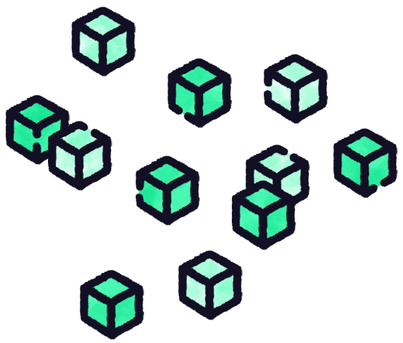
Embedding Vectors for Retrieval and Generation

Cătălin Tudose

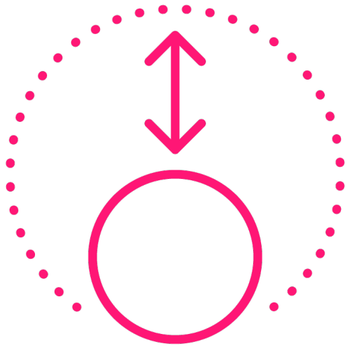
Java Champion, PhD in Computer Science, Java and Web Technologies Expert

www.catalintudose.com | www.linkedin.com/in/catalin-tudose-847667a1

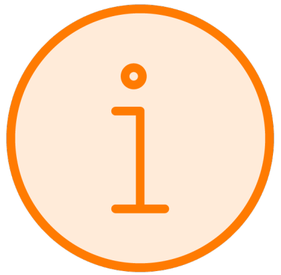
The Role of Embeddings



Unstructured data

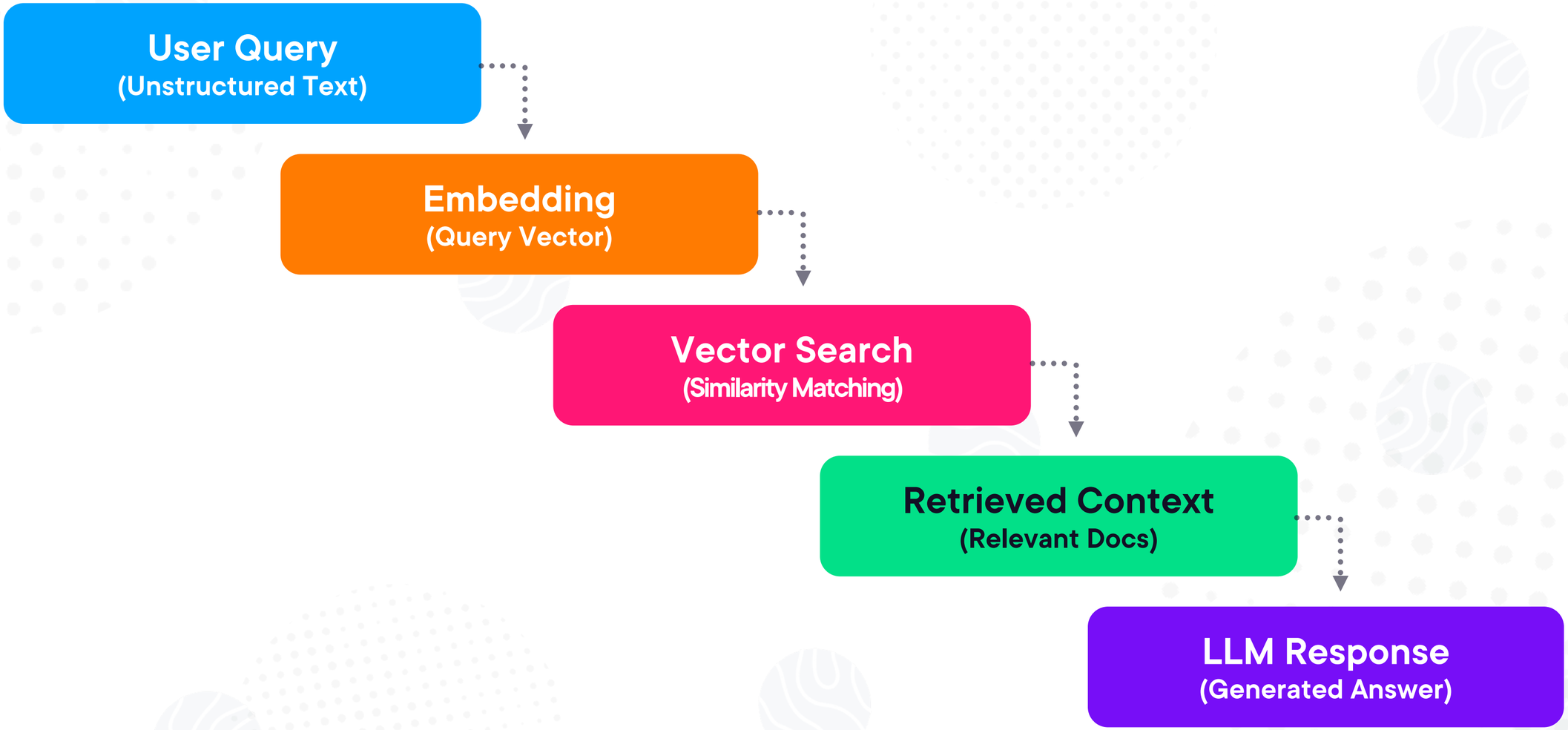


Text to numerical vectors



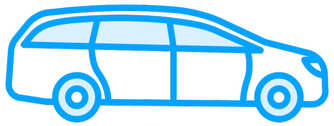
Semantic information

The RAG Retrieval Pipeline



Traditional Search Engines

Car



Automobile

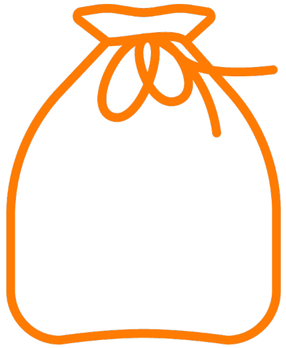
Embeddings

Meaning

Semantic Search

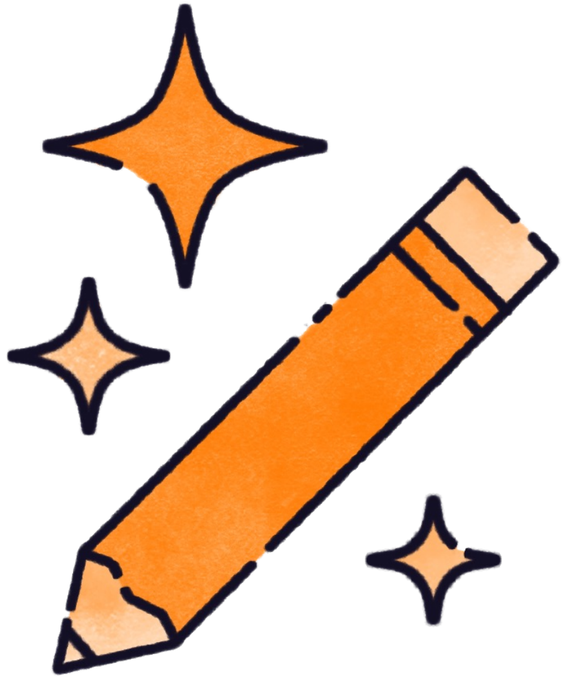
Search query

Compare embedding

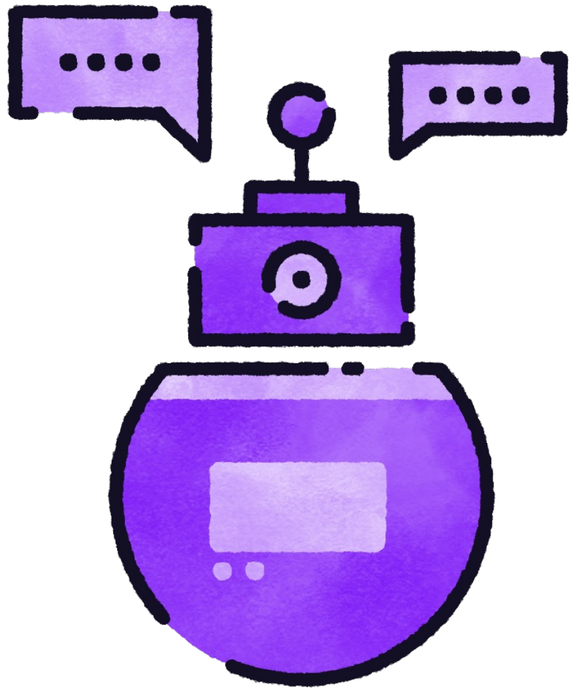


Semantically similar content

Semantic Search



AI writing tools



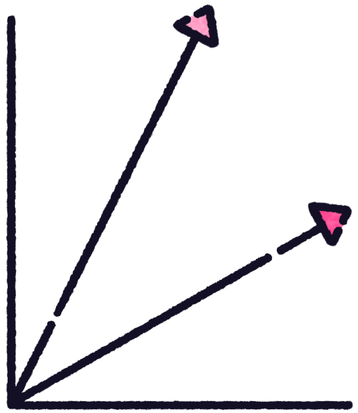
Chatbots for content creation

Similarity Matching

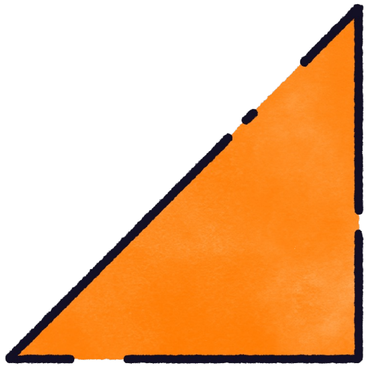
Distance between vectors encodes similarity

Similar meanings, closer embeddings

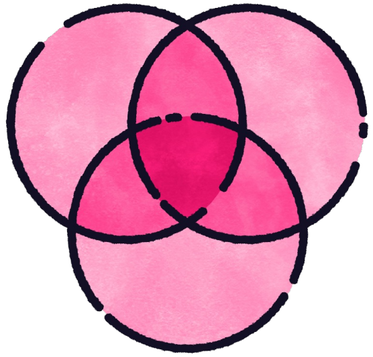
Similarity Matching



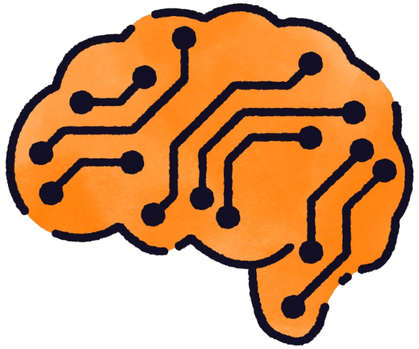
Cosine similarity



Euclidean distance

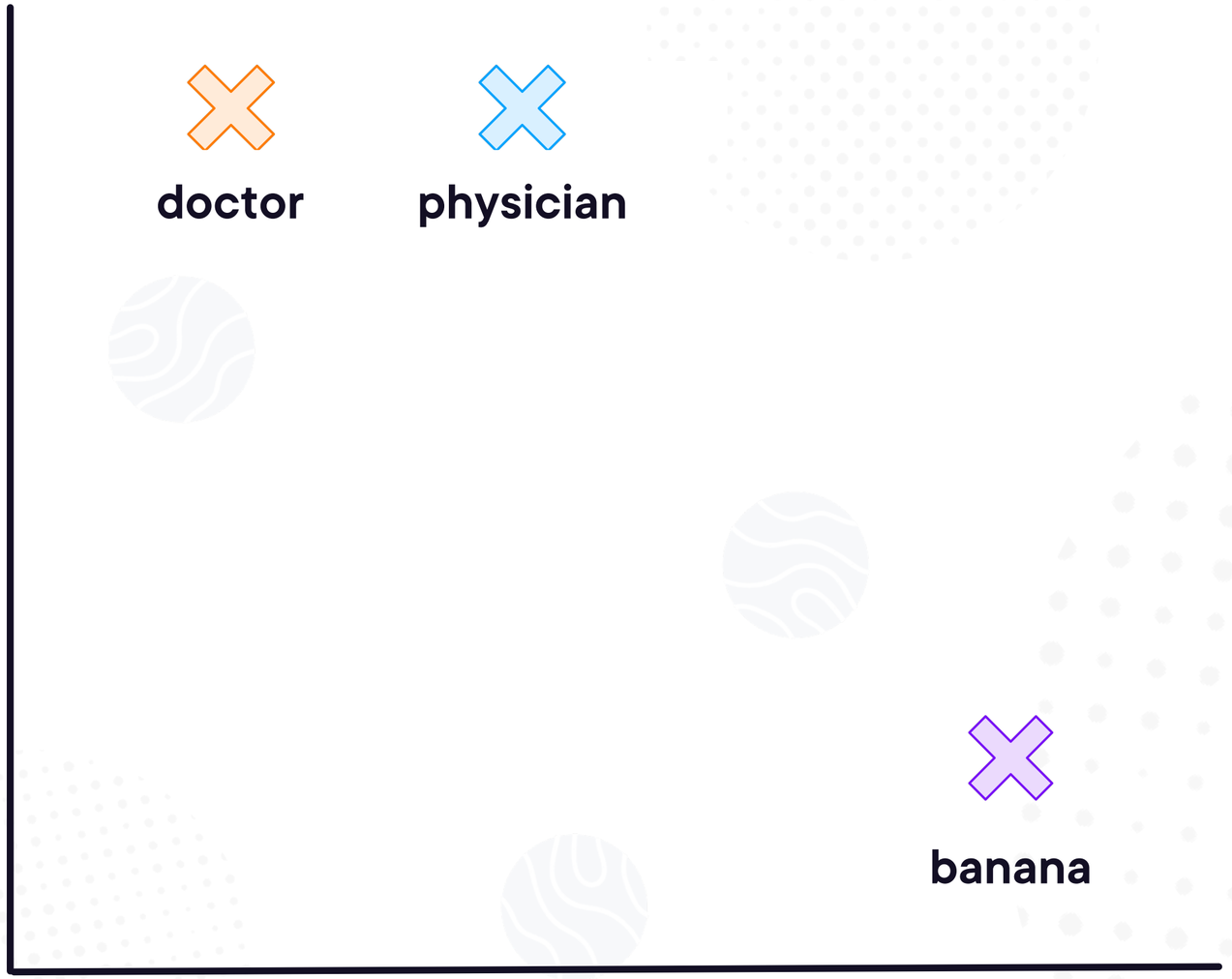


Keyword overlap

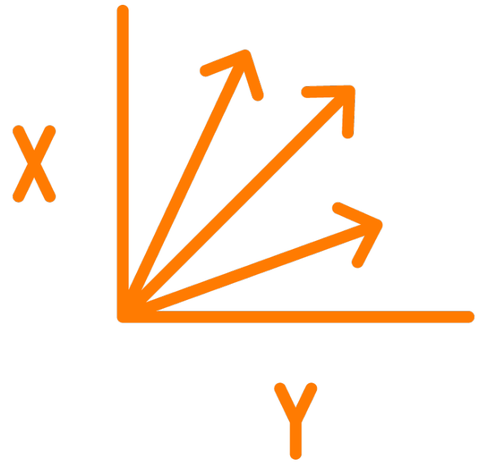


Conceptual similarity

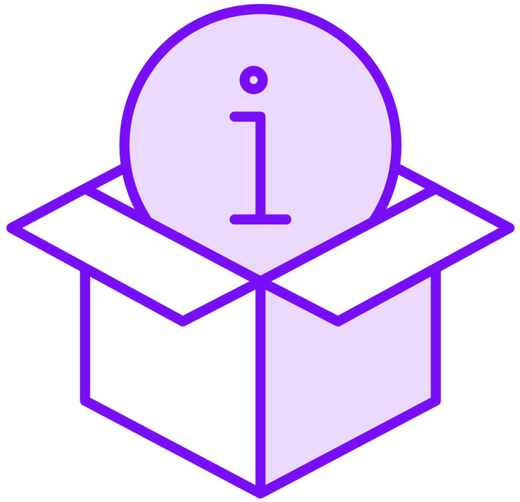
Similarity Matching



Relevance Ranking



Higher vector similarity



Most contextually information first

Relevance Ranking

Climate
change effects

Global warming
impacts on agriculture

Political debates on
energy policy

The Three Steps

Semantic search

Similarity matching

Relevance ranking

Embedding and Real-world Challenges

Vector Quality

Usefulness depends on quality

Poorly trained embeddings fail to capture semantic relationships

Domain mismatch

Granularity matters

Benchmarking against real tasks

Storage and Scalability

Storage requirements grow quickly

Specialized vector databases

Approximate Nearest Neighbor (ANN)

Retrieval latency

Updating Embeddings

Model updated, stored embeddings no longer compatible

Dynamic data requires periodic re-embedding

Versioning strategies

Quality, scalability, maintainability

Vector Drift

Embedding model is updated or fine-tuned

Queries no longer match previously indexed documents

Problematic in long-lived systems

Large-Scale Storage

Millions or billions of embeddings

Storage footprint grows rapidly

Challenge of efficient retrieval

Specialized vector databases

Approximate Nearest Neighbor (ANN)

Updating Outdated Knowledge

Remain current

Embeddings are static

Outdated embeddings = stale information

Re-embedding new documents

Operational challenges