

Embeddings

Embeddings convert text into high-dimensional numerical vectors that capture semantic meaning. They are the foundation of retrieval in RAG, enabling similarity search between queries and stored documents.

01

Embeddings

❑ What it is

- Converts text (queries, documents) into **dense vector representations** (embeddings) that capture semantic meaning.

❑ Why it's important

- Enables the system to understand **semantic similarity** between user queries and stored documents.

❑ Examples

- OpenAI (text-embedding-ada-002)
- Hugging Face models
- Cohere, Google's BERT/USE

Vector Embeddings

$$\begin{array}{lcl} \text{Apple} & \rightarrow & \begin{bmatrix} 0.5 & 0.6 & 0 & 0.1 & 0.4 & \dots & 0.4 & 0 \end{bmatrix} \\ \text{Man} & \rightarrow & \begin{bmatrix} 0.1 & 0.3 & 0.4 & 0 & 0.5 & \dots & 0.5 & 1 \end{bmatrix} \\ \text{Computer} & \rightarrow & \begin{bmatrix} 0.4 & 0.5 & 0.4 & 0.1 & 0 & \dots & 0 & 0 \end{bmatrix} \end{array}$$

02