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 generativo.

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.