# RAG Yöntemi ile Vektör Veritabanı Oluşturma ve Sorgulama

Tarih: 2025-09-06 15:54

Kapsam: Metinlerin embedding'e dönüştürülerek Chroma vektör veritabanına aktarılması ve sorgu ile en yakın sonuçların getirilmesi. Referans: Chroma Resmî Dokümantasyon, RAG Yapısı – Pinecone.

## Mimari Özeti

Data → Embedding (paraphrase-multilingual-MiniLM-L12-v2) → Chroma (rag\_docs) → Retrieval → Reranker (cross-encoder/ms-marco-MiniLM-L-6-v2) → LLM (OpenAI/HF).

## Değerlendirme Sonuçları (Özet)

- Soru sayısı: 6

- Ortalama yanıt süresi: ~13384 ms

- Cevaplar sonunda kullanılan kaynaklar [Kaynak: path, Parça: i] biçiminde listelendi.

## Soru–Cevap Tablosu

|  |  |  |  |
| --- | --- | --- | --- |
| Soru | Cevap (kısaltılmış) | Süre (ms) | Kaynaklar |
| embedding nedir? | Embedding, ham metinleri belirli bir formata dönüştürerek, bu metinlerin sayısal temsillerini oluşturan bir süreçtir. Bu sayısal temsiller, metinlerin anlamını ve ilişkilerini daha iyi temsil etmek için kullanılır. Örneğin, bir metin parçası, embedding yöntemi ile vektörlere dönüştürülerek, veri ... | 7425 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1] |
| RAG yaklaşımı neyi amaçlar? | RAG (Retrieval-Augmented Generation) yaklaşımı, bilgiye erişimi ve metin üretimini birleştirerek daha doğru ve bilgilendirici içerikler oluşturmayı amaçlar. Bu yaklaşım, öncelikle büyük veri setlerinden bilgi çekmeyi ve ardından bu bilgiyi kullanarak daha anlamlı ve bağlama uygun yanıtlar ... | 5317 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1] |
| Chroma nedir ve ne işe yarar? | Chroma, ham metinleri (PDF, HTML, TXT, MD) okuyarak bunları parçalara (chunk'lara) bölüp, embedding'ler üreten ve bu verileri kalıcı olarak saklayan bir vektör veritabanıdır. Chroma'nın başlıca işlevleri şunlardır: - Ham verileri işleyerek anlamlı parçalara ayırma. - Bu parçaları embedding ... | 6537 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1] |
| What is an embedding? | Embedding, ham metinleri okuyarak ve bunları belirli parçalara (chunk'lara) bölerek oluşturulan vektör temsilleridir. Bu vektörler, metinlerin anlamını ve bağlamını sayısal bir biçimde temsil eder. Embedding'ler, Chroma gibi veritabanlarına kalıcı olarak kaydedilir ve sorgulama işlemlerinde ... | 26748 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1] |
| How does RAG reduce hallucinations? | RAG (Retrieval-Augmented Generation), yanıltıcı bilgileri azaltmak için aşağıdaki yöntemleri kullanır: 1. \*\*Veri Tabanı Kullanımı\*\*: RAG, sorgulara yanıt vermek için önceden oluşturulmuş bir vektör veri tabanından bilgi çeker. Bu, modelin daha doğru ve güvenilir bilgilere erişmesini sağlar. 2. ... | 7264 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1] |
| What is a vector database? | Vektör veritabanı, ham metinleri okuyup, bunları parçalara (chunk) bölerek embedding'ler üreten ve bu verileri kalıcı olarak saklayan bir sistemdir. Bu tür veritabanları, metin verilerini sorgulamak için kullanılır. Örneğin, Chroma vektör veritabanı, kullanıcıların metinleri analiz etmesine ve ... | 27017 | [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 1]; [Kaynak: C:\Users\Halil Turan\rag-proje\data\example\_en.md, Parça: 0] |

## Notlar & Geliştirme Önerileri

- Chunking 800/100 ile başlandı; 600–1000 aralığı veri yapısına göre denenebilir.

- Aday sayısı (candidates) 10–20, top\_k 3–5 önerilir.

- Kalite için veri kapsamı artırılabilir; domain-özel prompt şablonu eklenebilir.

- Performans için caching ve daha hafif LLM’ler (veya OpenAI küçük model) tercih edilebilir.