

Sistema RAG com HuggingFace

Pequena Aplicação em Python

Thiago Sales

Problema

- Sistemas precisam responder perguntas com base em dados próprios.
- Modelos de linguagem não têm acesso direto a documentos locais.
- Respostas ficam desatualizadas ou fora de contexto sem recuperação.

Solução Desenvolvida

- Implementação de um sistema RAG (Retrieval-Augmented Generation).
- Uso de embeddings + FAISS para busca vetorial eficiente.
- Modelo HuggingFace (FLAN-T5) gera respostas usando o contexto recuperado.

Funcionamento

- Documentos → transformados em embeddings.
- FAISS → indexa e recupera textos semelhantes.
- LLM → gera resposta baseada no contexto encontrado.
- Resultado: respostas mais precisas, atualizadas e contextualizadas.