

RAG ASSIGNMENT DOCUMENTATION

N26 AI Engineer role

02/11/2025

Gowtham B C
bcgowtham17@gmail.com

RAG ASSIGNMENT DOCUMENTATION

Declaration:

Since Gemini key wasn't given, I ran through **free tier quota** for embedding and LLM and couldn't develop or build using cloud. Hence local ollama docker was used to complete the assignment.

While 90% of the code is same and 100% of concept is same, it is totally free as well.

I will try to make the entire solution live as well: <https://rag.solventumrnd.com/> (Not in scope of assignment)

Prerequisite

Run all these commands to ensure, Cuda enabled Nvidia graphics card and docker is up and running on hardware. Stable internet and 20GB of hardware memory

1. Nvidia-smi
2. nvcc -V
3. docker
4. open docker desktop application

If Nvidia graphics card is not configured, please visit docker compose file and comment the line from 72 to 78 i.e. Deploy section
By default, GPU is enabled it since it is too slow on CPU.

Steps to run:

Please follow every step carefully:

1. have signed in GitHub account. <https://github.com/Gowtham171996/N26-Gowtham-AI-Engineer>
open <https://github.com/Gowtham171996/N26-Gowtham-AI-Engineer/archive/refs/heads/master.zip>
2. Extract the zipped file
3. Open command prompt in folder path and hit DIR
4. Make sure n26rag-compose.yml file is visible. If not you are in wrong path.

RAG ASSIGNMENT DOCUMENTATION

5. Run "docker compose -f n26rag-compose.yml up -d"

(Make sure with Cuda GPU enabled or only CPU) Read prerequisites.

6. Wait for 10mins (pip requirements are very bulk in size, please wait)

7. Open the URLs:

- <http://localhost:8000/> (for main application)
- <http://localhost:6333/dashboard> (for qdrant)
- <http://localhost:8000/docs> (for docs)

8. The UI will load with files ingested. if not click "Trigger full ingestion" button

9. Ask a query to the model and observe the links it pulled information from.

10. Want to modify ingestion? In "data" folder add a new file (txt, word doc, pdf with text)

11. Delete the vector collection by clicking on UI

12. Perform step 7 to ingest new file.

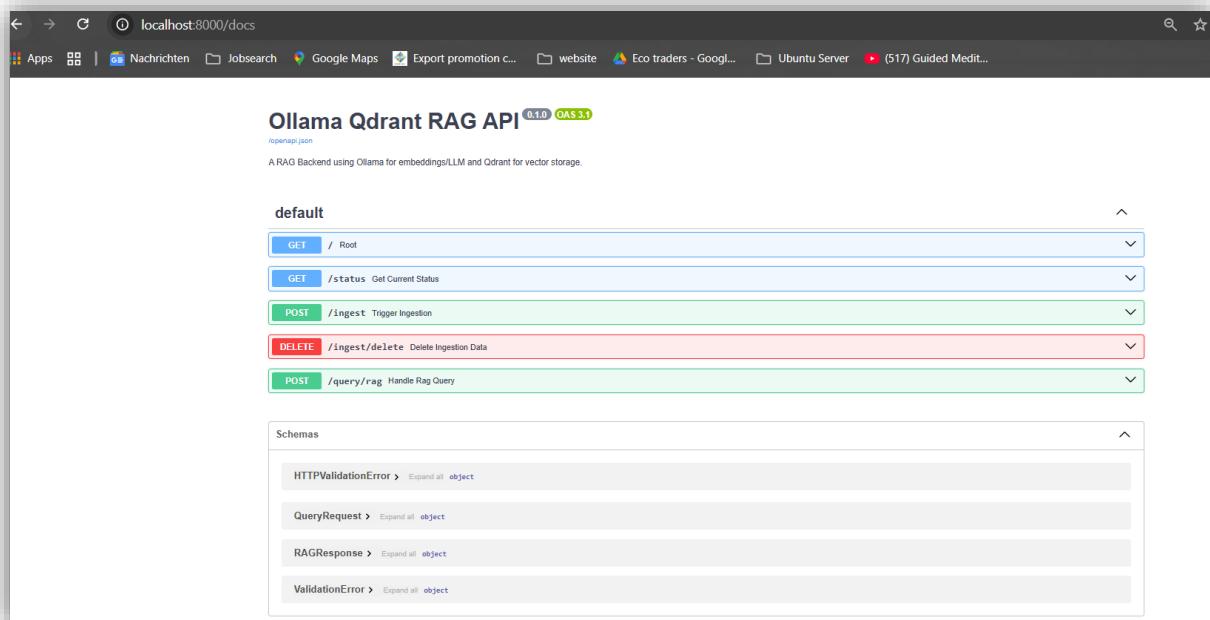
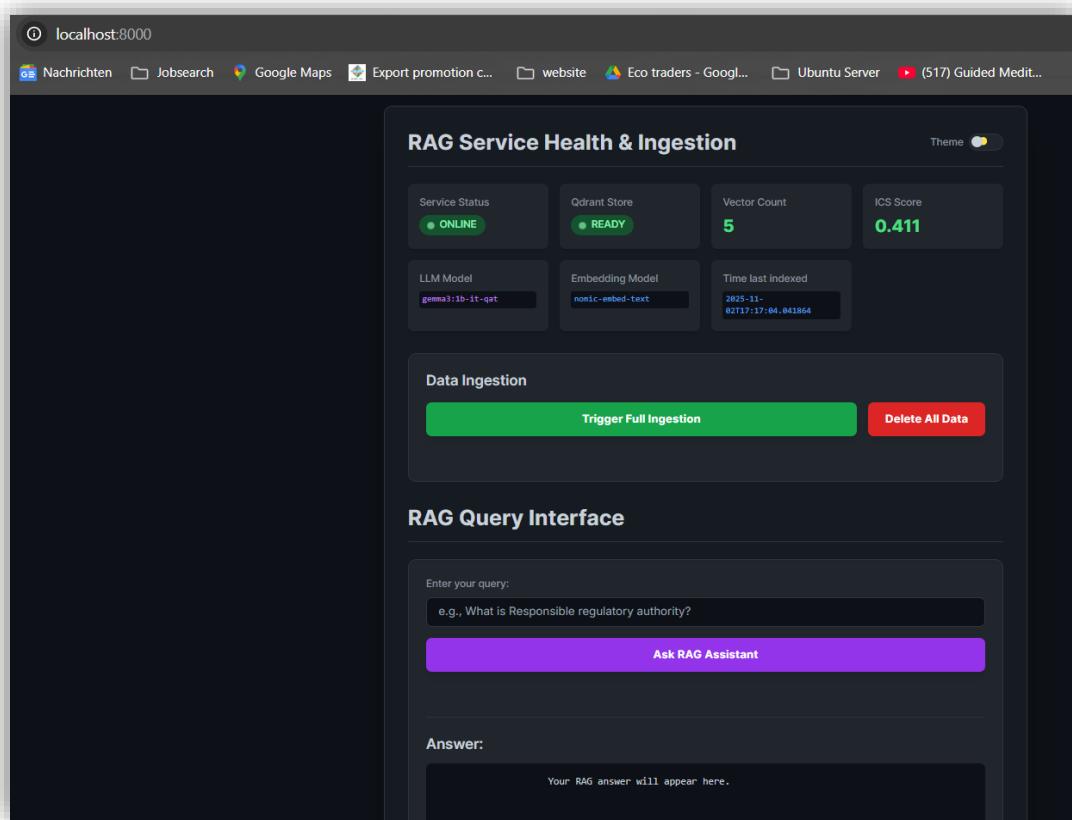
13. Perform step 8 with newly added file.

Known Issues:

1. Twice ingestion is taking place, even though ingestion is persistent.
2. Docker build images are heavy. (For Cuda enabled environment, along with reranker algorithm. PIP requirements are too much.)

RAG ASSIGNMENT DOCUMENTATION

Gallery



RAG ASSIGNMENT DOCUMENTATION

RAG Query Interface

Enter your query:

What is the responsible regulatory authority for N26?

Ask RAG Assistant

Answer:

The legal authority responsible for N26 is the Federal Financial Supervisory Authority (BaFin). You can find more information on the BaFin's website: [https://www.bafin.de] (https://www.bafin.de)

Sources:

Source 1:
File: legal documents.txt
Similarity Score: 0.7052

Source 2:
File: banks documents.txt
Similarity Score: 0.5922

Source 3:
File: money beam.txt
Similarity Score: 0.5273

Source 4:
File: money beam.txt
Similarity Score: 0.4633

Source 5:
File: banks documents.txt
Similarity Score: 0.3213

The screenshot shows the Qdrant web interface at localhost:6333/dashboard#/collections. The interface has a dark theme. On the left is a sidebar with icons for collections, embeddings, and other features. The main area is titled "Collections". It contains a search bar and a table with the following data:

Name	Status	Points (Approx)	Segments	Shards	Vectors Configuration (Name, Size, Distance)	Actions
ai_knowledge_base	green	5	6	1	default 768 Cosine	⋮