Semantic search: how to query a database using natural language

Enrico Zimuel, *Tech Lead & Principal Software Engineer*

March 26, 2025 - CloudConf, Turin (Italy)

Agenda

- What is semantic search?
- Vector and embedding
- k-nearest neighbor (kNN) algorithm
- Vector database: Elasticsearch
- Example in Python

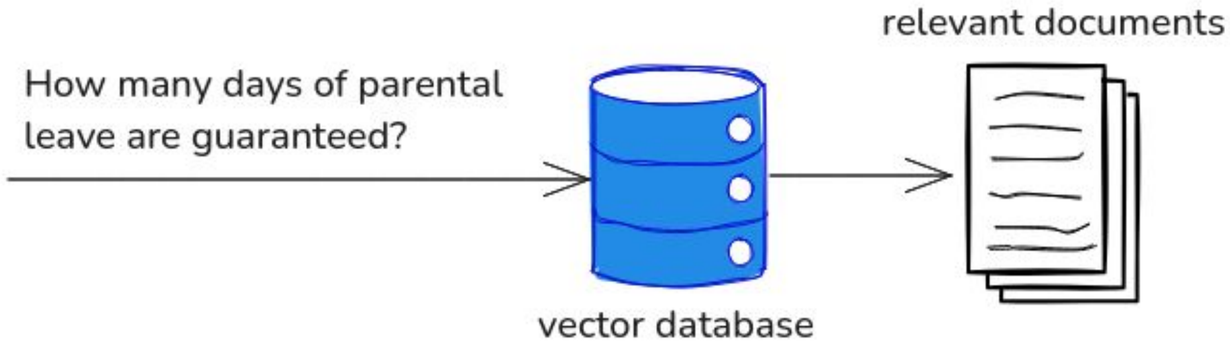
elastic

What is semantic search?

- **Semantic search** is a new approach of searching using the meaning of sentences
- Instead of searching for term frequency (TF-IDF) we search for semantic similarity between words or sentences
- Es. imagine a company database with HR documents, a typical semantic search query can be:
  How many days of parental leave are guaranteed?

Result of a semantic search

- The result of a query is a **set of documents** ordered by their semantic similarity to the query
- These **relevant documents** generally include all the information needed to answer the query



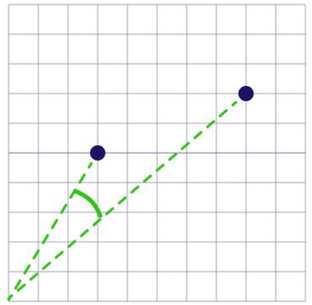
elastic

Embedding and vectors

- In semantic search the documents (text, image, sound, etc) are converted in **vectors** (list of numbers)
- The search operation compares the **similarity** (distance) between vectors
- The result is a list of documents whose vector representations are close to the query’s vector

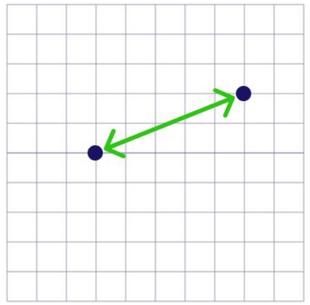
Similarity metrics

- Two vectors are (semantically) similar if they are close to each other
- There are many ways to measure the similarity



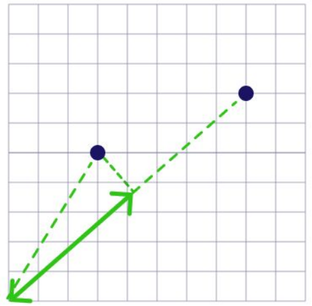
Cosine Distance

$$1 - \frac{A \cdot B}{||A|| \quad ||B||}$$



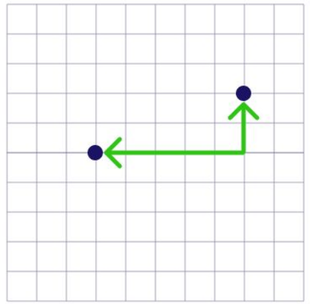
Squared Euclidean (L2 Squared)

$$\sum_{i=1}^{n} (x_i - y_i)^2$$



Dot Product

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$



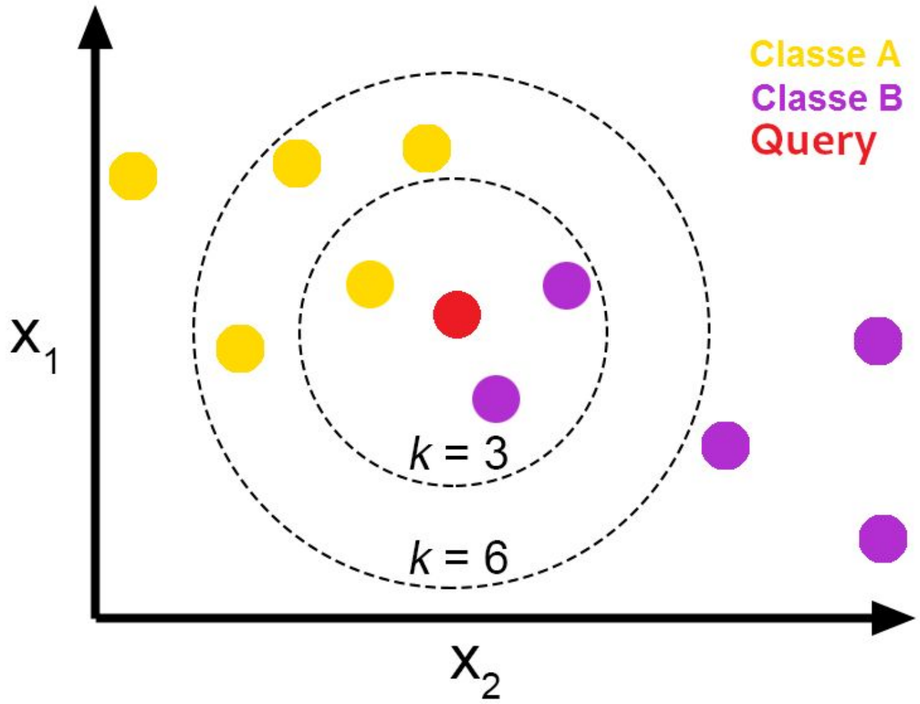
Manhattan (L1)

$$\sum_{i=1}^{n} |x_i - y_i|$$

elastic

kNN search

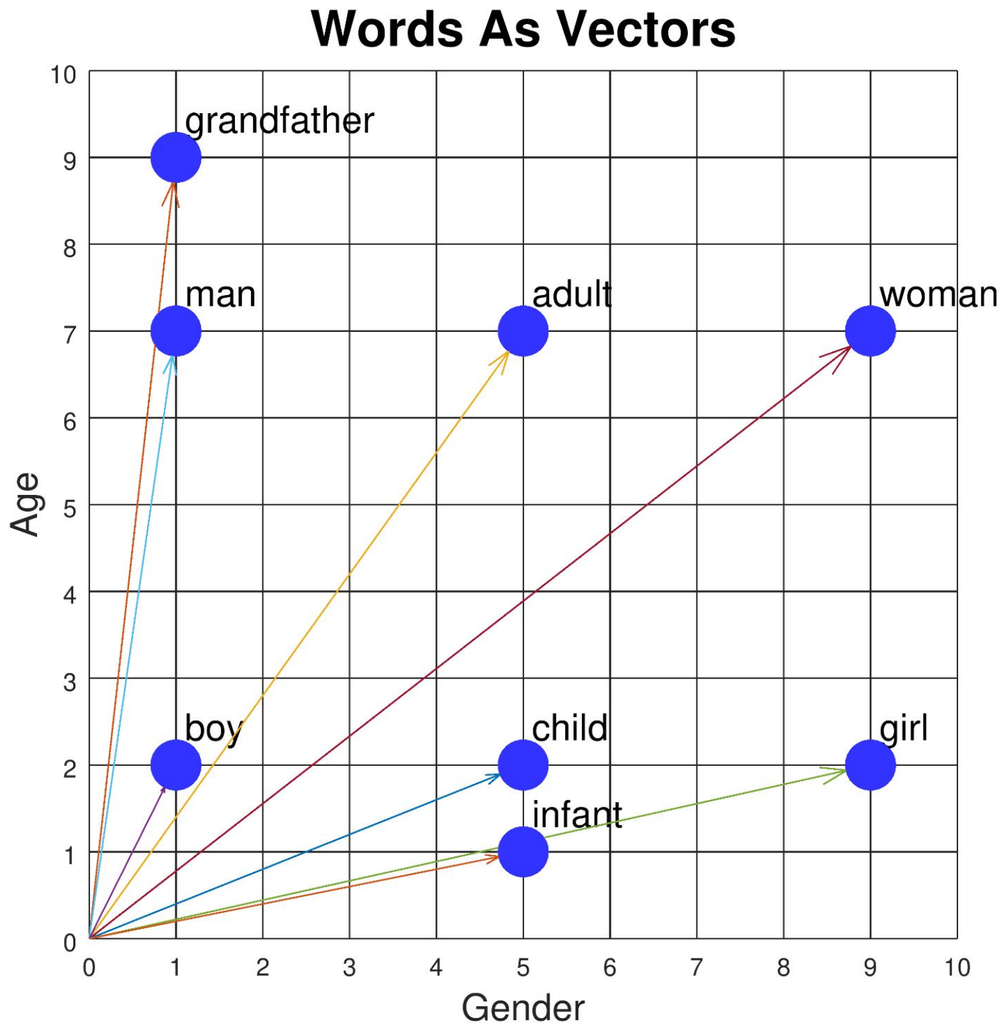
- **k-nearest neighbor** (kNN) search finds the k nearest vectors to a query vector, as measured by a similarity metric



elastic

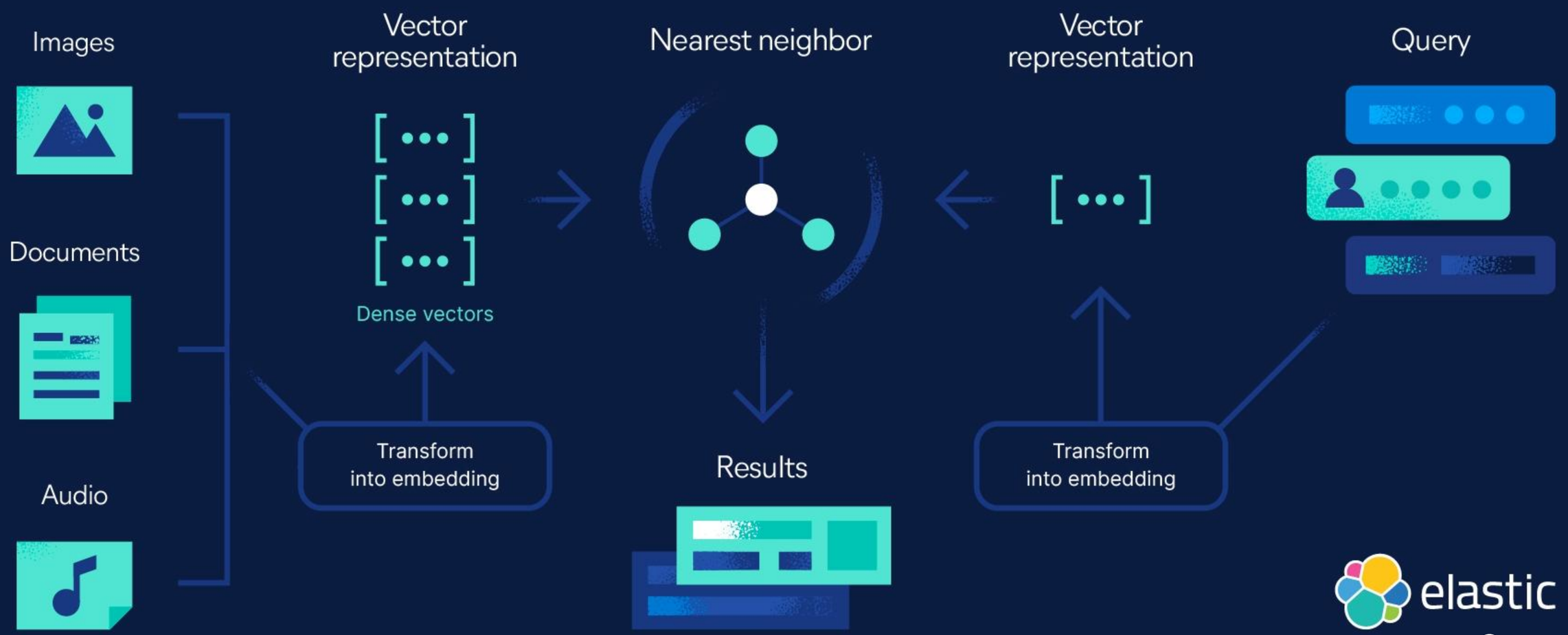
Embedding

- **Embedding** is the operation that transforms a document into a vector
- Typically, we use embedding models with **deep neural network** (e.g. LLM)
- These models distribute similar information across multidimensional spaces
- https://projector.tensorflow.org/



elastic

Vector search data workflow



elastic

elastic

Elasticsearch

- Elasticsearch is an Open Source, Distributed, RESTful Search Engine and Vector Database
- Vector database:
  - dense and sparse vectors
  - kNN (Approximate, Exact)
  - Reciprocal Rank Fusion (RRF)
  - semantic_text
  - Inference APIs
  - ELSER and E5 models (more to come)
- Try locally:
  - **curl -fsSL https://elastic.co/start-local | sh**

DEMO

https://github.com/ezimuel/semantic-search-examples

elastic

References

- kNN search in Elasticsearch, official documentation
- Elasticsearch as vector database, Elastic Search Labs
- Elasticsearch search relevance, Elastic Search Labs
- Carlos Delgado, How to choose between exact and approximate kNN search in Elasticsearch, Elastic Search Labs
- K-Nearest Neighbors Algorithm: Classification and Regression Star, History of Data Science, Accessed: 10/23/2023
- Cormack, Clarke, Buttcher, Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods, SIGIR '09: Proceedings of the 32nd international ACM
- E.Zimuel, Retrieval-Augmented Generation for talking with your private data using LLM, AI Heroes 2023 conference, Turin (Italy)

Thanks!

More information: www.elastic.co

Contact information: enrico.zimuel (at) elastic.co

elastic