RAG-SupportBot: Mimari Dokümantasyon

Bu belge, cihaz üstünde (offline) çalışan belge tabanlı teknik destek asistanı olan RAG-SupportBot projesinin mimari kararlarını, kullanılan teknolojileri ve bileşenlerin işleyişini detaylı şekilde açıklamak amacıyla oluşturulmuştur.

Kullanılan Ana Modeller

Embedding Modeli: `sentence-transformers/all-MiniLM-L6-v2`

Bu model, küçük boyutlu ve hızlı olmasının yanı sıra semantic search kalitesi açısından dengeli sonuçlar verir. MVP aşamasında yüksek hız ve düşük kaynak tüketimi sağladığı için tercih edilmiştir. Daha sonra `nomic-embed-text-v1` veya `instructor-xl` gibi daha güçlü embedding modellerine geçilmesi planlanmaktadır.

Vector Store: FAISS + LlamaIndex

FAISS, yüksek performanslı bir vektör benzerlik arama kütüphanesidir. LlamaIndex ile entegre edilerek belge parçalarının vektörleştirilmiş hallerini saklamak ve sorgulara göre en benzer içerikleri bulmak için kullanılır. Veriler `db/faiss_index` klasöründe JSON tabanlı formatta saklanır.

MVP Modülleri

embedder.py

Belgeleri 500 token'lık anlamlı parçalara ayırır, her parçanın embedding vektörünü üretir ve FAISS vektör veritabanına kaydeder.

retriever.py

Kullanıcıdan gelen sorunun embedding vektörünü üretir, FAISS veritabanından en benzer içerikleri getirir. Geri dönen parçalar, puanlanmış halde modele aktarılmaya hazırdır.

Gelecek Genişletmeler

- `generator.py`: Belge parçalarından nihai cevap üreten LLM bileşeni
- `query_rewriter.py`: Kısa veya eksik soruları yeniden yazan modül
- `document_grader.py`: Getirilen belgelerin kalite kontrolünü yapan zincir
- `fallback_router.py`: Cevap üretilemediğinde alternatif yolları yöneten yapı