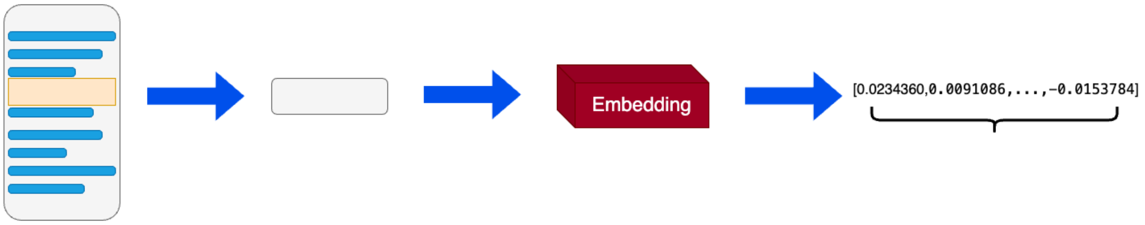
1142 words | 3 minutes

# Word vectors and vector knowledge base

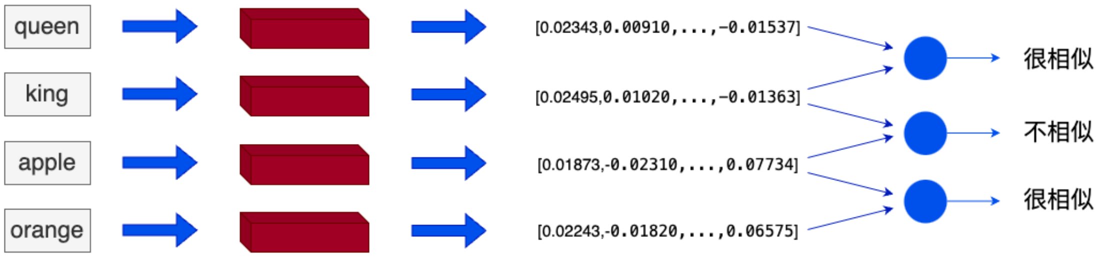
## 1. Word vector

### 1. What is word vector



In machine learning and natural language processing (NLP), embeddings are a technique for converting unstructured data, such as words, sentences, or entire documents, into real number vectors that can be better understood and processed by computers.

The main idea behind embeddings is that similar or related objects should be close to each other in the embedding space.



For example, we can use word embeddings to represent text data. In word embeddings, each word is converted into a vector that captures the semantic information of the word. For example, the words "king" and "queen" will be very close in the embedding space because they have similar meanings. And "apple" and "orange" will also be close because they are both fruits. On the other hand, the words "king" and "apple" will be far apart in the embedding space because they have different meanings.

### 2. Advantages of word vectors

In terms of RAG (Retrieval Augmented Generation), word vectors have two main advantages:

- Word vectors are more suitable for retrieval than text. When we search in a database, if the database stores text, we mainly find relatively matching data by searching for keywords (lexical search) and other methods. The degree of matching depends on the number of keywords or whether the query is completely matched. However, word vectors contain the semantic information of the original text. We can directly obtain the similarity between the question and the data at the semantic level by calculating the dot product, cosine distance, Euclidean distance and other indicators between the question and the data in the database.
- Word vectors have stronger comprehensive information capabilities than other media. When traditional databases store multiple media such as text, sound, images, and videos, it is difficult to build associations and cross-modal query methods for the above multiple media; however, word vectors can map multiple data into a unified vector form through a variety of vector models.

### 3. General method of constructing word vectors

When building a RAG system, we can often construct word vectors by using an embedding model. We can choose:

- Use Embedding APIs from various companies;
- Use the embedding model locally to build word vectors from the data.

## 2. Vector Database

### 1. What is a vector database?

Vector database is a solution for efficient calculation and management of large amounts of vector data. Vector database is a database system specifically used for storing and retrieving vector data (embedding). It is different from traditional relational model-based databases, and it mainly focuses on the characteristics and similarities of vector data.

In a vector database, data is represented as vectors, each of which represents a data item. These vectors can be numbers, text, images, or other types of data. Vector databases use efficient indexing and query algorithms to speed up the storage and retrieval process of vector data.

### 2. Principles and core advantages of vector database

The data in the vector database uses vectors as the basic unit to store, process and retrieve vectors. The vector database obtains the similarity with the target vector by calculating the cosine distance, dot product, etc. When processing large or even massive amounts of vector data, the efficiency of the vector database index and query algorithm is significantly higher than that of traditional databases.

## 3. Mainstream vector databases

- Chroma : It is a lightweight vector database with rich functions and simple API. It has the advantages of simplicity, ease of use and lightweight, but its functions are relatively simple and it does not support GPU acceleration, so it is suitable for beginners.
- Weaviate : is an open source vector database. In addition to similarity search and maximum marginal relevance (MMR) search, it also supports hybrid search combining multiple search algorithms (based on lexical search, vector search), thereby improving the relevance and accuracy of search results.
- Qdrant : Qdrant is developed in Rust language, with extremely high retrieval efficiency and RPS (Requests Per Second), and supports three deployment modes: local operation, deployment on local server and Qdrant cloud. And data can be reused by setting different keys for page content and metadata.

Next Chapter >
2. Use the Embedding API