Vector Databases and Large Language Models

Sam Partee - Principal Engineer, Applied AI

Vector Embeddings

What are vector embeddings and how are they created?

Vectors

- Commonly represent unstructured data
  - Audio, text, images, etc
- Usually of high-dimension in the form of a **dense** embedding.
- Packed with information
- Easy to use API to create

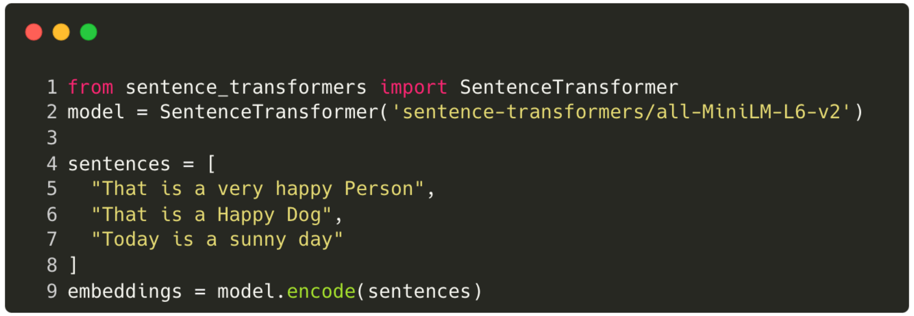
Hugging Face

OpenAI

cohere

Vector Embedding Creation

- Simple creation APIs
- Example with HuggingFace Sentence Transformer



```python
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v2')

sentences = [
  "That is a very happy Person",
  "That is a Happy Dog",
  "Today is a sunny day"
]
embeddings = model.encode(sentences)
```

redis

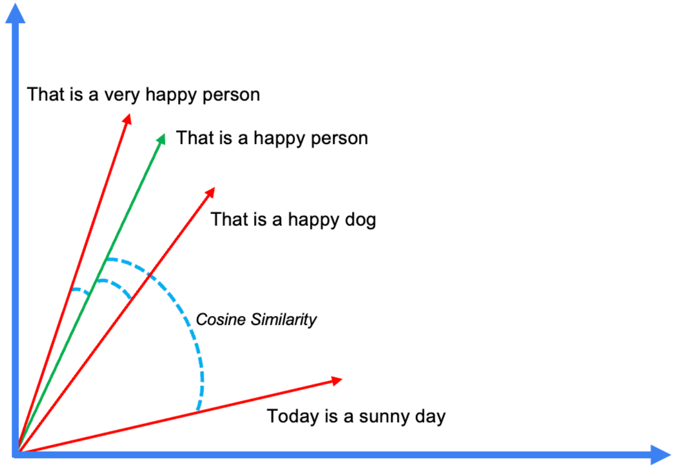
Vector Similarity Search

How are vector embeddings used for similarity search?

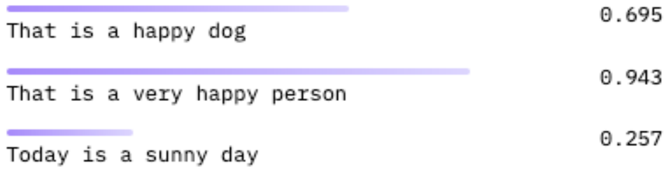
- 3 semantic vectors = **Search Space**
  - “today is a sunny day”
  - “that is a very happy person”
  - “that is a very happy dog”

- 1 Semantic vector = **Query**
  - “That is a happy person”

Goal: Find most similar vector to the query



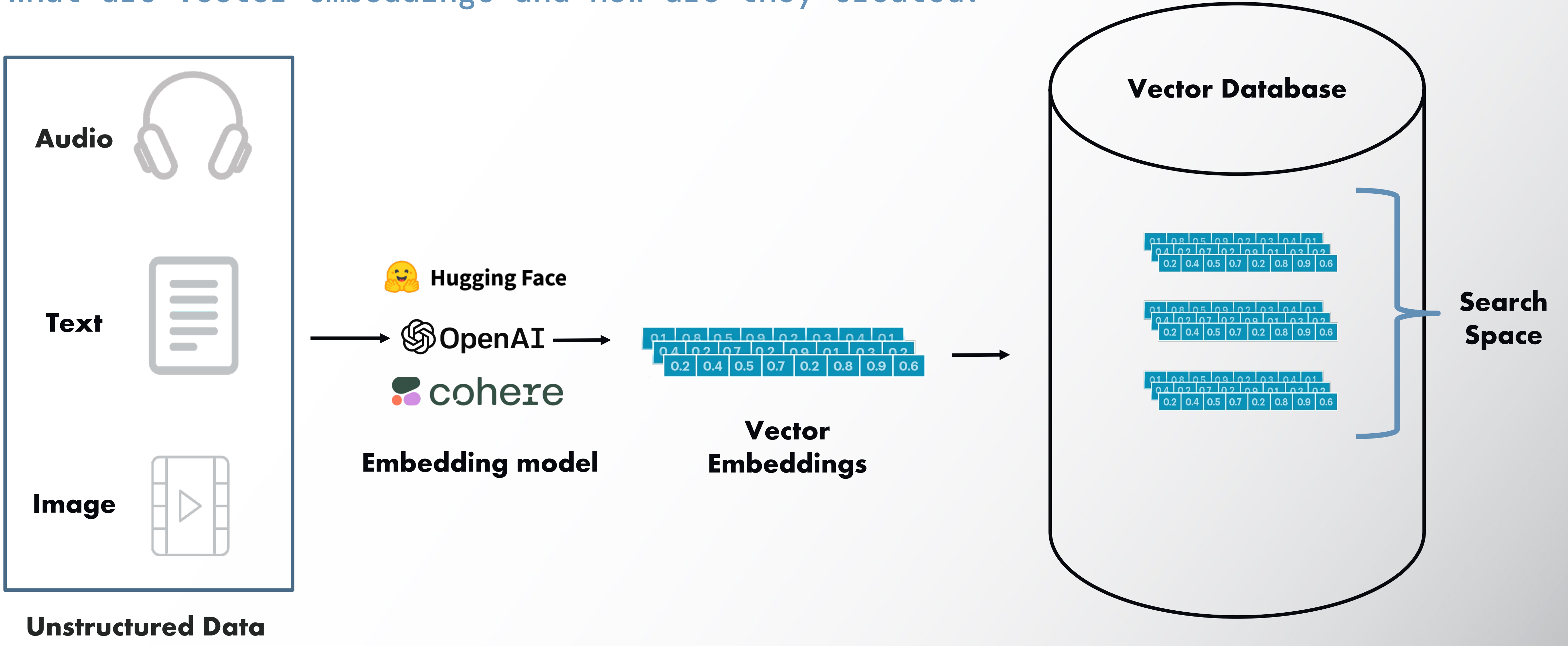
How? Calculate the distance (ex. Cosine Similarity)



https://mlops.community/vector-similarity-search-from-basics-to-production/

Vector Database

What are vector embeddings and how are they created?



Audio

Text

Image

Unstructured Data

Hugging Face

OpenAI

cohere

Embedding model

0.1 0.8 0.5 0.9 0.2 0.3 0.4 0.1

0.4 0.2 0.7 0.2 0.9 0.1 0.3 0.2

0.2 0.4 0.5 0.7 0.2 0.8 0.9 0.6

Vector Embeddings

Vector Database

Search Space

Redis as a Vector Database

Redis + RediSearch = Vector Database

Features

- Index types

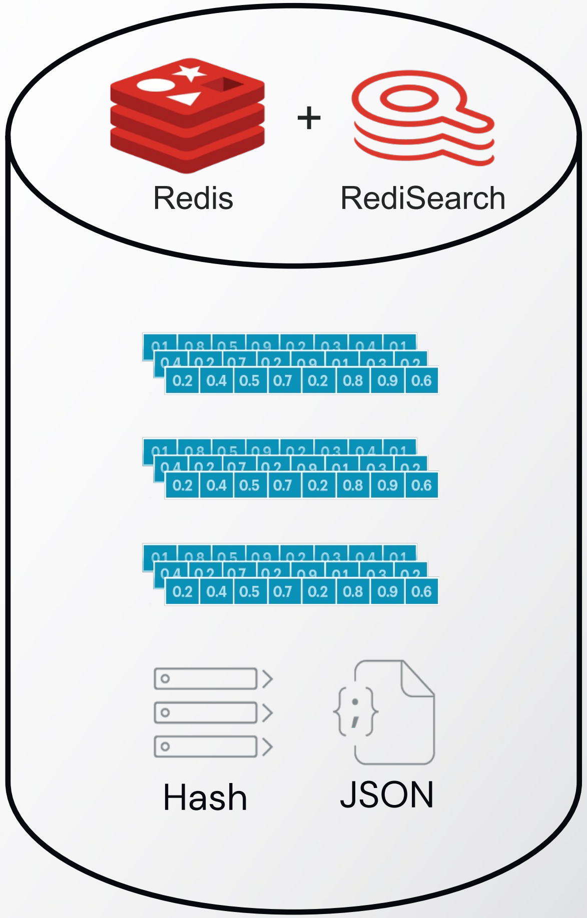
  Flat HNSW

- Distance metrics
  - L2, Cosine, Internal Product
- Integrations

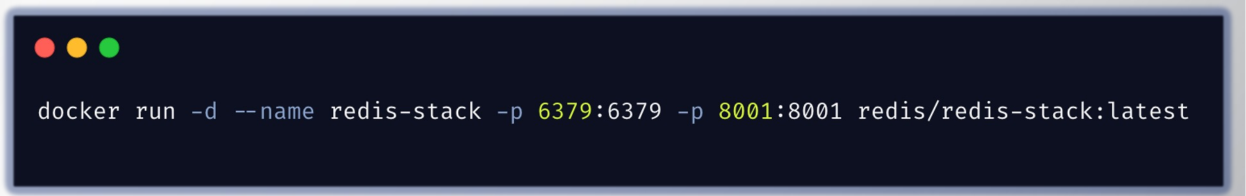
  Jina LangChain Relevance AI OpenAI Retrieval Plugin

- Coming Soon
  - **GPU index with NVIDIA RAFT integration**
  - LLamaIndex

  NVIDIA



Try it out! (w/ OpenAI Cookbook example)



```
docker run -d --name redis-stack -p 6379:6379 -p 8001:8001 redis/redis-stack:latest
```

LLM + Vector DB Use Cases

Because large was not large enough

redis

Vector Database

Use cases with Large Language Models

Context Retrieval

- Search for relevant sources of text from the “knowledge base”
- Provide as “context” to LLM

LLM “Memory”

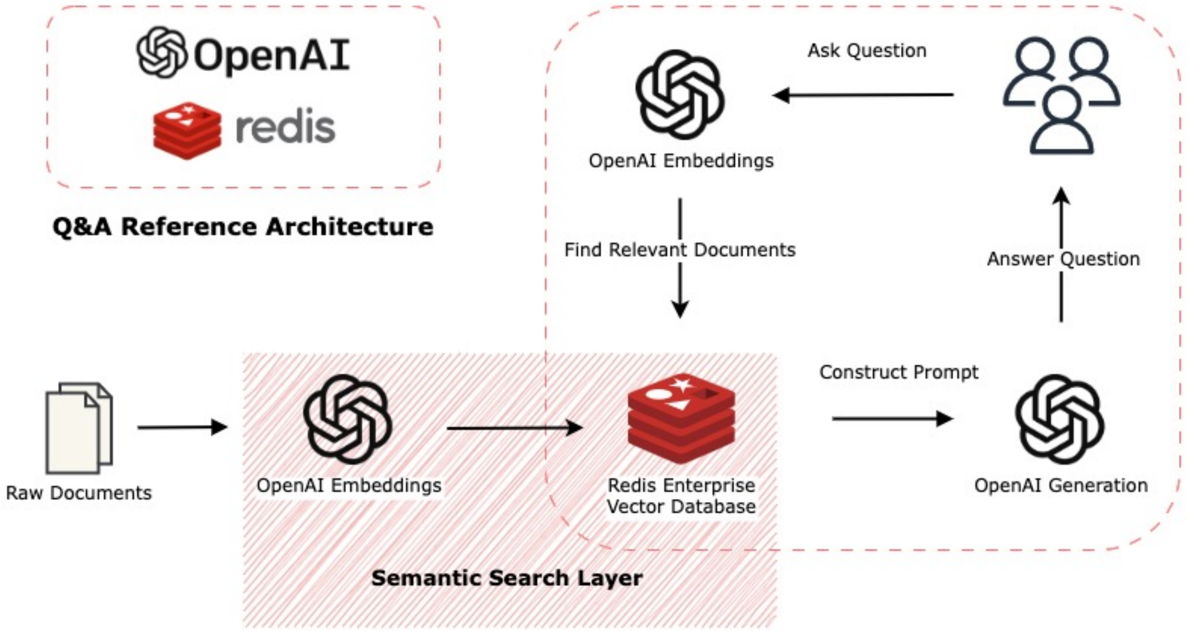
- Persist embedded conversation history
- Search for relevant conversation pieces as context for LLM

LLM Cache

- Search for semantically similar LLM prompts (inputs)
- Return cached responses

Context Retrieval

For Question and Answering systems



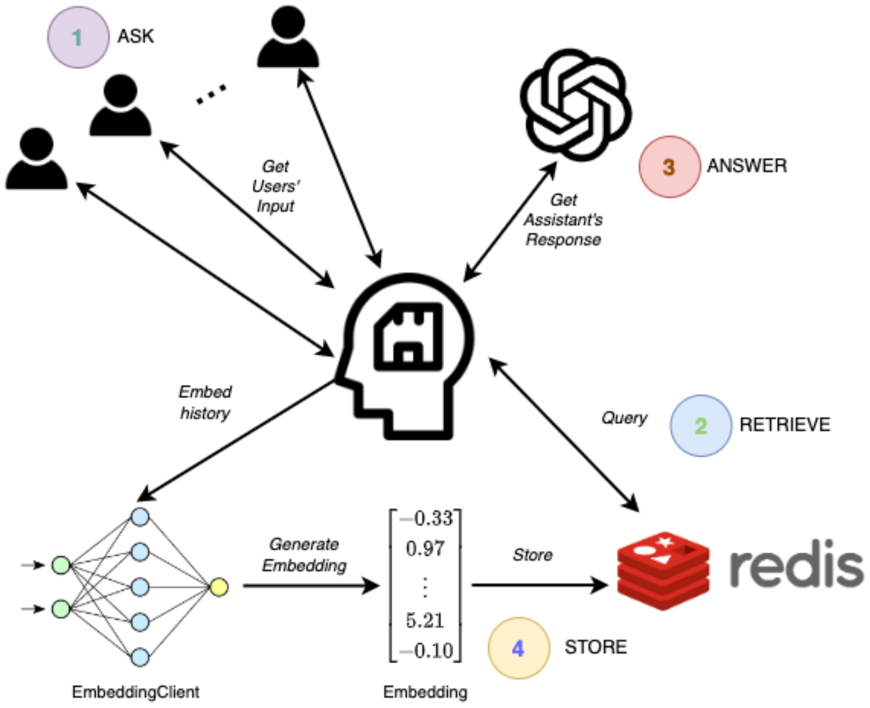
- Description
  - Vector database is used as an external knowledge base for the large language model.
  - Queries are used to detect similar information (context) within the knowledge base
- Benefits
  - **Cheaper and faster** than fine-tuning
  - **Real-time updates** to knowledge base
  - **Sensitive data** doesn't need to be used in model training or fine tuning
- Use Cases
  - Document discovery and analysis
  - Chatbots

Document QnA Example: https://github.com/RedisVentures/redis-openai-qna

Chatbot Example w/ Langchain: https://github.com/RedisVentures/redis-langchain-chatbot

Long-term Memory for LLMs

Contextual Memory without limits

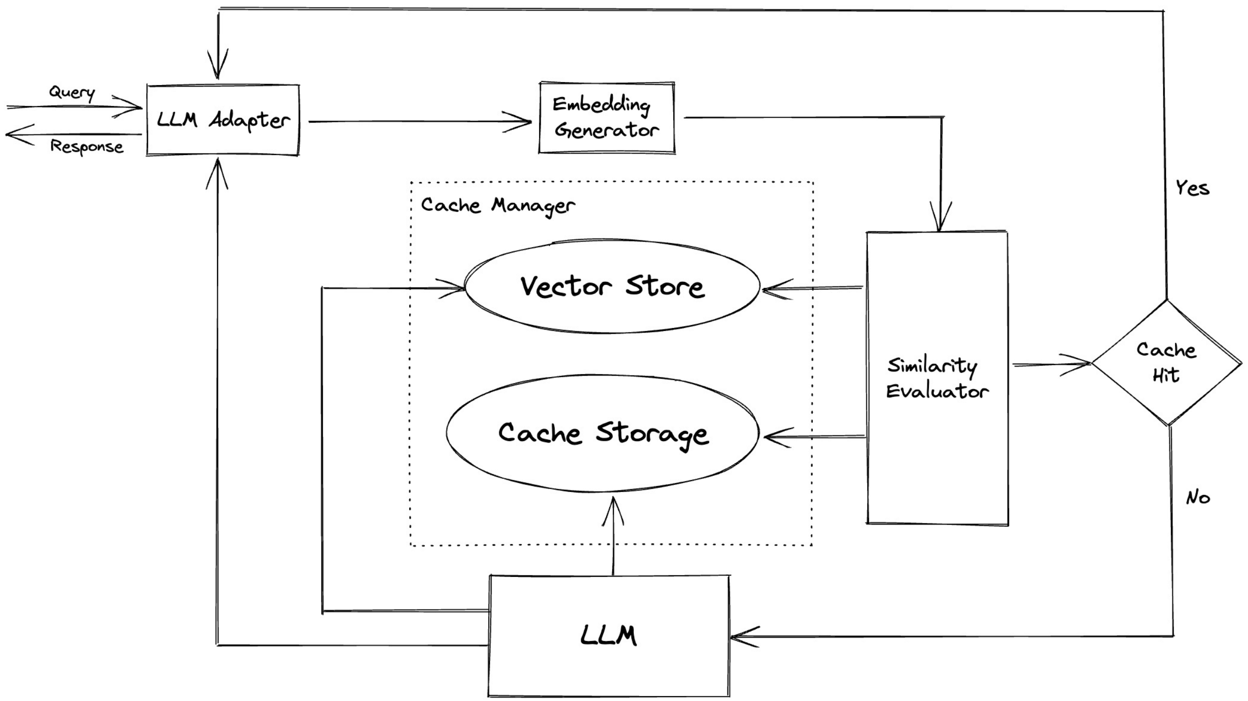


Repository: https://github.com/continuum-llms/chatgpt-memory

- Description
  - Theoretically infinite, contextual memory that encompasses multiple simultaneous sessions
  - Retrieves only last K messages relevant to the current message in the entire history.
- Benefits
  - Provides **solution to context length limitations** of large language models
  - Capable of **addressing topic changes** in conversation without context overflow
- Use Cases
  - Chatbots
  - Information retrieval
  - Continuous Knowledge Gathering

LLM Query Caching

Speed up Applications and Save Cost



https://github.com/zilliztech/GPTCache

- Description
  - Vector database used to cache similar queries and answers
  - Queries embedded and used as a cache lookup prior to LLM invocation
- Benefits
  - **Saves on computational and monetary cost** of calling LLM models.
  - Can **speed up applications** (LLMs are slow)
- Use Cases
  - Every single use case we've talked about that uses an LLM.

Build on Redis Vector Database

Context Retrieval

LLM “Memory”

LLM Cache

https://github.com/RedisVentures/

redis.com/solutions/use-cases/vector-database/

Contact: sam.partee@redis.com or @sampartee on Twitter

redis