向量数据库

技术扩展

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【商标声明】

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

技术扩展
Embedding
DMC

版权所有：腾讯云计算（北京）有限责任公司

技术扩展
Embedding

最近更新时间：2024-12-27 17:22:53

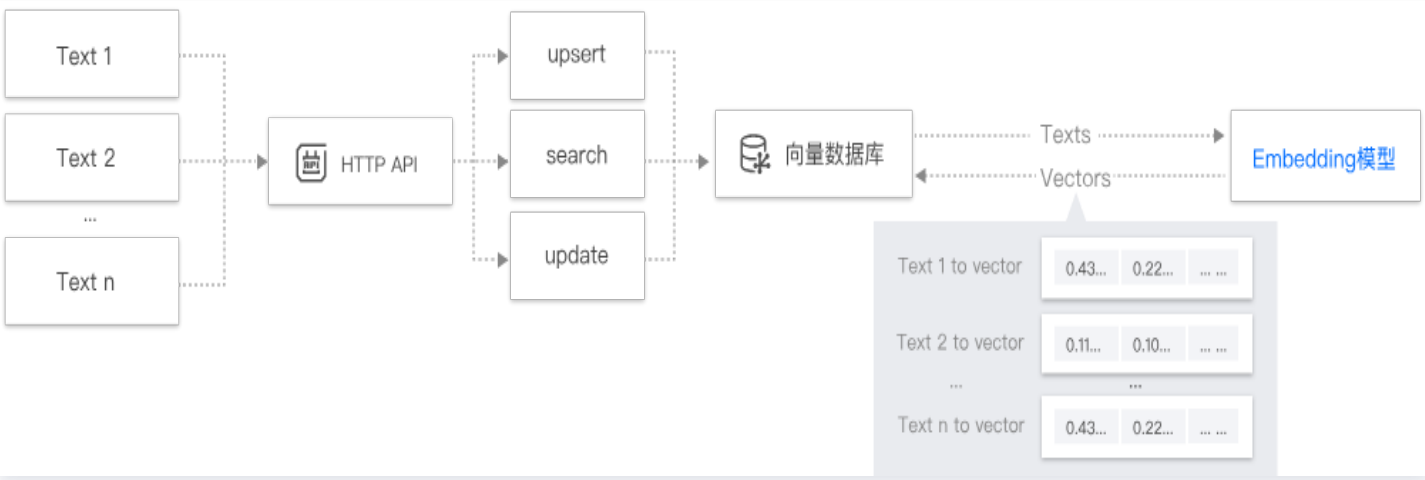
Embedding 功能是腾讯云向量数据库（Tencent Cloud VectorDB）提供将非结构化数据转换为向量数据的能力，目前已支持文本 Embedding 模型，能够覆盖多种主流语言的向量转换，包括但不限于中文、英文。开启 Embedding 功能并在创建 Collection 时配置模型，在插入、更新和相似性检索数据时直接传入原始文本，向量数据库会自动将原始文本进行转换，生成对应的向量数据后插入数据库或进行相似性计算，大幅提高业务接入效率。

快速接入

如果您想快速体验 Embedding 能力，腾讯云向量数据库（Tencent Cloud VectorDB）提供了 Python SDK 与 HTTP 的快速接入教程，请参见 连接并写入原始文本 。

Embedding 实现架构

腾讯云向量数据库（Tencent Cloud VectorDB）通过 HTTP API 将这些非结构化文本数据送入向量数据库，向量数据库将原始文本数据转交给 Embedding 模型进行向量化，再将转换后的向量数据以及原始文本一并存储在向量数据库中。其整个实现架构，如下图所示。



模型信息

腾讯云向量数据库（Tencent Cloud VectorDB）快速测试并分析来源于 Massive Text Embedding Benchmark（MTEB）上排名靠前的模型，选择出综合性能较好、适合不同应用场景的模型。当前，Embedding 功能支持的模型如下表所示。您可以依据数据集的语言类型、向量维度、以及综合性能得分选择合适的模型。

> **说明：**

如您正在选择模型阶段或认为当前模型效果不佳，可参考以下逻辑选择合适的 Embedding 模型：

1. 如果需要向量化的原始文本长度超过500字符数，且不希望做拆分，推荐选用 BAAI/bge-m3，该模型提供8K Token 的输入长度，可满足长文本向量化的需求，但请注意：即使该模型支持8K的 token 输入，但在实际的业务场景中，为了您的效果和性能最佳，建议每次 token 不超过2K。
2. 如果您的数据以中文文本为主，从效果方面考虑选择 bge-large-zh-v1.5；从性能、效果、成本方面综合考虑可选择 bge-base-zh-v1.5。使用 bge-base-zh-v1.5的存储空间对比 bge-large-zh-v1.5会降低约25%，且向量生成平均性能 bge-base-zh-v1.5对比 bge-large-zh-v1.5约有1倍优势。
3. 如果您的数据以混合语言的数据组成，推荐选用 BAAI/bge-m3，该支持100多种语言且在混合语言下具有较好的效果。

模型名	适用语言类型	Dimensions（维度）	最大 Token 数量
bge-large-zh-v1.5（推荐）	中文	1024	512
bge-base-zh-v1.5	中文	768	512
bge-base-zh	中文	768	512
bge-large-zh	中文	1024	512
BAAI/bge-m3	多语言	1024	8192
m3e-base	中文	768	512
text2vec-large-chinese	中文	1024	512
e5-large-v2	英文	1024	512
multilingual-e5-base	多语言	768	514

计费说明

腾讯云向量数据库（Tencent Cloud VectorDB）默认开通 Embedding 功能。在使用 Embedding 功能时，腾讯云向量数据库（Tencent Cloud VectorDB）将会根据输入文本的 **Token** 数量进行计费。具体计费信息，请参见 计费概述 。

说明：

> 在 Embedding 模型中，Token 是指文本数据处理的基本单元。通常在文本中，一个 Token 可以是一个字或词，也可以是一个标点符号。在将文本输入到 Embedding 模型中进行向量化时，文本数据会被切分成一系列的 Token 序列，每个 Token 都会依据在模型中预先建立的词汇表的映射关系与唯一的数字 ID 相关联，实现将所有 Token 映射到一个固定维度的向量空间，完成文本的向量化。

发布地域

当前 Embedding 功能支持地域包含：北京、上海、广州、南京、新加坡、硅谷。

相关 API

您需要在建表时，做相关配置，才能在写入、更新、检索数据直接写入原始文本，应用 Embedding 功能进行向量化。相关 API，如下表所示。

相关 API	含义	Embedding 信息
/collection/create	创建集合	指定 Embedding 模型，配置输入文本的字段名及其输出的向量字段。
/document/upsert	插入数据	插入原始文本信息，将原始文本直接向量化，将原始文本与向量数据一并存入数据库。
/document/update	更新数据	更新之前写入的文本信息，自动向量化后存入数据库。
/document/search	检索数据	检索数据时，可根据输入的文本信息，自动向量化并检索与其最相似的数据。

版权所有：腾讯云计算（北京）有限责任公司

DMC

最近更新时间：2024-10-31 17:43:02

数据库管理（Database Management Center，DMC）是一个高效，安全，可靠的数据库一站式管理平台，为用户提供库表级操作、实时监控、实例会话管理、SQL 窗口、数据管理为一体的数据库管理服务。腾讯云向量数据库支持通过 DMC 可视化管理数据库，帮助您更直观、更高效、更友好地操作数据库。更多信息，请参见 向量数据库管理查看首页 。

快速导入示例数据

一键导入测试数据，帮助您快速上手操作向量数据库并体验 Embedding 功能，显著提高操作效率。

可视化库表管理

管理实例下所有库表信息，便捷执行创建、清空、删除操作，减少了手动输入的时间和错误率，尤其是对于不熟悉命令行的用户来说，可视化界面更容易上手。

数据操作

在线执行数据插入、精确查询、相似度查询等操作，可视化直观地查看执行结果。

索引管理

查看集合下的索引类型、索引状态等信息，支持一键重建索引，高效运维管理。

版权所有：腾讯云计算（北京）有限责任公司

登录 DMC

使用腾讯云账号登录 向量数据库控制台，可一键登录 DMC。具体操作，请参见 登录 DMC 。

版权所有：腾讯云计算（北京）有限责任公司