**Chapter – 1: Embeddings & Similarity Metrics – The Ultimate Guide**

**📘 1. Concept: What Are Similarity Metrics?**

**Similarity metrics** measure **how close two vectors (embeddings)** are to each other.  
They're used to compare texts, images, or multimodal data once represented numerically as embeddings.

For example:

* Find which document best matches a query
* Detect duplicate records
* Rank search results

**🧮 2. Common Similarity Metrics (with Examples)**

| **Metric** | **Range** | **Focus** | **Use Case** |
| --- | --- | --- | --- |
| **Cosine** | −1,1-1, 1−1,1 | Angle (direction only) | Semantic search, RAG |
| **L2 Distance** | [0, ∞) | Magnitude + direction | Clustering, kNN, anomalies |
| **Dot Product** | −∞,∞-∞, ∞−∞,∞ | Magnitude \* cosine(angle) | Used in neural attention |
| **Jaccard** | 0,10, 10,1 | Set similarity (discrete items) | Tags, categories, keywords |
| **Manhattan (L1)** | [0, ∞) | Sum of absolute differences | Simpler geometrical distance |

**✅ Example: Cosine Similarity**

Let:

python

CopyEdit

A = [1, 2]

B = [2, 3]

**Step-by-step:**

1. Compute dot product: A.B = 1\*2 + 2\*3 = 8
2. Compute magnitudes:
   * ||A|| = sqrt(1² + 2²) = sqrt(5)
   * ||B|| = sqrt(4 + 9) = sqrt(13)
3. Cosine similarity:
   * cos\_sim = 8 / (sqrt(5) \* sqrt(13)) ≈ 0.993

📌 **Interpretation**: Close to 1 = very similar direction = similar meaning

**✅ Example: L2 Distance**

python

CopyEdit

A = [1, 2], B = [2, 3]

L2 = sqrt((1-2)² + (2-3)²) = sqrt(2) ≈ 1.41

📌 **Interpretation**: Lower = more similar. Best when embeddings are **not normalized**.

**✅ Example: Dot Product**

python

CopyEdit

A = [1, 2], B = [2, 3]

Dot = 1\*2 + 2\*3 = 8

Used when **magnitude matters** (e.g., attention layers).

**🛠️ 3. Applications in AI Systems**

| **Use Case** | **How Similarity Helps** |
| --- | --- |
| **RAG Systems** | Retrieve semantically close docs |
| **Chatbots** | Thread context retrieval |
| **Duplicate Detection** | Compare text pairs with cosine similarity |
| **Product Recommendations** | Find similar items using L2 or cosine |
| **Multilingual Search** | Cross-lingual embeddings comparison |

**⚖️ 4. Pros & Cons of Similarity Metrics**

| **Metric** | **Pros** | **Cons** |
| --- | --- | --- |
| **Cosine** | Ignores magnitude, fast | Doesn’t work with zero vectors |
| **L2 Distance** | Works well when unnormalized | Sensitive to scale/magnitude |
| **Dot Product** | Differentiable, good in neural nets | Biased by vector length |
| **Jaccard** | Good for sets, interpretable | Not suitable for continuous vectors |

**🧭 5. When to Use Which Metric?**

| **Use Case** | **Recommended Metric** | **Why** |
| --- | --- | --- |
| Text Search / RAG | **Cosine Similarity** | Focuses on semantic direction |
| Clustering Raw Vectors | **L2 Distance** | Measures actual geometric distance |
| Deep Learning Attention | **Dot Product** | Magnitude & direction matters |
| Tag / Keyword Similarity | **Jaccard** | Set overlap-based |
| Sparse or Zero vectors | **L2 or Jaccard** | Cosine fails if any vector is zero |

**🌌 6. Embeddings: Deeper Concepts You Must Know**

| **Topic** | **Key Insight** |
| --- | --- |
| **Chunking Strategy** | Affects how relevant your search is |
| **Dimensionality** | Higher isn't always better — use model based on task |
| **Normalization** | Essential for cosine similarity |
| **Vector Stores** | Use Qdrant, FAISS, Chroma, Pinecone to scale vector search |
| **Hybrid Search** | Combine keyword + semantic (vector) search |
| **Instruction-Tuned Models** | Like instructor-xl for task-specific embeddings |
| **Thread-aware Retrieval** | For chatbots, embed query + past thread for relevance |
| **Evaluation Metrics** | Precision@k, MRR, Recall — measure how good your RAG system is |

**🔍 Example Code (Cosine Similarity on Documents)**

python

CopyEdit

from sklearn.metrics.pairwise import cosine\_similarity

from sentence\_transformers import SentenceTransformer

model = SentenceTransformer('all-MiniLM-L6-v2')

docs = [

"India won the 2011 World Cup",

"India has a nuclear power policy",

"India is a democratic country"

]

query = "Who won the 2011 ODI World Cup?"

# Embed

doc\_embeddings = model.encode(docs, normalize\_embeddings=True)

query\_embedding = model.encode([query], normalize\_embeddings=True)

# Compute similarity

scores = cosine\_similarity(query\_embedding, doc\_embeddings)

print(scores)

**Output (example):**

lua

CopyEdit

[[0.81, 0.22, 0.13]]

📌 Highest score → Most relevant doc.

**🧠 Bonus: Chunking Best Practices**

* Prefer RecursiveCharacterTextSplitter or sentence-aware splitters
* Ideal chunk size: **300–500 tokens**, with 10–50 token overlap
* For large docs: use **sliding window + summarization** or **refinement chains**

**✅ Summary Table: Embedding Power Moves**

| **Strategy** | **Description** |
| --- | --- |
| Use cosine for RAG | Standard and effective |
| Normalize vectors | Mandatory for cosine similarity |
| Smart chunking | Improves context relevance |
| Use domain-specific models | Improves retrieval accuracy |
| Hybrid search | Boosts recall with metadata filters |
| Eval with Recall@k | Optimizes search quality |
| Use vector stores | Efficient large-scale search |

## Chapter – 2: Query Search & Retrieval in FAISS Vector Store

### 🔰 Overview

In Retrieval-Augmented Generation (RAG), once all corpus texts are converted to vectors and stored in a vector store like **FAISS**, the **similarity search process** kicks in at query time to fetch the most relevant context.

### ⚙️ Components Involved

1. **FAISS Index**  
   Stores dense embedding vectors with an internal positional index (e.g., 0, 1, 2…).
   * Example:  
     0 → [0.12, 0.67, 0.89]  
     1 → [0.45, 0.23, 0.91]
2. **Docstore**  
   Stores actual documents mapped to UUIDs.
   * Example:  
     'uuid\_123' → "India won T20 World Cup in 2007"
3. **Index-to-Docstore-ID Mapping**  
   Maps FAISS index positions to docstore UUIDs.
   * Example:  
     0 → 'uuid\_123'  
     1 → 'uuid\_456'

### 🧠 Query Search Workflow

#### 1. **Convert User Query to Embedding**

* Use the **same embedding model** used during corpus vector creation.
* Converts query text to a **dense vector**.

#### 2. **Search in FAISS Index**

* Pass query vector to faiss\_index.search(query\_vector, top\_k).
* FAISS computes similarity (e.g., cosine, L2) with **all stored vectors**.
* Returns:
  + Top k vector indices
  + Corresponding similarity scores

#### 3. **Map FAISS Index to UUIDs**

* Use index\_to\_docstore\_id to convert returned indices to UUIDs.

#### 4. **Fetch Documents**

* Use UUIDs to get actual texts from docstore.

### ✅ Final Output

* You receive **Top-K most similar documents** to the user query.
* These can be fed as context to an LLM for further reasoning or answer generation.