## Milvus使用手册

~~NC5主机，代码位置：~/pwork/milvus~~

轻量版：DS主机，代码位置：~/pwork/milvus

容器版：HZ1主机，代码位置：~/pwork/milvus

云端：zilliz

《向量数据库：大模型驱动的智能检索与应用(2025)》

1. 嵌入模型

下载bge嵌入模型，模型大小：1.5G

apt install git-lfs

git lfs install

git clone https://huggingface.co/BAAI/bge-large-zh-v1.5

使用bge嵌入模型

编写：bge.py

from langchain\_community.embeddings import HuggingFaceBgeEmbeddings

bge\_embeddings = HuggingFaceBgeEmbeddings(model\_name="/root/model/bge-large-zh-v1.5", model\_kwargs={'device': 'cuda:0'})

计算嵌入向量

v=bge\_embeddings.embed\_query("青蛙是食草动物")

1. 轻量Milvus

pip install pymilvus ；这个是LITE轻量版

编写：rag.py

读取faq文档，目录为：en，72条数据

本地轻量库：milvus\_demo.db

调用insert接口将数据插入向量数据库

调用search接口检索语义距离最近的N个向量

组成上下文，调用DeepSeek生成RAG应答

cuda版本：建库时长1.71秒，检索时长：0.025秒

cpu版本：建库时长22.6秒，检索时长：0.19秒

1. 容器Milvus

pip install pymilvus ；作为客户端

安装docker

curl https://raw.githubusercontent.com/milvus-io/milvus/master/scripts/standalone\_embed.sh -o standalone\_embed.sh

bash standalone\_embed.sh start

镜像：milvusdb/milvus，1.82G

端口：19530

编写：standalone.py

cpu版本：建库时长33.0秒，检索时长：0.54秒

1. 云端Milvus

访问点：https://in03-d2d7b3b4b519433.serverless.gcp-us-west1.cloud.zilliz.com

Token：2434ed4371b207dd8afce193e17f8113409b3ad82b220ff8c730db10a5df92dab668e1c8a04dadbfe6aad86e7d817ce139dcecb6

编写：zcreate.py

编写：zinsert.py

编写：zsearch.py

zilliz云端：建库时长30.3秒，检索时长：0.32秒