alura

LangChain: Criando chatbots inteligentes com RAG

Armazenamento Vetorial com FAISS/Chroma

Instrutor(a): Leonardo Pena

Seja bem vindo (a)!

Aula

Armazenamento Vetorial

O PROBLEMA: COMO O COMPUTADOR ENTENDE O SIGNIFICADO?

Quando humanos leem palavras como "carro", "automóvel" e "veículo", entendemos que são conceitos relacionados. Mas para um computador, são apenas sequências de caracteres sem relação óbvia.

A solução: Embeddings Vetoriais

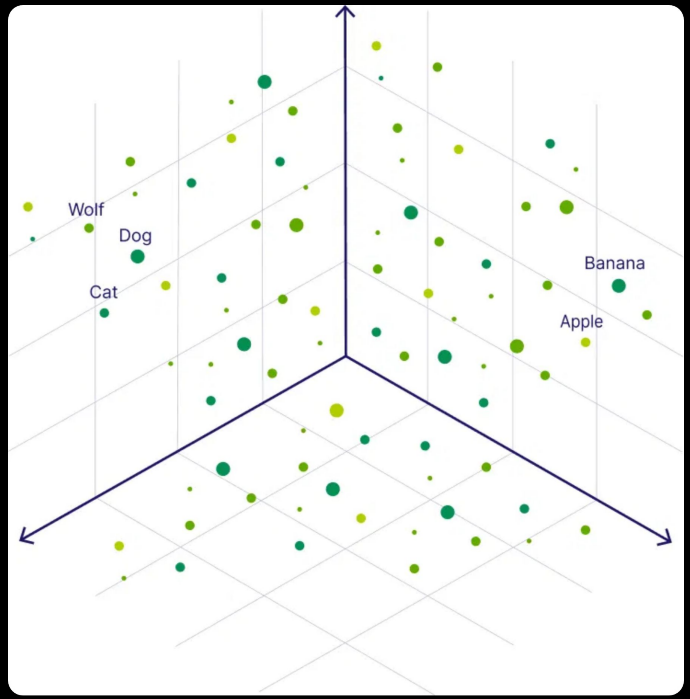
O QUE É EMBEDDINGS

Embeddings: São representações numéricas de texto em um espaço de alta dimensão.

Textos com significados semelhantes são posicionados próximos nesse espaço vetorial, permitindo que computadores comparem informações com base na semântica.

ESPAÇO VETORIAL

Os embeddings criam um **espaço multidimensional** onde cada palavra ou documento é um ponto. Neste espaço, a **proximidade geométrica** representa a **similaridade semântica .**



Visualização 3D de um espaço vetorial semântico (reduzido de centenas de dimensões)

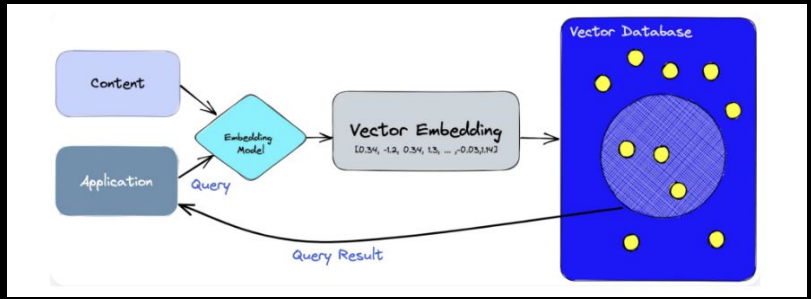
ESPAÇO VETORIAL

Quando humanos leem palavras como "carro", "automóvel" e "veículo", entendemos que são conceitos relacionados. Mas para um computador, são apenas sequências de caracteres sem relação óbvia.

Quanto mais próximos os vetores, mais relacionados os conceitos!

BANCOS DE DADOS VETORIAIS

São sistemas especializados em **armazenar, indexar e buscar vetores** de alta dimensão de forma eficiente.
Diferente dos bancos de dados tradicionais, eles são otimizados para **busca por similaridade**, não por correspondência exata.



BANCOS DE DADOS VETORIAIS

FAISS (Facebook AI Similarity Search)

Biblioteca de alto desempenho desenvolvida pelo Facebook Research para busca eficiente de vetores similares.

- Alta Performance
- Escalabilidade
- Otimizado para GPU
- Código Aberto

Chroma

Banco de dados vetorial de código aberto projetado especificamente para aplicações de IA e RAG, com foco em facilidade de uso.

- Fácil Integração
- Metadados
- API Python
- Persistência

Outras Opções

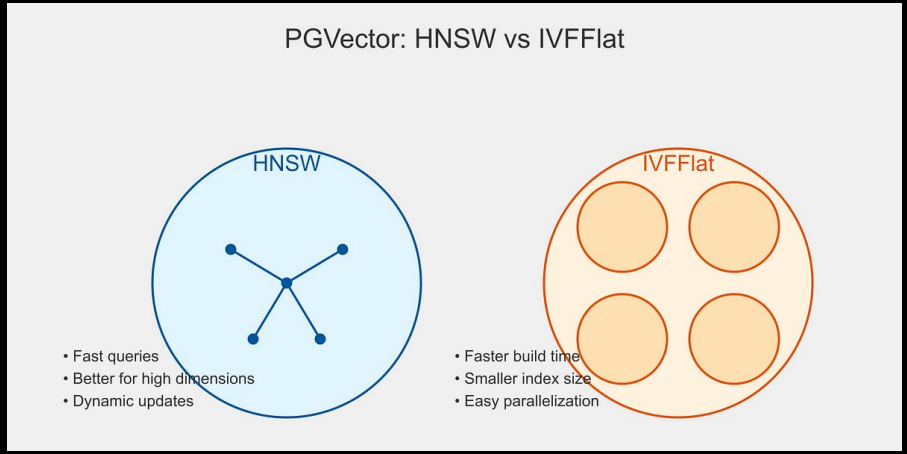
- Pinecone
- Weaviate
- Milvus
- Qdrant
- Vespa
- pgvector

A escolha do banco vetorial depende do seu caso de uso específico!

TIPOS DE ÍNDICES

Existem diferentes métodos de organização de vetores, cada um com seus trade-offs:

- **Flat (Força Bruta):** Compara a consulta com todos os vetores, garantindo precisão perfeita, mas sendo muito lento para grandes volumes de dados.
- **IVF (Inverted File Index):** Agrupa vetores em clusters e busca apenas nos mais relevantes, sendo significativamente mais rápido que o Flat.
- **HNSW (Hierarchical Navigable Small World):** Constrói um grafo multicamadas para uma busca extremamente rápida e eficiente, ideal para sistemas em larga escala.



ARMAZENAMENTO LOCAL VS NUVEM

Ambas são excelentes opções, mas têm características distintas que as tornam mais adequadas para diferentes cenários.

FAISS

\+ **Vantagens**: Extremamente rápido e eficiente para grandes volumes de dados. Suporte a GPU para aceleração.

— **Desvantagens**: Curva de aprendizado mais íngreme. Não possui persistência nativa. Sem suporte a metadados.

Ideal para:

- Grandes volumes de dados (milhões/bilhões de vetores)
- Quando a performance é crítica
- Quando há recursos computacionais abundantes (GPU)
- Projetos onde você pode construir sua própria camada de metadados

Chroma

\+ **Vantagens**: Fácil de usar e integrar. Suporte nativo a metadados e filtragem. Persistência integrada.

— **Desvantagens**: Menos otimizado para volumes muito grandes. Menos opções de configuração avançada.

Ideal para:

- Projetos de RAG de pequeno a médio porte
- Quando a facilidade de uso é prioritária
- Quando você precisa de filtragem por metadados
- Prototipagem rápida e MVPs

FILTROS POR METADADOS

É a capacidade de **refinar a busca vetorial** aplicando **filtros** sobre informações adicionais associadas a cada documento, como data, autor ou tipo de documento.

Isso permite **restringir a busca a um subconjunto específico de dados**, aumentando a relevância.

São informações adicionais associadas a cada vetor que permitem **filtrar e refinar buscas.**

Combinando busca semântica com filtros de metadados, criamos um sistema de recuperação muito mais poderoso.

FILTROS POR METADADOS

- **Busca Híbrida**: Combine a busca por similaridade semântica com filtros tradicionais (data, autor, categoria, etc).
- **Precisão Aumentada**: Restrinja resultados a um subconjunto específico de documentos que atendam a critérios exatos.
- **Performance Otimizada:** Reduza o espaço de busca aplicando filtros antes da comparação vetorial.
- **Organização Lógica**: Agrupe documentos por departamento, projeto, tipo ou qualquer outra classificação relevante.

FILTROS POR METADADOS

Exemplo de Uso

Busca em uma base de conhecimento corporativa:

"Encontre documentos sobre estratégias de marketing" + {departamento: "Marketing", ano: 2023, tipo: "Relatório"}

Busca em catálogo de produtos:

"Sapatos confortáveis para caminhada" + {categoria: "Calçados", disponível: true, preço: "< 200"}

APLICAÇÕES PRÁTICAS

Os bancos de dados vetoriais estão revolucionando diversas indústrias ao permitir buscas baseadas em significado, não apenas em palavras-chave.

Assistentes de IA Corporativos

Chatbots e assistentes virtuais que respondem perguntas com base em documentação interna e bases de conhecimento.

"Qual é o procedimento para aprovação de despesas acima de R$5.000?"

Busca Semântica Avançada

Sistemas de busca que entendem a intenção por trás das consultas, não apenas as palavras exatas.

"Restaurantes aconchegantes para um jantar romântico"

Recomendação de Produtos

Sistemas que recomendam produtos similares por características semânticas extraídas de descrições.

"Calças confortáveis para viagem" após busca por "jeans leve"

Pesquisa Médica e Científica

Ferramentas que ajudam pesquisadores a encontrar estudos relevantes em vastos repositórios de artigos.

Encontrar estudos sobre "inflamação pulmonar" mesmo com termos como "pneumonite"

Busca Multimodal

Sistemas que permitem buscar imagens ou vídeos com base em descrições textuais ou vice-versa.

"Encontre imagens de pôr do sol sobre montanhas com lagos"

Detecção de Fraudes

Sistemas que identificam padrões anômalos em transações financeiras ou comportamentos de usuários.

Identificar tentativas de phishing com linguagem nunca vista antes