Solução: Mini-aplicação RAG com HuggingFace

Participantes:
Guilherme Araujo
João Vitor Corazzari
Vitor Hugo Alvarenga Facciolo

Responder perguntas com base em uma base de textos específica

Usuários precisam de respostas fundamentadas em um conteúdo específico, como:

- Apostilas
- Documentos internos
- Notas de aula
- Políticas corporativas
- Bases de conhecimento privadas

LLMs tradicionais (“puros”) não possuem acesso direto a esses conteúdos:

- Não conhecem documentos privados ou recentes.
- Podem gerar respostas genéricas, desatualizadas ou imprecisas.
- Não conseguem citar informações específicas sem acesso ao material.

Surge a necessidade de combinar busca documental com geração de texto, por meio de:

- Indexação e recuperação: localizar trechos relevantes dentro da base de textos.
- RAG (Retrieval-Augmented Generation): o modelo lê os documentos encontrados e produz uma resposta context
- Precisão e rastreabilidade: respostas fundamentadas em evidências reais do conjunto de documentos.
- Atualização dinâmica: sempre responder considerando a versão mais atual da base de conhecimento.

Trabalho AF #11 – Tópico: RAG (Retrieval Augmented Generation)

Solução: Mini-aplicação RAG com HuggingFace

Base de conhecimento:

- Textos armazenados em lista ou arquivo .txt com conteúdos sobre IA e RAG.

Embeddings com SentenceTransformer:

- Uso do modelo all-MiniLM-L6-v2 para transformar cada texto em vetores semânticos.

Busca vetorial com FAISS:

- Criação de um índice para localizar rapidamente os trechos mais parecidos com a pergunta.

LLM para geração de resposta:

- Ex.: TinyLlama-1.1B-Chat.
- Recebe a pergunta + contexto recuperado e produz uma resposta fundamentada.

Execução, Resultados e Próximos Passos

Exemplo de pergunta:

- “O que é RAG em Inteligência Artificial?”

Funcionamento observado:

- Sistema recuperou 2 trechos relevantes da base.
- O LLM gerou uma resposta explicando o conceito apenas com o contexto recuperado.

Limitações:

- Base de textos pequena.
- Modelo de linguagem pequeno → respostas simples.

Próximos passos:

- Ampliar a base (PDFs, artigos, apostilas).
- Testar modelos mais robustos (Gemma, Llama, etc.).
- Criar interface web (Gradio ou Streamlit).