Laporan Progres 4

Embedding

Oleh: Mahendra Zidane Rainafa (24523142)

A. Pendahuluan

Tahapan ini berfokus pada pembangunan sistem memori jangka panjang (*Ingestion*) untuk mengatasi keterbatasan pengetahuan umum pada LLM standar dengan memanfaatkan database vektor . Proses teknisnya melibatkan pembacaan dokumen PDF dari Google Drive, konversi teks menjadi representasi vektor matematika menggunakan model Google Gemini, dan penyimpanannya ke dalam Pinecone. Mekanisme ini sangat kritikal untuk memungkinkan sistem melakukan pencarian semantik (*semantic search*) yang akurat terhadap data-data privat atau dokumen spesifik yang tidak dimiliki oleh model publik.

B. Skema Workflow

Sesuai instruksi pengembangan Workflow 3, arsitektur *pipeline* data yang dibangun adalah sebagai berikut : Google Drive → Text Splitter → Embedding Model → Pinecone (Upsert)

1. **Source:** Mengunduh file PDF secara otomatis dari folder tertentu di Google Drive.
2. **Processing:** Memecah dokumen panjang menjadi potongan-potongan kecil (*Chunking*) agar muat dalam konteks window model.
3. **Embedding:** Menggunakan model Google Gemini (models/text-embedding-004) untuk mengubah teks menjadi vektor. *(Catatan: Menggantikan OpenAI sesuai ketersediaan resource)*.
4. **Storage:** Menyimpan vektor dan metadata ke dalam index Pinecone.

C. Langkah Pengerjaan

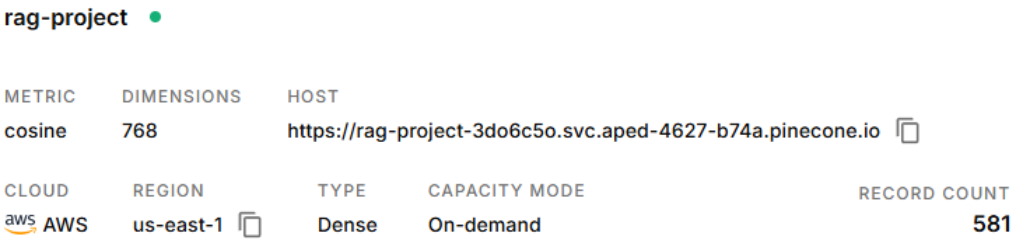
1. **Konfigurasi Database:** Membuat Index baru di dashboard Pinecone dengan dimensi yang sesuai untuk model text-embedding-004.
2. **Integrasi Google Drive:** Menghubungkan kredensial Google Drive API ke n8n dan mengkonfigurasi node untuk mencari serta mengunduh file PDF target.
3. **Implementasi Embedding:** Membangun workflow di n8n yang membaca file, memecahnya (*splitting*), dan mengirimkannya ke API Gemini untuk proses embedding.

4. **Penyimpanan Data:** Menambahkan node **Pinecone Vector Store** dengan mode *Upsert* untuk menyimpan hasil vektor ke database.
5. **Eksekusi:** Menjalankan workflow untuk memproses data latih (dokumen PDF) hingga status berhasil.

D. Hasil Implementasi

1. Screenshot Hasil Embedding:

2. Screenshot Data Masuk ke Pinecone:



3. Link GitHub Workflow:

   https://github.com/kevvy01/Project-RAG/blob/main/workflows/workflow-3-embedding.json