

RAG

Enzo Bognar

Por que usar RAG (Retrieval-Augmented Generation)?

- LLMs possuem conhecimento limitado ao que foi treinado.
- Sem acesso externo, modelos podem alucinar informações.
- Muitos problemas reais dependem de documentos, relatórios, PDFs, textos.
- Precisamos de um método para consultar informações externas com precisão.

Sistema RAG implementado em Python + HuggingFace

- **Tecnologias utilizadas:**

SentenceTransformers / HuggingFace → gerar embeddings

FAISS → busca vetorial

FLAN-T5 (HuggingFace Transformers) → gerar resposta

Google Colab → ambiente de execução

- **Funcionamento do sistema:**

Usuário faz uma pergunta.

Pergunta é convertida em embedding.

FAISS busca os documentos mais semelhantes (similaridade vetorial).

Esses documentos são enviados ao modelo gerativo.

O modelo responde usando as informações recuperadas.

- **Resultado esperado:**

O modelo responde com base em fatos da base de conhecimento, evitando alucinações.

Demonstração e Resultados

Conclusão:

O sistema funciona como um pipeline RAG básico, mostrando recuperação + geração em funcionamento real.

```
[9] ✓ 10s
query = "Para que serve o RAG?"
answer, retrieved = rag_answer(query)

print("⭐ Pergunta:", query)
print("\n📋 Documentos Recuperados:\n", retrieved)
print("\n🤖 Resposta do Modelo:\n", answer)

...
Both `max_new_tokens` (=256) and `max_length` (=150) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main\_classes/text\_generation)
⭐ Pergunta: Para que serve o RAG?

📋 Documentos Recuperados:
LMs podem usar RAG para acessar informações externas sem precisar memorizar tudo nos pesos do modelo.
FAISS é uma biblioteca do Facebook para busca vetorial em grandes coleções de embeddings.

🤖 Resposta do Modelo:
LMs can use RAG to access external information without needing to memorize all the weights of the model. FAISS is a library from Facebook for vector search in large embedding collections.
```