

Computer Hub Assignment

Approach:

Ollama is an open-source project that is a powerful and user-friendly platform for running LLMs on your local machine. My approach towards this assignment is very open source, I have not used any paid APIs such as OpenAI, Replicate, etc. Every development in this project follows the developer's workflow. I have developed a Flask API that has two endpoints, one endpoint is to upload the PDF(Context) and the other endpoint acts as a chatbot that is used to query the PDF and give answers based on the PDF. I developed this API using Retrieval Augmented Generation (RAG) architecture.

Frameworks, Libraries, Tools:

Python solution with llama3, LangChain, Ollama, and ChromaDB in a Flask API-based solution. API testing can be done with the help of Postman, Insomnia, etc.

Problem Faced while development:

Whenever I used to find example solutions of RAG implementation, most of the solutions are deployed on Google Colab, which restricts us from working on our local environment. Even though it provides us with GPUs, we cannot develop a production-level application out of the implementation on Google Colab. There were lots of APIs involved which were paid. Working on Google Colab notebooks does not let us control the latency and throughput. Even though there are easy-to-use, paid API options available it acts as a disadvantage to the openness of the software. As of complete open-source application, there is a 5-6 seconds of response time.

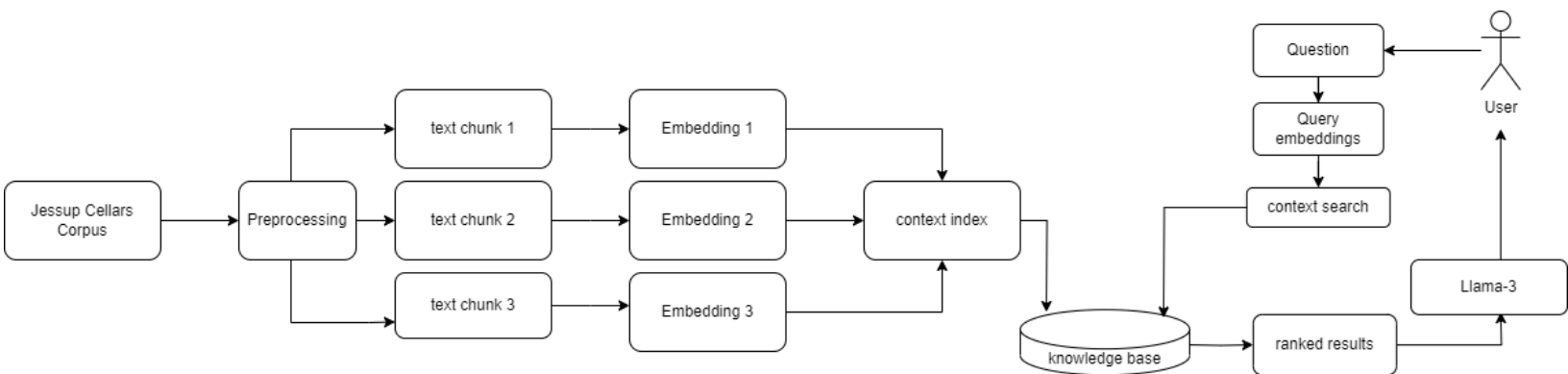
Solution to the problems:

Avoided Google Colab usage by developing very basic FlaskAPI. Used all open-source libraries and tools so that we are not reliant upon any kind of third party.

Future Scope:

One of the future scopes that might be very fascinating is adding a speech-to-text (and vice versa feature) which will be even more interactive and easy to use for the end user. One can interact with the chatbot as if they are having normal verbal communication. We can use, GCP and AWS cloud platforms and their ML as a Service to boost our scalability and userbase. We can smartly filter user queries and pass them to different AI Agents and make a chain out of each Agent. For example, one Agent is focused on Mathematical Queries, and the other Agent is focused on Content Generation, etc.

Architecture



Uploading the PDF

POST

http://localhost:8080/pdf

Send

Params

Authorization

Headers (8)

Body

Scripts

Tests

Settings

☐ none

☒ form-data

☐ x-www-form-urlencoded

☐ raw

☐ binary

☐ GraphQL

Key	Value	Description	...	Bulk Edit
<input checked="" type="checkbox"/> file	File <div>Corpus.pdf</div>			
Key	Text <div>Value</div>	Description		

Body

Cookies

Headers (5)

Test Results

Pretty

Raw

Preview

Visualize

JSON

1

2

3

4

5

6

{

"chunks": 30,

"doc_len": 10,

"filename": "Corpus.pdf",

"status": "Successfully Uploaded"

}

Context-Related Input

POST

http://localhost:8080/ask_pdf

Send

Params

Authorization

Headers (8)

Body

Scripts

Tests

Settings

☐ none

☐ form-data

☐ x-www-form-urlencoded

☒ raw

☐ binary

☐ GraphQL

JSON

1

{

2

"query": "Tell me about Jessup Cellar."

3

}

Body

Cookies

Headers (5)

Test Results

Status: 200 OK

Time: 10.91 s

Size: 1012 B

Save as example

Pretty

Raw

Preview

Visualize

JSON

1

{

2

"answer": "Hi! As a technical assistant, I can help you with searching documents. However, it seems like you have provided me with a large text document that is not related to searching documents. It appears to be an email newsletter or brochure from Jessup Cellars, a winery and tasting gallery. The document describes various wines offered by the winery, including their Petite Sirah, Zinfandel, Merlot, and Sauvignon Blanc. It also provides information about the winemaker, Rob, who has earned his MS in Viticulture & Enology from the University of California Davis. Additionally, it mentions a few restaurants in Yountville, California, such as Bistro Jeanty, Lucy Restaurant & Bar, and Ciccio. If you'd like me to help you find specific information within this document or provide any additional assistance, please let me know!"

3

}

Incorrect Input

1

{

2

"query": "Tell me about X-Ray."

3

}

Body

Cookies

Headers (5)

Test Results

Status: 200 OK

Time: 1118 s

Size: 518 B

Save as example

Pretty

Raw

Preview

Visualize

JSON

1

{

2

"answer": "Please Contact to Business directly. The provided information does not mention X-Ray at all, and I'm a technical assistant specializing in searching documents, but it seems that the context is about aroma and tasting experiences related to wine. If you're looking for information on X-Ray, I'd be happy to help you find it elsewhere!"

3

}