**Web Trafik Loglarına Dayalı Yapay Zeka Destekli Soru-Cevap Sistemi Geliştirme Raporu**

**Aşama 1: Veri Hazırlığı ve Ön İşleme**

Bu aşamada ilk önce Apache formatında veri setini oluşturuldu. Oluşturulan veri seti rastgele veriler eksik olacak şekilde .csv dosyasına kaydedildi. ( web\_traffic\_logs\_creating )

Veri seti; IP Adresi, Zaman Damgası, HTTP Yöntemi, Erişilen Sayfa Yolu, Durum Kodu, Yanıt Boyutu sütunlarından oluşmaktadır.

Ön işleme kısmında eksik veri bulunan Zaman Damgası, HTTP Yöntemi sütunlarındaki satırlar silindi. Yanıt Boyutu sütunundaki NaN değerler ise ortalaması ile dolduruldu. Bu noktada satırlar silinebilir de.

Gerekli verilerin seçilmesi noktasında ise tüm sutunlar anamlı bilgiler oluşturmak için kullanıldı.

Verilerin vektörleştirmesi kısmında nasıl vektörleştirileceğine dair sorunlar yaşandı.

İlk önce tüm veriler str haline getirildi. Model oluşturulduktan sonra bu yöntemin yanlış olduğu, doğru sonuçlar alınmadığı gözlendi.

Sonrasında veriler teker teker vektörleştirildi: IP adreslerine ve Erişilen sayfa yollarına Label Encoding işlemi uygulandı. Zaman damgası Unix cinsinden zaman çevrildi. HTTP metodları One Hot Encoding işlemi yapılarak sayısal verilere dönüştürüldü. Durum kodları ve Yanıt boyutları sayısal veriler oldukları için onlara bir işlem uygulanmadı.

Vektörel veri tabanı olarak FAISS seçildi.

Yüklenecek veriler seçildi. Path sutunu üzerinden ilk model oluşturulduğu için TF-IDF ile o sutun ayrı olarak fit edildi. Seçilen sutunlar bir numpy dizisine dönüştürüldü. Dizi, vektörlerden daha iyi sonuçlar almak için ölçeklendirildi.

Ölçeklendirilen veriler veri tabanına yüklendi.

Bu aşamada uyumsuz indis problemleri yaşandı. Vektörleştirme adımları gözden geçirilerek bu problem ortadan kaldırıldı.

**Aşama 2: RAG Modelinin Kurulumu**

Vektör veritabanı ile (FAISS) kullanıcının sorgusu için k-NN arama yapılmaktadır.

Sorgu vektörünün en yakınındaki sonuçları getirmesi üzerine sistem kurulmuştur. Bu da sorulan bazı direkt soruların cevaplanmamasına yol açar. Örneğin ‘Bu data frame kaç satırlık bir veriden oluşmaktadır?’ sorusuna cevap vermez. Bunun cevaplanabilmesi için farklı arama algoritmaları kullanılabilir.

Bağlam oluşturma kısmında önemli bir hata ile karşılaşıdı.

Model için ilk önce Path sütunu kullanıldı. Bu aşamada sorgularda doğru sonuçlar alınmadı. Bu model kullanılıyorken IP adresleri üzerinden sorgu yapılamadı. Bu kısım data frame’i kullanacak şekilde ( context = " ".join([str(df.iloc[i].to\_dict()) for i in indices]) ) değiştirildi. Bunun sonucunda sorgularda tüm sutunlar üzerinden sorgu yapılabilir hale gelindi.

Yanıt oluşturma kısmında GPT dil modeli kullanıldı. İlk olarak ücretsiz verilen sorgu hakkı kullanıldı.

Bittiğinde ise api key için ödeme yapılması gerektiğine dair hatalar alındı. Bu hatalar giderildi.

**Aşama 3: Sistem Entegrasyonu ve Test**

Olusturulan fonksiyonlar ve kod parçaları RAG modeli altında birleştirildi.

K-NN aldoritmasına uygun doğru sorular sorulduğunda büyük ölçüde doğru sonuçlar veren bir Q&A sistemi geliştirildi.

Bu sistem sonuçları kesin olan sorulara cevap vermemektedir. ‘Delete metodunu kullanıp about.html adresine erişen kullanıcıların IP adresleri nedir?’ gibi bağlamlı sorular için cevaplara erişir. Bu kısımda da dikkat edilmesi gereken nokta ise: Bu sorunun birden fazla cevabı olabilir. Ancak algoritma gereği en yakın sonuçları veya sonucu getirir.

**Aşama 4: Performans Değerlendirmesi**

pandas\_query dosyasında web\_traffic\_logs\_analysis dosyasındaki sözel sorguların pandas sorguları bulunmaktadır. Burada sayısal sorular ile doğru cevaplar alınmazken, bağlamlı sorulara doğru cevaplar alınmaktadır.

Farklı algoritmalar ve kütüphaneler (PyTorch gibi) ile modeller kurularak direkt sayısal sonuçlar veren bir sistem geliştirilebilir. Böylece kullanıcıyı soru noktasında kısıtlamayan daha sağlıklı bir sistem meydana getirilmiş olur.