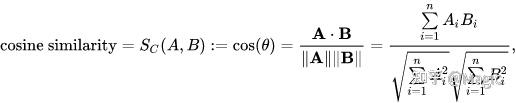
1. 什么是向量数据库？

简单下个定义，因为喂给Transformer的知识首先需要做embedding(嵌入，插入)，所以用于存储embedding之后数据的数据库即可称为向量数据库。

向量数据库需要解决什么问题？因为向量数据库是基于embedding之后的向量的存储与检索。所以首先需要提供存储能力，其次，更重要的是检索。 即如何根据一个query快速找到相关的embedding内容。 关于检索，主要是计算两个向量之间的相似度。

推荐的计算两个向量之间距离推荐的算法为： 余弦相似度函数， 其表达式如下:



1. 常见的向量数据库

向量检索是一个革命性的技术，使得开发者和工程师可以将知识或数据向量化之后实现更有效的存储、检索以及推荐。通过对比这些向量两两之间的相似性，可以实现快速、直观、无缝的信息检索。

Pinecone: 一个全托管的矢量数据库：Pinecone是一个专门为工程师与开发者设计的向量数据库。 作为一个全托管的服务，它减轻了工程师以及运维人员的负担，使得客户可以聚焦于数据内在价值的抽取。 免费版可以支持500w的向量存储，其用法简单，价格低廉，可以快速支持向量检索业务的验证与尝试。在特性上Pinecone也具有高速、准确以及可拓展等特性，此外也具备对单级元数据过滤和尖端稀疏-密集索引等高级功能。

1. 下面是一些使用案例

名称 描述

GPT-4 Retrieval Augmentation 如何通过检索增强来增强GPT4的能力

Generative Question-Answering 生成式问答

Semantic Search 语义搜索，构建一个简单的语义搜索

Weaviate: weaviate 是一个开源的矢量数据库，具有健壮、可拓展、云原生以及快速等特性

Weaviate 是一个开源的向量数据库，可以存储对象、向量，支持将矢量搜索与结构化过滤与云原生数据库容错和可拓展性等能力相结合。 支持GraphQL、REST和各种语言的客户端访问。

Redis: redis 通过RedisSearch 模块，也原生支持向量检索。

RedisSearch 是一个Redis模块，提供了查询、二级索引，全文检索以及向量检索等能力。如果要使用RedisSearch，需要首先在Redis数据上声明索引。

Qdrant: Qdrant 是面向下一代的生成式AI向量数据库，同时也具备云原生的特性。

Milvus: 面向下一代的生成式AI向量数据库，支持云原生。

Chroma: 一个开源的向量数据库。 可以快速基于Python和JavaScript构建内存级LLM应用

Zilliz, 构建在Milvus之上的数据基础设施。

1. 能力对比

数据库名称 是否开源 社区影响力 编程语言 核心特性 适用场景

Pinecone 否 未知 向量存储与检索、全托管 Saas类业务场景

weaviate 是 5.3k star Go 同时支持向量与对象的存储、支持向量检索与结构化过滤、具备主流模式成熟的使用案例。高速、灵活，不仅仅具备向量检索，还会支持推荐、总结等能力

qdrant 是 6.3k star Rust 向量存储与检索、云原生、分布式、支持过滤、丰富的数据类型、WAL日志写入

milvus 是 17.7k star Go 极高的检索性能: 万亿矢量数据集的毫秒级搜索非结构化数据的极简管理丰富的API跨平台实时搜索和分析可靠：具有很高的容灾与故障转移能力高度可拓展与弹性支持混合检索统一的Lambda架构社区支持、行业认可

Chroma 是 4.1k star python 轻量、内存级

综上所属， 在几个主流的向量数据库中， milvus在大规模、检索性能、社区影响力等方面都具备绝对优势，其分布式架构也更Match下一代存储的理念。Weaviate在使用案例上，有很多现成的例子，跟当前GPT前沿热门项目贴合比较紧秘，但在大规模生产环境使用场景中，还需要接受考验。Chroma是一个很轻量级的数据库，底层使用了clickhouse、duckdb等存储引擎