

RAG: Retrieval-Augmented Generation

Combinando recuperação de informações e geração de texto para respostas mais precisas e contextualizadas

O Desafio: Modelos Limitados por Conhecimento Estático

Problemas Comuns

Modelos de linguagem tradicionais enfrentam limitações significativas quando precisam responder perguntas sobre domínios específicos ou informações atualizadas.

- Conhecimento congelado no momento do treinamento
- Incapacidade de acessar bases de dados corporativas
- Respostas genéricas sem contexto específico
- Alto custo para retreinar com novos dados

Impacto no Negócio

Essas limitações resultam em sistemas que não conseguem aproveitar o conhecimento interno das organizações de forma eficiente.

- Respostas imprecisas ou desatualizadas
- Impossibilidade de integrar documentação técnica
- Necessidade de intervenção manual constante
- Baixa confiabilidade em aplicações críticas



A Solução: Pipeline RAG Implementado

Desenvolvemos um sistema completo que combina recuperação semântica e geração contextualizada, permitindo respostas precisas baseadas em conhecimento específico.

01	02	03
Base de Conhecimento Documentos estruturados contendo informações técnicas sobre Python, FAISS, Hugging Face e RAG, prontos para consulta.	Embeddings Semânticos Utilizamos SentenceTransformers (all-mpnet-base-v2) para transformar textos em vetores densos de alta qualidade que capturam significado semântico.	Indexação FAISS Construímos um índice vetorial eficiente usando FAISS IndexFlatL2, permitindo busca rápida por similaridade entre milhões de vetores.
04	05	
Recuperação Inteligente Convertemos a query do usuário em embedding e recuperamos os top-k documentos mais relevantes com base em distância L2.	Geração Contextualizada O modelo FLAN-T5 recebe o contexto recuperado junto com a pergunta, gerando respostas fundamentadas e precisas.	

Componentes Tecnológicos

sentence-transformers Embeddings de alta qualidade para captura semântica	FAISS Busca vetorial eficiente em larga escala	Hugging Face Transformers Modelos pré-treinados para geração de texto
---	--	---

Esta arquitetura permite que o sistema acesse dinamicamente informações relevantes e gere respostas contextualizadas, superando as limitações dos modelos tradicionais. O resultado é um assistente inteligente capaz de responder perguntas técnicas com precisão, utilizando conhecimento específico do domínio sem necessidade de retreinamento constante.