Survey on Vector Databases

Andreas og Eivind
IN4120
18.11.2024

Vector databases

Allow the user to look up vectors similar to a given input vector

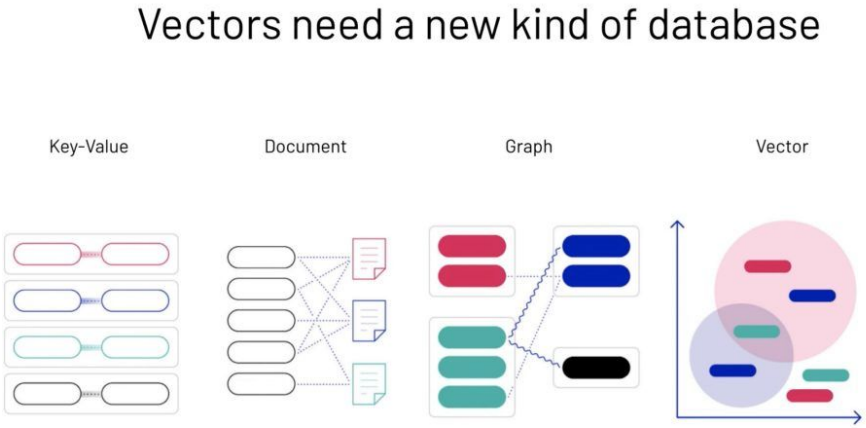
Essentially implement approximate nearest neighbor algorithms

This presentation:

- Why we need dedicated vector databases
- Storage and search methods used in vector databases for ANN
- Evaluating which method is right for a given use case
- Use cases for vector databases

Why do we need dedicated vector databases?

- Traditional database technologies
  - Scalars vs vectors
  - Dimensionality curse
  - No concept of semantic similarity
- Key question **how do you sort vectors in a meaningful way?**



Pinecone

Storage

- Sharding
  - Hash-based sharding
  - Range-based sharding
- Partitioning
  - List partitioning
  - Range partitioning
- Caching
  - Least Recently Used policy
- Replication
  - Leaderless replication
  - Leader-follower replication

ANN methods (some review from the lectures)

- Treebased
  - K-means Tree
- Hashbased (or clusters)
  - Deep Hashing
- Graph based
  - Navigable Small World

Performance evaluation

Speed recall tradeoff

- How quickly can we get the similar vectors?

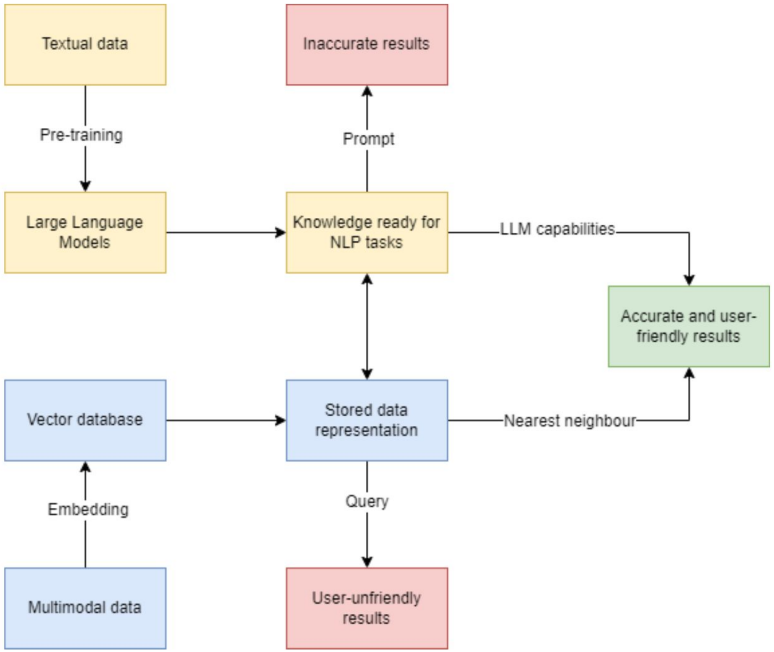
  vs.

- How well does the database get the vectors that are actually most similar to the query?

Vector databases and LLMs

Vector databases allow LLMs to store embeddings outside of itself. A couple of example use-cases:

- RAG
- Distributed training



Vector database products