



Construindo um Vector Database com HuggingFace

Transforme textos em vetores e crie buscas semânticas inteligentes usando modelos de embedding e FAISS



Componentes Principais



Modelo de Embedding

Transformers da HuggingFace convertem texto em vetores

- all-MiniLM-L6-v2
- multi-qa-mpnet-base-dot-v1



Base Vetorial

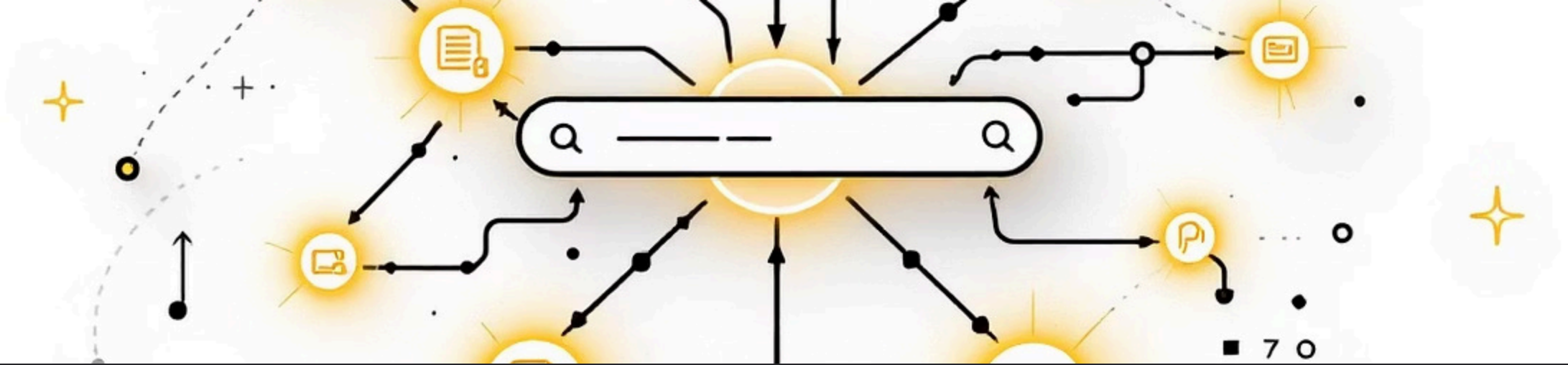
FAISS armazena vetores + metadados para busca eficiente

- Nearest Neighbors
- Alternativas: ChromaDB, Milvus



Pipeline de Uso

Coleta → Embeddings → Indexação → Consulta → RAG



Implementação Passo a Passo

Stack Tecnológico

- sentence-transformers
- faiss-cpu
- FastAPI (opcional)
- LangChain (opcional)

Arquitetura Simples

- data/ - textos base
- embeddings.py - geração
- vector_store/ - índice FAISS
- app.py - consultas

1

Gerar Embeddings

SentenceTransformer.encode
transforma textos em vetores
numpy

2

Indexar FAISS

IndexFlatIP com normalização
L2 para similaridade de cosseno

3

Consultar

Busca top-k vetores similares e
retorna textos correspondentes