Storing Audio Data in Vector Databases: Options and Implementation

Your Name

March 6, 2025

1 O que é uma base de dados vetorial?

Segundo a *Cloudfare* [1], uma base de dados vetorial é uma coleção de dados armazenados como representações matemáticas. Estes sistemas facilitam a memória de *inputs* anteriores, permitindo o uso de inteligência artificial para prever *outputs* futuros e recomendações. Ao invés dos dados serem identificados com técnicas de *pattern matching* ou *indexing*, os vetores são comparados com base em sua similaridade matemática.

No âmbito do nosso trabalho, um exemplo prático para a utilização destes sistemas é, por exemplo, uma plataforma de streaming de música que possui um algoritmo que recomenda músicas semelhantes àquelas que o utilizador já escutou. Para tal, o algoritmo compara os vetores das músicas que o utilizador já ouviu com os vetores de todas as músicas disponíveis na plataforma, e recomenda aquelas que apresentam maior similaridade.

Estas bases de dados permitem que os programas façam comparações, identifiquem relações e compreendam o contexto, o que permite a criação de sistemas avançados de inteligência artificial, como os modelos de linguagem de larga escala (LLMs).

2 Comparação entre Sistemas de Bases de Dados Vetoriais

Atualmente, existem diversos sistemas de bases de dados vetoriais que permitem armazenar e pesquisar dados de forma eficiente. Nesta secção, iremos comparar as opções mais populares ao nosso dispor, destacando as suas principais características, vantagens e limitações, assim como explicar o porque de escolhermos o Weaviate para o nosso sistema de armazenamento e recuperação de áudio.

2.1 Weaviate

Overview: Weaviate: Open-source, combina pesquisa vetorial com filtragem estruturada e suporta múltiplos tipos de dados (texto, imagens, áudio). Possui

uma API GraphQL flexível e escalabilidade horizontal.

2.2 Pinecone

Overview: Pinecone: Serviço totalmente gerido, fácil de usar e com alto desempenho, mas é um sistema fechado e não open-source.

2.3 Qdrant

Overview: Otimizado para pesquisas vetoriais filtráveis, oferece bom desempenho e eficiência, mas tem uma comunidade menor e menos integrações.

2.4 Chroma

Overview: Focado em aplicações de LLM e RAG, é simples de integrar e utilizar, mas ainda não é tão maduro nem escalável quanto outras soluções.

3 Porque é que escolhemos o Weviate?

Após analisar as opções, optámos pelo Weaviate para o nosso sistema de armazenamento de áudio por várias razões:

1. **Capacidade de pesquisa híbrida:** permite pesquisa por similiaridade vetorial assim como filtragem estruturada, o que é essencial para a nossa aplicação.
2. **Flexibilidade de esquema:** permite representar ficheiros de áudio com diferentes propriedades e relações.
3. **API GraphQL:** que simplifica a interação com a base de dados e permite uma integração mais fácil com outras aplicações.
4. **Open-Source:** o Weaviate tem uma comunidade ativa de desenvolvimento com atualizações regulares e melhorias.
5. **Documentação de qualidade:** agiliza o processo de desenvolvimento

Embora o Pinecone seja mais simples de configurar e escalar automaticamente, não é open-source, o que limita a personalização e controlo sobre os dados. Milvus poderia ser uma alternativa com maior desempenho em bases de dados extremamente grandes, mas o Weaviate apresentou o melhor equilíbrio entre funcionalidades, flexibilidade e facilidade de uso.

Professor, considera que esta é a melhor escolha para o nosso caso, ou acha que deveríamos considerar outra solução?

4 Representação de ficheiros de audio em vetores

4.1 Como podemos converter os ficheiros de audios em vetores ?

4.2 Plano de implementação

4.3 1º: Setup da base de dados

4.4 2º: Desenho do esquema da base de dados

4.5 3: Processar o audio

4.6 4º: Recolher os metadados do audio e converter em vetores

4.7 5º: Alimentar a base de dados com os dados que recolhemos

Usar uma API?

4.8 6º: Desenvolver queries ao sistema... desenvolver os parametros de similiaridade

4.9 7º Desenvolver uma GUI?

5 Desafios e Limitações:

Trabalhar com vetores de audio apresenta vários desafios:

- Variabilidade de qualidade do audio
- Custo computacional do pré-processamento e extração de características do audio
- Equilibrar a precisão do vetor com os requisitos de armazenamento/desempenho

References

[1] Cloudflare. What is a vector database?, n.d. Accessed on March 4, 2024.