Lecture #40 : RAG System Retrievers

- New opportunities
- for serious students only (free of cost ⇒ incubation center)
- Be grateful to Allah
- discussion about future projects
- Benefits of this field
- grab the opportunity & work for it.

[Proper communication & Tracking]

⇒ Good intentions — Dua — Mehnat

- just path -
- continuous effort
- different courses announcement + interactive sessions
-

Previously lectured:
- vector store
- local vs cloud
- Practically implemented

[improve step by step]

⇒
- Many New Courses
- proper interactive LMS

Retrieval

vector store → search engine

need to interact through ——

⇒ To interact with data (vector store)

interface ⇒ retrieval.

- enable us to interact with ingested data.

Retriever (finds similarity) → methods
1. Dot Product
2. cosine similarity
3. Maximum Margin Retriever

Query → vector form → go to vector store through retriever → find n similar vectors

demo of project references

Vector Store as Retriever

vector store use a vector store to retrieve documents..

light-weight wrapper around the vector store class to make it confirm retriever interface.

1. Semantic Search (cosine similarity calculate)
2. MMR (Maximum margin relevance) → filter out selected text
   - method used to avoid redundancy while retrieving relevant items to query.

vector store
1. vector
2. text associated
3. Meta-data

- will get associated text & chunk as output

Semantic Search :-

MMR :- (ensure relevancy and diversity)

All docs ⟶ fetch k most similar doc ⟶ k most diverse

other retrievals
- SVM
- TF/IDF (more focus on word) (most repeated & important)

Practical demo

ingestor
- Retrieval
  - query
  - perform retrieval (Similarity Search)

addressing specificity, working with metadata

filtering

must, should,

→ other retrieval example :- SVM, tf idf

qdrant

need analyzing power
not knowledge ≈ RAG

LLM

(Selecting brain for the system)

Why so many models?

- Total model
- Text-based models
- context window
- analyzing power, cost
- which model you use.

So mini model are more feasible

understand its complete image

- model
- cost
- accuracy