alura

LangChain: Criando chatbots inteligentes com RAG

Embeddings de Alta Performance

MERGULHE EM TECNOLOGIA_

Instrutor(a): Leonardo Pena

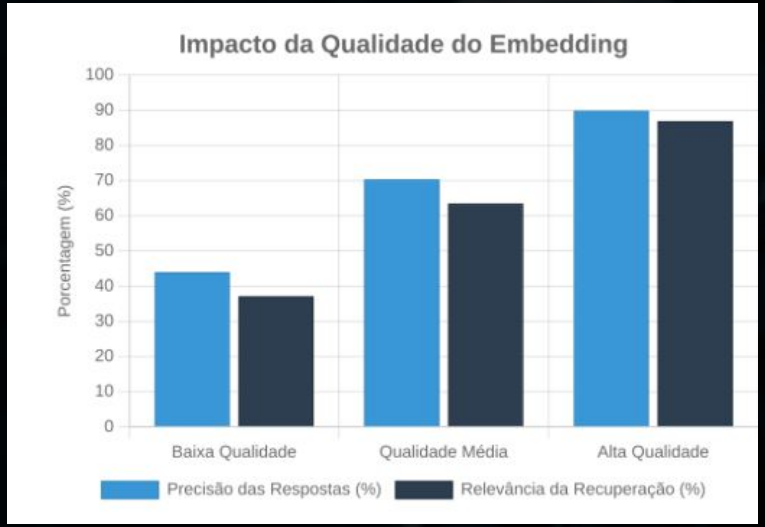
Seja bem vindo (a)!

Aula

Embeddings de Alta Performance

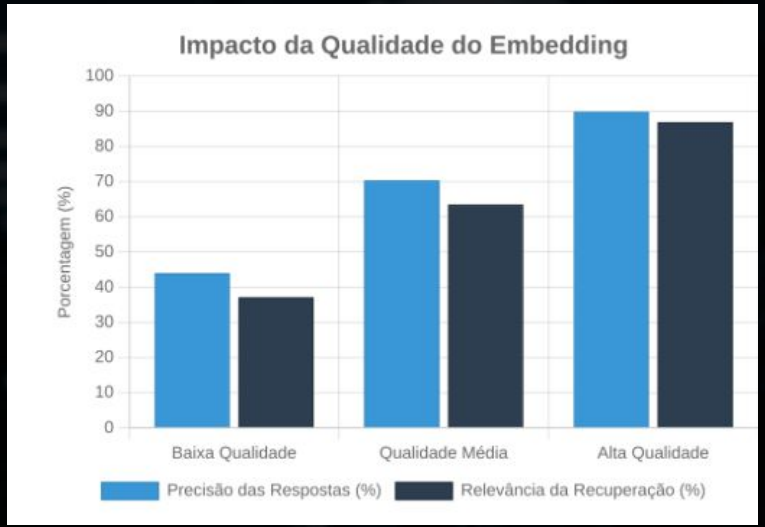
IMPORTÂNCIA DA QUALIDADE DOS EMBEDDINGS

- A qualidade da representação vetorial impacta diretamente a relevância das informações recuperadas em sistemas RAG.
- Embeddings de alta qualidade resultam em maior precisão e factualidade nas respostas geradas pelo LLM.



IMPORTÂNCIA DA QUALIDADE DOS EMBEDDINGS

- Fatores que influenciam a qualidade: modelo utilizado, tamanho do contexto, domínio específico e pré-processamento dos dados.
- A escolha do modelo de embedding deve considerar o equilíbrio entre desempenho e custo para cada aplicação específica.



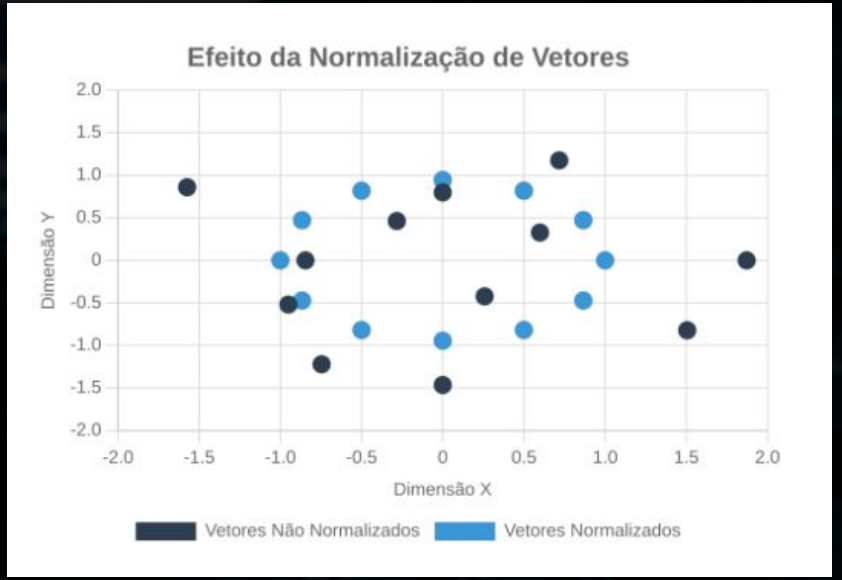
Modelos Proprietários vs. Open-Source

A escolha do modelo para gerar embeddings se divide em duas categorias principais:

Proprietários	Open-Source
Fácil uso via API	Controle total e privacidade
Ótimo desempenho geral	Possibilidade de fine-tuning
Custo por uso	Gerenciamento de infraestrutura
Dados enviados para serviços externos	Otimização para domínios específicos
Ex: OpenAI, Cohere	Ex: BGE, Sentence-Transformers

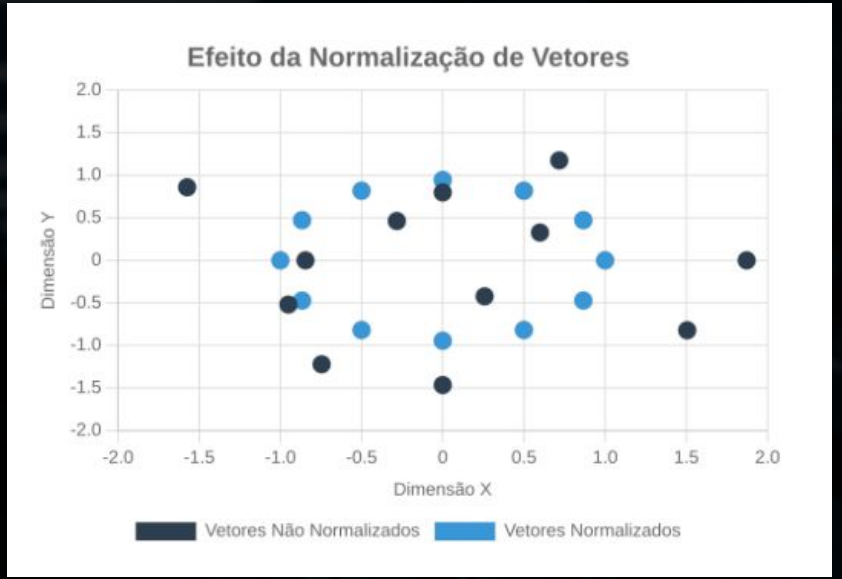
Normalização de Embeddings

- Normalização é o processo de **ajustar os vetores** para que todos tenham a mesma magnitude (comprimento).
- Vetores normalizados garantem **consistência na comparação** e melhoram o desempenho de métricas de similaridade, especialmente a similaridade de cosseno.



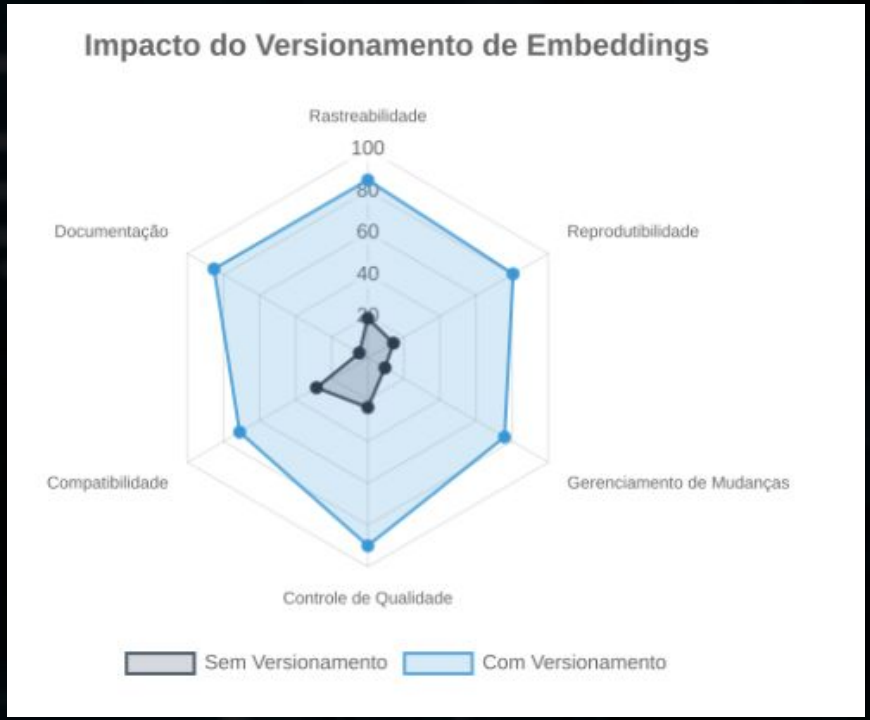
Normalização de Embeddings

- A normalização **elimina a influência do tamanho do vetor**, focando apenas na direção (ângulo) entre os vetores.
- Implementação simples: dividir cada componente do vetor pela **norma euclidiana** do vetor (raiz quadrada da soma dos quadrados).



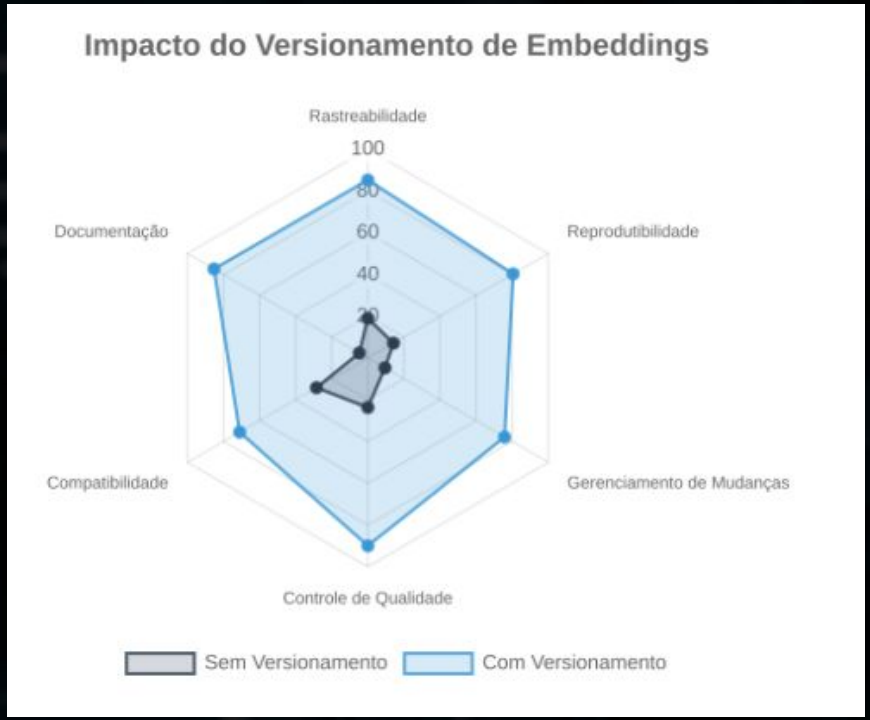
Versionamento de Embeddings

- O versionamento de embeddings é a prática de rastrear metadados sobre como os vetores foram gerados.
- Informações cruciais incluem: modelo utilizado, parâmetros de configuração, pré-processamento aplicado e data de geração.



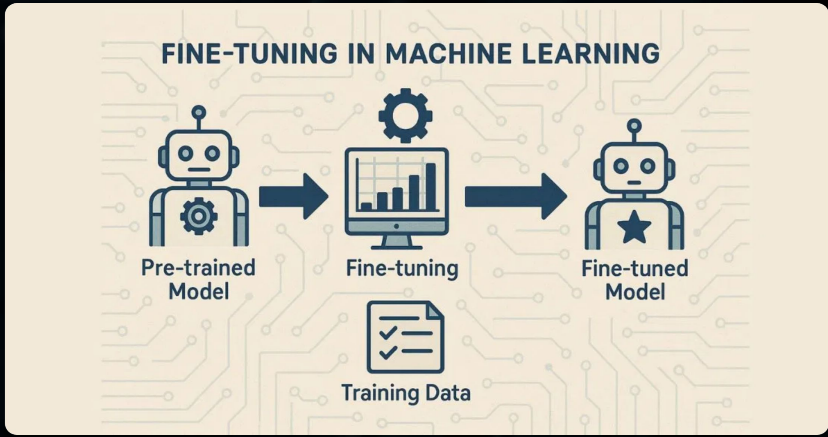
Versionamento de Embeddings

- Benefícios: rastreabilidade, reprodutibilidade dos resultados e gerenciamento controlado de atualizações.
- Permite identificar quando é necessário regenerar embeddings após atualizações significativas nos modelos ou dados.



Otimização com Fine-tuning

- Para domínios específicos, o fine-tuning de modelos de embedding pode melhorar significativamente o desempenho.
- O processo consiste em treinar adicionalmente o modelo com dados do domínio para que ele aprenda a entender melhor as nuances e terminologia específicas.



Otimização com Fine-tuning

Domínios que mais se beneficiam do fine-tuning:

- Jurídico: Terminologia legal e precedentes
- Saúde: Termos médicos e relações clínicas
- Financeiro: conceitos econômicos e mercado
- Técnico: documentação e código específico