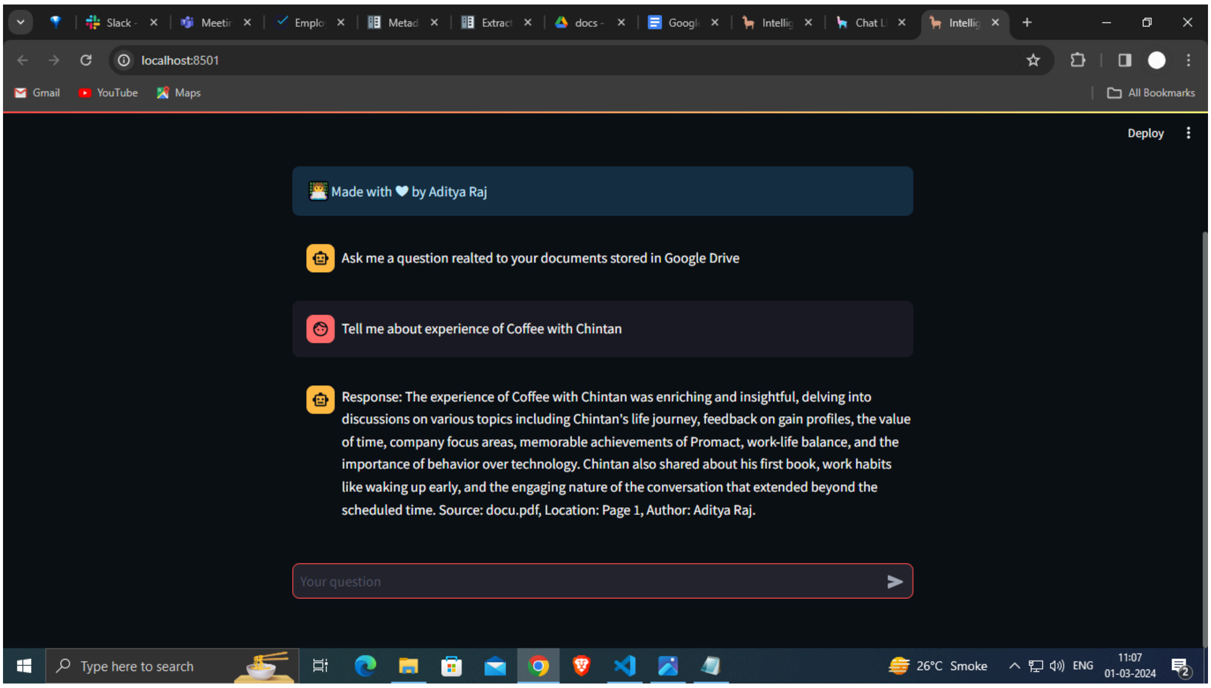
Here i have tested my project where multiple files are located and asking query related to one of the pdf, Here is the result along with meta data: :

Response:[Generated by LLM]

Source:[File name]

Location:[Page number from where the response is generated]

Author: [Name of file creator]



Made with ❤ by Aditya Raj

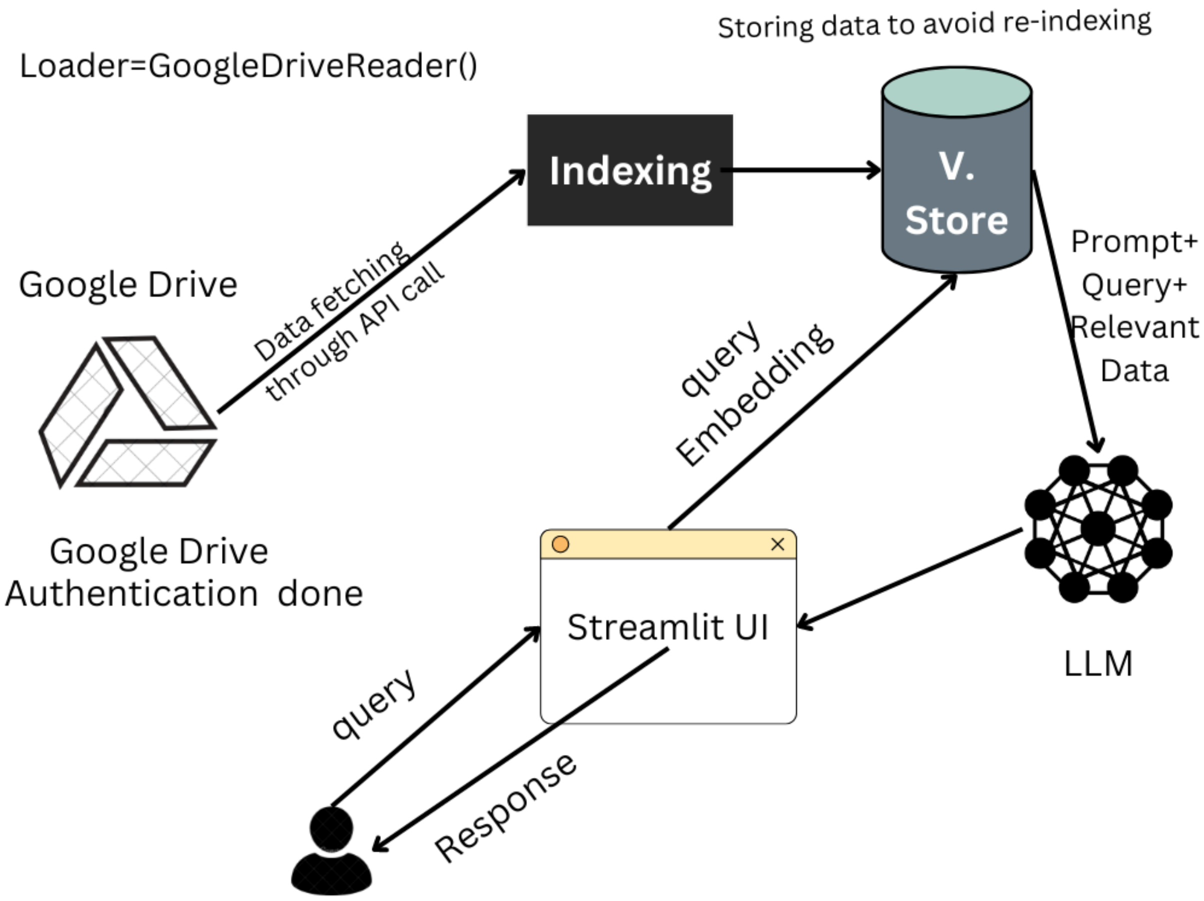
Ask me a question realted to your documents stored in Google Drive

Tell me about experience of Coffee with Chintan

Response: The experience of Coffee with Chintan was enriching and insightful, delving into discussions on various topics including Chintan's life journey, feedback on gain profiles, the value of time, company focus areas, memorable achievements of Promact, work-life balance, and the importance of behavior over technology. Chintan also shared about his first book, work habits like waking up early, and the engaging nature of the conversation that extended beyond the scheduled time. Source: docu.pdf, Location: Page 1, Author: Aditya Raj.

Your question

Flow chart of Intelligent_Document_Finder:


Loader=GoogleDriveReader()
Storing data to avoid re-indexing
Indexing
V. Store
Google Drive
Data fetching through API call
Prompt+ Query+ Relevant Data
query Embedding
Google Drive Authentication done
Streamlit UI
LLM
query
Response

Flow chart created on Canva

Documented by: Aditya Raj

Thank You!