Vector Database – wyspecjalizowane bazki zaprojektowane do store’owania i manipulacji wysoko-przestrzennymi wektorami danych. 80% danych jest nieustrukturayryzowane i nie pasuje do kontkestu relacyjnej bazy danych. Wektorowe bazy danych pomagają rozwiazywac problem z efektywnym store’owaniem danych i indeksowaniem by wykonywać query szybko.

Unstructured data są przetwarzane na początku w procesie vector embeddingu. Przykładem takiej danej mogą być zdjęcia, co jest 2D RGB obiektem, video, audio. Takie dane trafiają do embedding modelu, w wyniku czego otrzymujemy arraykę z liczbami, co jest vector embed. Dla przykłądu procesem vector emedingu będzie zakdowanie słowa „HELLO WORLD” „HELLO MOSHI” takie dwa słowa są wrzucane do embeddingującego algorytmu, który w wyniku przetwarzania daje arrayke np. [ 1 1 0 ] albo [1 0 1] gdzie [1 – Hello 1 – Moshin 1 – World] jak wartość 0 to znaczy ze słowa brak. Obraz zawierający tekst, zrzut ekranu, diagram, Czcionka

Opis wygenerowany automatycznie

Inny przykład z owockami/ warzywami/ kwiatami. Kategorie mniej wiecej grupia się w tym samym obszarze. Clustrowanie i Classyfikacja to co cos co się wykonuje w procesie czyli określenie labelek, klas danych grup jak np. ten przykład z owocami warzywami.

Obraz zawierający tekst, diagram, linia, Czcionka

Opis wygenerowany automatycznie

Można wyliczać odleglosci miedzy takimi grupami stosując rozne algorytmy co pozwala okreslic odleglosc miedzy punktami.

Obraz zawierający tekst, diagram, linia, Czcionka

Opis wygenerowany automatycznie

W procesie indexingu te same kategorie są ze sobą grupowane.

Przykłady vector database: Pinecone, Milvus, ScaNN, Weaviate, Vespa, FAISS, Chroma, Redis, Annoy, PgVector.

USE case’y:  
LLM – polegaja na dataset na kotrym były trenowane. Mogą nie posiada aktualnych wiaodmosci, albo specyficznej wiedzy. Vector database może pomoc rozszerzyć możliwości, albo uaktualnić infomracje.  
Semantic search - wyszukując koncpetu jak picture of waving nie chodzi o słowa kluczowe, ale o abstrakcje czynności, dla tego interpolacja wynikow jest bardziej oczekiwana do czego nadaja się wektorowe bazki. Porownujac ze sobą wkeotry, czyli wykonując vector embedding i obliczając podobniestwo miedzy wektorami można wyciagnac interesujące nas rezultaty.  
Personalizaowane rekomendacje – DLRM by Facebook/ Neural Collaborative Filtering (NCF) / Variational Auto-encoder Collaborative Filtering (VAE-CF). W takim przypadku tez sie nada bo bazujac na rekomendacji szuka odpowiedniej zwrotki.  
Wykrywanie anomalii – gdy cos znajduje się w zasięgu malicious zachowan.

Początek formularza

**What problem does a vector database address in large language models (LLMs)?**

**ANSWER: Memory issues**

Dół formularza