🧾 🎯 **Project Title:** **RETRIEVAL-AUGMENTEDGENERATION - RAG WEB ASSISTANT**  
📅 **Project Timeline:** August 2024 – January 2025  
🎥 YouTube Demo: Not available  
📦 GitHub Source Code: <https://github.com/IvanSicaja/2024.12.31_GitHub_Retrieval---Augmented-Generation---RAG-Web-Assistant>   
----------------------------------------------------------------------------------------------------------------

🏷️ My Personal Profiles: ⬇︎  
🎥 Video Portfolio: To be added  
📦 GitHub Profile: <https://github.com/IvanSicaja>  
🔗 LinkedIn: <https://www.linkedin.com/in/ivan-si%C4%8Daja-832682222>  
🎥 YouTube: <https://www.youtube.com/@ivan_sicaja>  
 ----------------------------------------------------------------------------------------------------------------

### 📚🔍 Project description: ⬇︎⬇︎⬇︎

### 💡 App Purpose

Imagine you have a **topic** (text, web page, company…) and popular **LLMs** (e.g. **OpenAI ChatGPT**) know a little bit or completely nothing about it, and you want to get a beautifully structured **answer** based on your topic.

This project is exactly about that — about **RAG - Retrieval Augmented Generation**, which uses a **custom knowledge base** to generate structured answers in beautiful text form.  
I am a person who is really hard to impress because I know a lot from different fields, e.g., **Autonomous Robotics**, **Robotics Engineering** (Mobile robots and drones), **Computer Vision**, **NLP**, **AI** (**LLMs, Transformers**), **CAD/CAM**, **3D Printing**, **Software Development**, **Physics**, **Mathematics**… I am completely impressed by **RAG's architecture**. It looks at the end that **“Attention is all we need.” 😉**

An **intelligent text generation and retrieval system** has been developed here by integrating **state-of-the-art Natural Language Processing models**. This system demonstrates expertise in integrating **Natural Language Understanding (NLU)**, **Information Retrieval (IR)**, and **Generative AI techniques** to solve complex problems.

### 🧠 How It Works

This project involved the following:

* **Text Preprocessing:** Utilized **spaCy** for tokenization, lemmatization, and removal of stop words and punctuation to ensure uniformity in text analysis.
* **Knowledge Base Construction:** Created a **knowledge base** from an **Excel dataset** and structured it into meaningful **chunks** for document retrieval.
* **Embedding Generation and Search:** Employed the **SentenceTransformer** model (**all-MiniLM-L6-v2**) to generate **embeddings** and implemented a **FAISS index** for fast and scalable **document similarity searches**.
* **Text Generation Pipeline:** Integrated the **LLaMA** model (**meta-llama/Llama-3.2-1B-Instruct**) for generating **context-aware responses** using **Hugging Face's Transformers library**.
* **Optimization for Performance:** Configured the environment to leverage **GPU acceleration** for efficient processing and utilized advanced **text generation techniques** like **temperature sampling**, **top-k**, and **top-p filtering** to balance creativity and relevance in responses.
* **End-to-End System Design:** Designed **functions** for **user query handling**, **document retrieval**, and generating **contextually rich responses** using retrieved documents.

All project files are available on my **GitHub page:**

* One **script** using **internal knowledge base**
* One **script** using **external “Excel” knowledge base**
* One **script** for **automatic question generation** to experiment with **pipeline parameters**
* **Original webpage text**
* **Possible pipeline parameters document**
* **All Python dependencies**
* **Questions to ask the model**

### ⚠️ Note

None.

### 🔧 Tech Stack

**RAG - Retrieval-Augmented Generation, NLP - Natural Language Processing, LLMs - Large Language Models, Transformers architecture, Text Preprocessing (spaCy), SentenceTransformers (all-MiniLM-L6-v2), Hugging Face, LLaMA (meta-llama/Llama-3.2-1B-Instruct), FAISS - Facebook AI Similarity Search, PyTorch, Python**

### 📣 Hashtags Section

**# #rag #retrievalaugmentedgeneration #nlp #transformers #llm #spacy #huggingface #llama #faiss #pytorch #python #machinelearning #ai #deeplearning #knowledgebase**