Sau khi đã tạo ra các vector (embedding) từ văn bản, câu hỏi tiếp theo là: **Làm thế nào để lưu trữ và tìm kiếm hàng triệu, thậm chí hàng tỷ vector này một cách hiệu quả?**

Câu trả lời chính là **Vector Database (Cơ sở dữ liệu Vector)**.

**Vector Database là gì?**

Hãy quay lại với ví dụ về thư viện. Nếu **Embedding** là việc gán cho mỗi cuốn sách một tọa độ trên bản đồ, thì **Vector Database** chính là **hệ thống định vị GPS siêu nhanh cho bản đồ đó**.

Thay vì phải đi bộ khắp thư viện để so sánh từng cuốn sách (tìm kiếm tuần tự), bạn chỉ cần nhập tọa độ của thứ bạn muốn tìm, và hệ thống GPS này sẽ ngay lập tức chỉ ra những cuốn sách ở gần tọa độ đó nhất.

Về mặt kỹ thuật, **Vector Database** là một loại cơ sở dữ liệu được tối ưu hóa đặc biệt cho việc lưu trữ và truy vấn các vector đa chiều. Nó sử dụng các thuật toán xấp xỉ láng giềng gần nhất (Approximate Nearest Neighbor - ANN) để tìm kiếm cực kỳ nhanh, ngay cả với lượng dữ liệu khổng lồ, mà không phải hy sinh quá nhiều độ chính xác.

**Quy trình Lưu và Tìm kiếm**

Quy trình tổng thể rất trực quan và bao gồm 3 bước chính:

1. **Embedding (Vector hóa):** Dùng một mô hình (như Sentence Transformer) để chuyển đổi dữ liệu của bạn (câu, đoạn văn, ảnh...) thành các vector số.
2. **Indexing (Lập chỉ mục):** Đưa các vector này vào một Vector Database. Database sẽ sắp xếp chúng vào một cấu trúc hiệu quả để tìm kiếm nhanh chóng.
3. **Searching (Tìm kiếm):**
   * Chuyển đổi câu hỏi hoặc dữ liệu truy vấn của bạn thành một vector (sử dụng **cùng một mô hình** embedding).
   * Dùng vector truy vấn này để tìm kiếm trong database.
   * Database sẽ trả về các vector (và dữ liệu gốc tương ứng) gần nhất với vector truy vấn của bạn.

**Mini-Project: Xây dựng một hệ thống tìm kiếm ngữ nghĩa đơn giản**

Dưới đây là một project hoàn chỉnh bằng Python sử dụng **FAISS** (Facebook AI Similarity Search), một thư viện cực kỳ mạnh mẽ và phổ biến để xây dựng Vector Database ngay trên máy của bạn.

**Các lựa chọn Vector Database khác**

* **FAISS:** Tuyệt vời để bắt đầu và cho các ứng dụng chạy trên một máy duy nhất.
* **ChromaDB, Qdrant:** Các lựa chọn mã nguồn mở phổ biến, có thể chạy trên máy của bạn hoặc trên server, cung cấp nhiều tính năng hơn như lọc metadata.
* **Pinecone, Weaviate:** Các dịch vụ cloud được quản lý (managed services). Bạn không cần lo về việc cài đặt hay vận hành, chỉ cần đẩy dữ liệu qua API. Đây là lựa chọn hàng đầu cho các ứng dụng lớn và chuyên nghiệp.