向量数据库

关键特性

腾讯云

文档目录

关键特性
AI 套件

最近更新时间：2024-03-21 11:16:51

什么是 AI 套件?

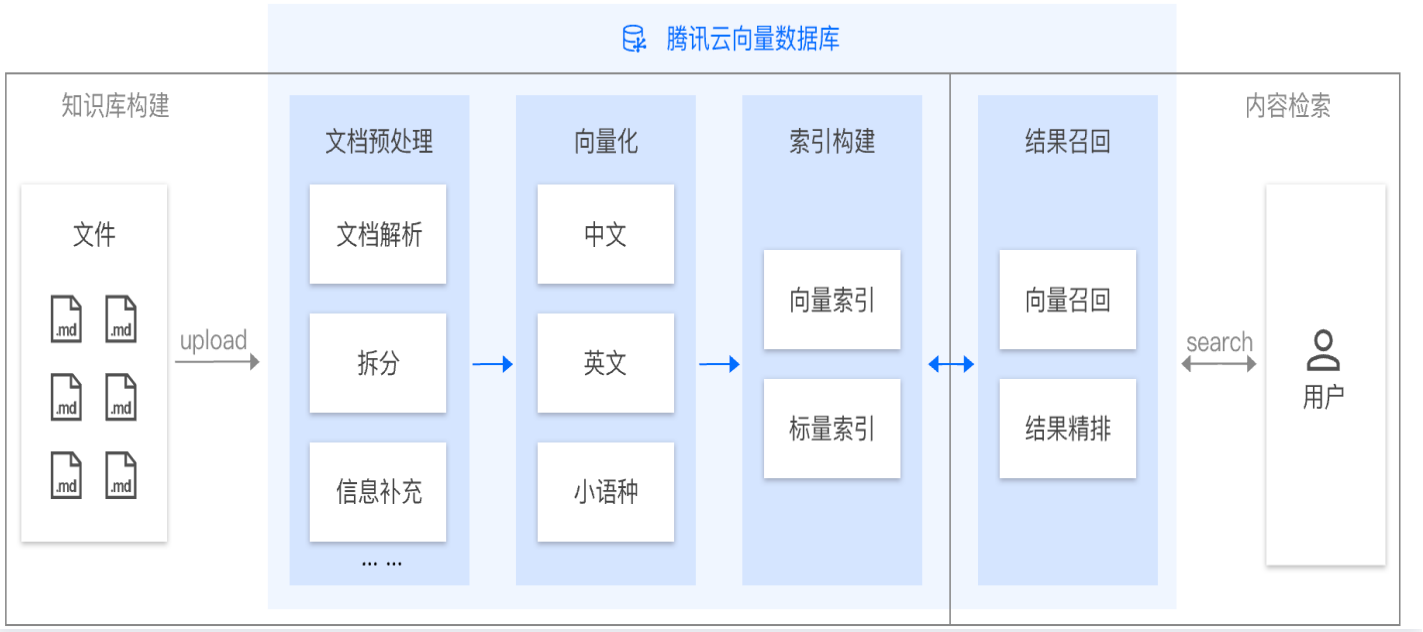
AI 套件是腾讯云向量数据库（Tencent Cloud VectorDB）提供的一站式文档检索解决方案，包含自动化文档解析、信息补充、向量化、内容检索等能力，并拥有丰富的可配置项，助力显著提升文档检索召回效果。用户仅需上传原始文档，数分钟内即可快速构建专属知识库，大幅提高知识接入效率。

快速接入

如需快速体验 AI 套件能力，请参见 使用 AI 套件快速上传文件并检索 。

设计思想

AI 套件检索方案提供完整的文档预处理和灵活的内容检索能力。用户只需上传 Markdown 格式的文档文件。腾讯云向量数据库将自动进行文本切分（Split）、信息补充、向量化（Embedding）和索引构建等一系列操作，完成知识库的建立。在进行检索时，会先基于切分后的内容进行相似度计算，并结合词（Words）向量进一步对检索结果进行精排，最终返回排名靠前的 Top K 条数据和其上下文内容。这种综合利用词级别做精排的检索方式，提供了更专业、更精确的内容检索体验。



基本概念

请先了解数据库设计的 逻辑结构 ，以便更好地理解 AI 套件相关的基本概念。

AI 类 Database

AI 类 Database 是专门用于 AI 套件上传和存储文件的向量数据库系统，可用于构建知识库。用户可以直接将文件上传至 AI 类 Database 下的 CollectionView 中，自动构建个性化的知识库。

> **说明：**
> - AI 类 Database 不支持直接对向量数据进行操作，已上传的文件不支持更新文件内容。
> - 为便于区别，腾讯云向量数据库将可直接操作向量数据的数据库称为 **Base 类 Database**。用户可以将向量数据上传至 Base 类 Database 中进行存储和管理，并可以直接对向量数据进行操作和处理。更多信息，请参见 Database 。

CollectionView

AI 类数据库文档组的集合视图，由多个 DocumentSet 组成，每个 DocumentSet 存储一组数据，对应一个文件数据。多个 DocumentSet 构成一个 CollectionView。

DocumentSet

相对 Document，DocumentSet 是 AI 类数据库中存储在 CollectionView 中的非结构化数据，是文件被拆分成多个 Document 的集合。每个DocumentSet 存储一组数据，对应一个文件，是 CollectionView 下存储文件的最小单元。

Metadata

文件元数据，指上传文件时所携带的文件元数据信息，可以包括文件的名称、作者、创建日期、文件类型等信息。所有元数据被自动解析为标量字段，以 `Key-Value` 格式存储。用户可根据元数据构建标量字段的 Filter 索引，以检索并管理文件。

Word

词语，是智能文档检索中最小的分割粒度，通常由一个或多个字符组成。在结果召回时，将对召回段落中所有 Words 进行相似性计算，以便于根据词向量进一步对检索结果做精排。

约束与限制

1. 当前支持导入数据库的**文件类型**包含： Markdown、PDF、Word、PPT，后续将逐步支持更多文件类型，请关注 产品动态 。

   > **说明：**
   > 2024-02-22 之前创建的实例，请 提交工单 申请升级实例版本，才能支持上传 PDF、Word、PPT。

2. 每次只能上传一个文件，Markdown 类型文件最大限制为 1MB，其余类型最大限制为 10MB。
3. 当前支持**地域**包含：北京、上海、广州、新加坡。

开发者工具

您可以通过 Python SDK 或 HTTP 的方式访问 AI 类 Database。具体信息，请参见下表。

类别	功能	Demo & API
Python SDK	将 AI 类 HTTP API 封装为 Python 函数或类	SDK AI Demo
HTTP	支持创建 AI 类数据库、集合、上传并检索文件	HTTP API

Embedding

最近更新时间：2024-06-06 09:30:11

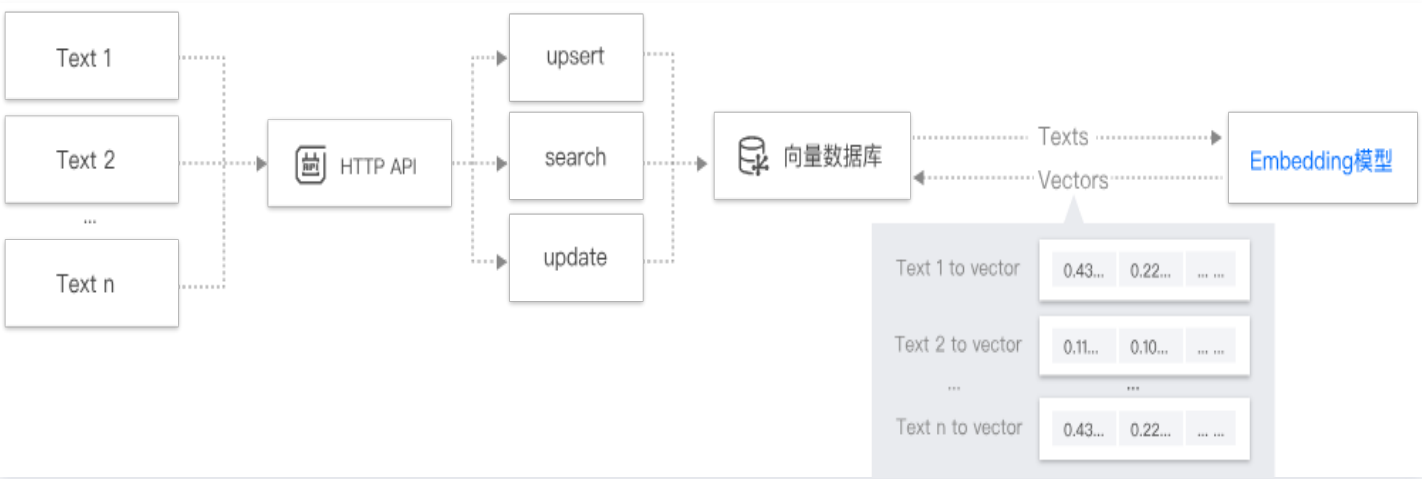
Embedding 功能是腾讯云向量数据库（Tencent Cloud VectorDB）提供将非结构化数据转换为向量数据的能力，目前已支持文本 Embedding 模型，能够覆盖多种主流语言的向量转换，包括但不限于中文、英文。开启 Embedding 功能并在创建 Collection 时配置模型，在插入、更新和相似性检索数据时直接传入原始文本，向量数据库会自动将原始文本进行转换，生成对应的向量数据后插入数据库或进行相似性计算，大幅提高业务接入效率。

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。

模型名	适用语言类型	Dimensions（维度）	最大 Token 数量	综合得分		
				Classification（分	Clustering（聚	Retrieval（检

				类）	类）	素）
bge-base-zh（推荐）	中文	768	512	67.06	47.64	69.53
m3e-base	中文	768	512	67.52	47.68	56.91
text2vec-large-chinese	中文	1024	512	60.66	30.02	41.94
e5-large-v2	英文	1024	512	75.24	44.49	50.56
multilingual-e5-base	多语言	768	514	65.35	40.68	40.68

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、新加坡。

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
	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

/collection/create		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。

DMC

最近更新时间：2023-09-27 22:07:21

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 数据库管理概述 。

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。

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 。