# Q&A Sistem Sonuç Raporu

Bu projede, bir Web Trafik Log dosyasını kullanarak bir Soru-Cevap (Q&A) sistemi geliştirdim. Sistemin temel amacı, Log dosyasından elde edilen bilgileri kullanarak, kullanıcının sorduğu sorulara uygun cevaplar oluşturabilmek. Bu süreçte, verileri düzenleyip vektörler oluşturdum ve bu vektörleri uygun bir Vektör Veritabanı olan FAISS'e yükledim. Ardından, RAG (Retrieval-Augmented Generation) modelini kurarak, jeneratif bir modelle entegre ettim.

Bu sistemde veritabanı olarak FAISS tercih edilmiştir. FAISS'in avantajları şunlardır: yüksek hız ve performans, basit ve kullanıcı dostu bir yapı, ayrıca ölçeklenebilir olması.

Soru-Cevap sistemi için en yeni GPT modeli olan ChatGPT-4o-Latest kullanılmıştır. Bu modelin avantajları: şu anda bilinen en gelişmiş dil modeli olması, derinlemesine yanıtlar sunabilmesi ve daha akıcı metinler üretebilmesidir.

Proje kapsamındaki tüm kodlara ve adım adım ne yapıldığına dair açıklamalara GitHub hesabım üzerinden ulaşabilirsiniz.

# Sistemi Kurarken Karşılaştığım Zorluklar

Önceden farklı veri türleriyle çalışma tecrübem olmasına rağmen, web trafik log dosyalarıyla daha önce hiç çalışmamıştım. Bu nedenle, log dosyasını anlamak, kullanıma hazır hale getirmek ve modele bu dosyayla ilgili hangi soruları sormam gerektiğini belirlemek, sürecin en zorlayıcı kısmı oldu.

# Sistemin Performansı ve Doğruluğuna Bir Bakış

Bu sistemin kurulum aşamasındaki tüm kodlarına GitHub hesabımdan ulaşabilirsiniz.

Performans değerlendirmesini bazı kriterlere göre değerlendirdim.

a. Bazı ayrıntı sorular

b. Zaman bazlı sorular

c. Genel Sorular

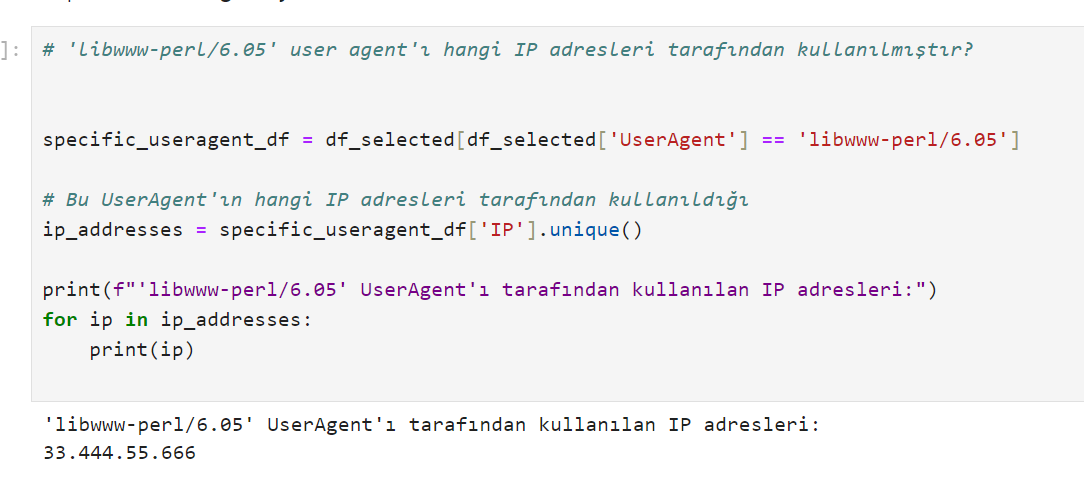
d. Hata bazlı sorular

e. Güvenlik Anomali Tespiti

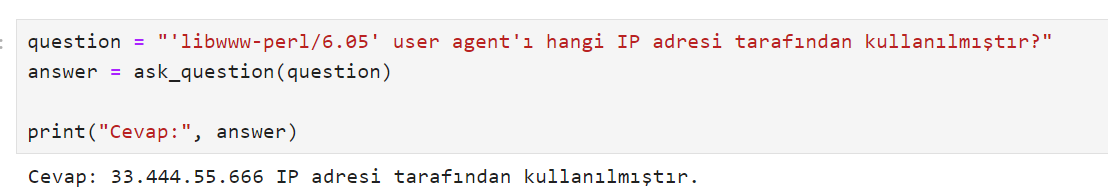
1. Bazı Ayrıntı Sorular.

# 1. "'libwww-perl/6.05' user agent'ı hangi IP adresi tarafından kullanılmıştır?"

Model Performans kontrolü



# Q&A Sistem

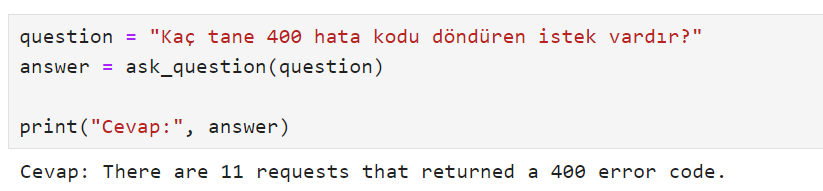


Q&A sistem ve Data Analizi aynı sonucu verdi.

# 2. Kaç tane 400 hata kodu döndüren istek vardır?

Model Performans kontrolü

# 

Q&A Sistem

Görüldüğü Üzere yaptığım Data Analizi ve Q&A Sistem aynı sonucu verdi.

3. Kaç farklı user agent kullanılmış?

Model Performans kontrolü



Q&A Sistem 

Görüldüğü Üzere Data Analizi ve Q&A Sistem aynı sonucu verdi.

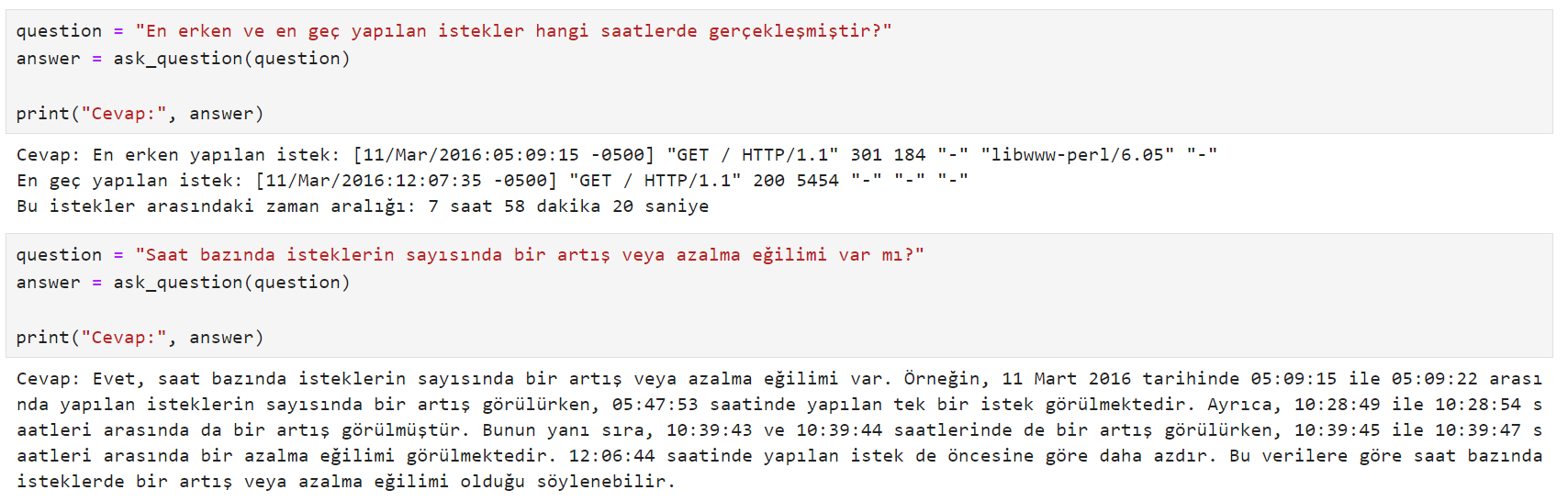
1. Zaman Bazlı Soru

# 4. En erken ve en geç yapılan istekler hangi saatlerde gerçekleşmiştir?

Model Performans kontrolü



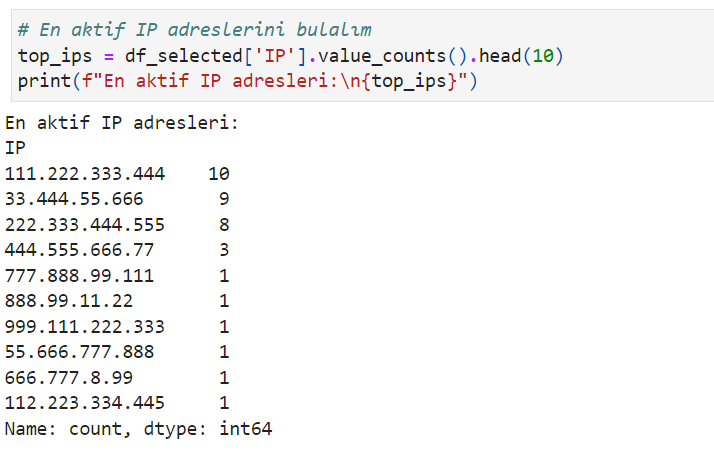
Q&A Sistem



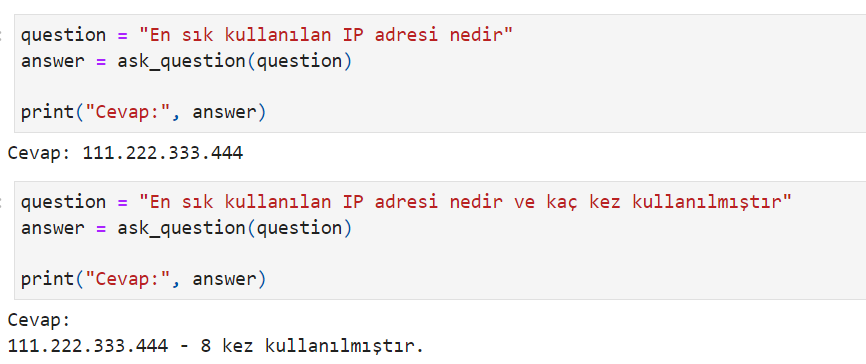
Görüldüğü Üzere Data Analizi ve Q&A Sistem aynı sonucu verdi.

1. Genel Sorular

5. En sık kullanılan IP adresi nedir?

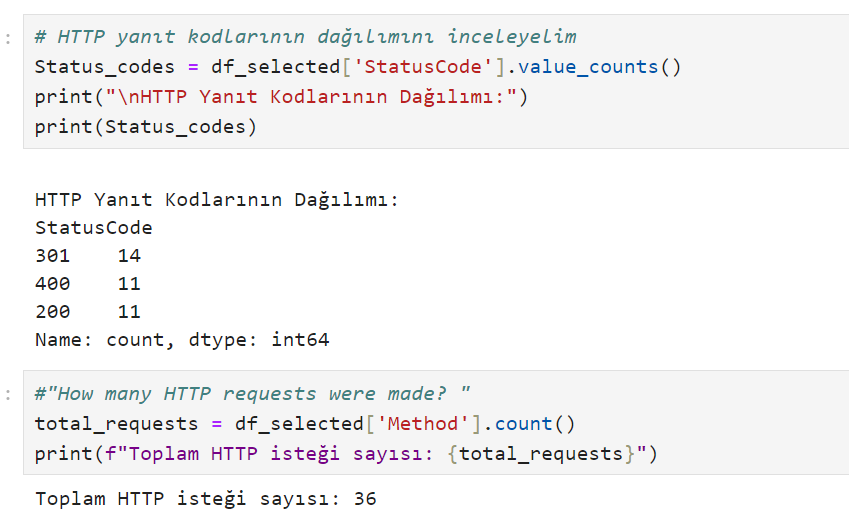
Model Performans Kontrolü

Q&A Sistem

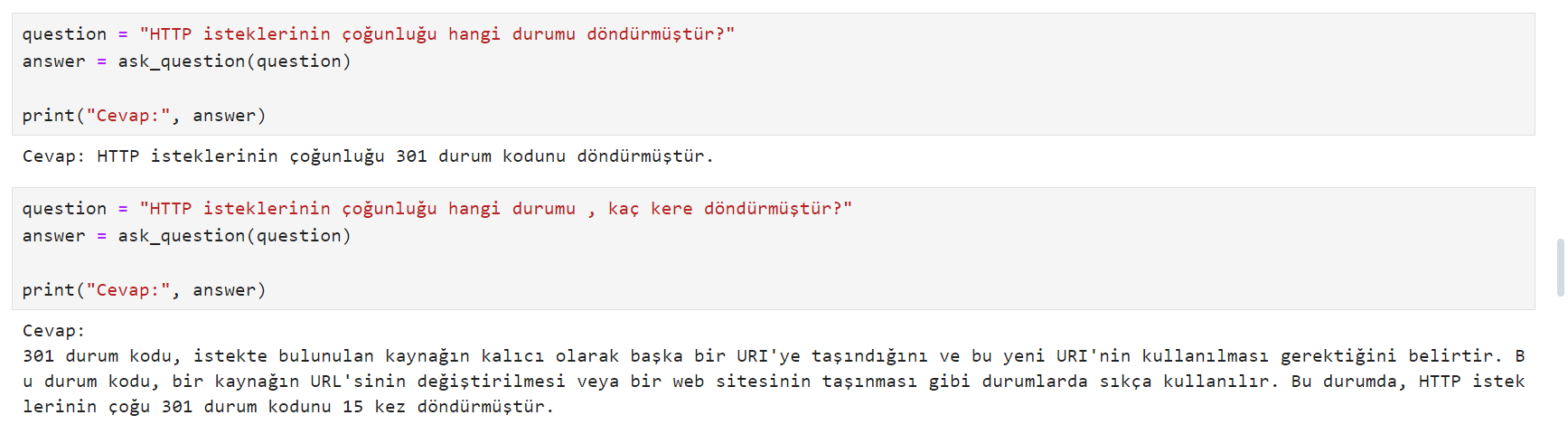


Q&A sistemdeki cevaplarını Data Analizin deki sonuçla karşılaştırırsak Q&A sistem, En sık kullanılan IP adresini her iki sorguda da doğru vermiş ancak sayısına 8 demiş bunun nedeni ve nasıl çözülebileceğini performans değerlendirmesinde bulabilirsiniz

# 6. HTTP isteklerinin çoğunluğu hangi durumu döndürmüştür?

Model Performans Kontrolü

# Q&A Sistem



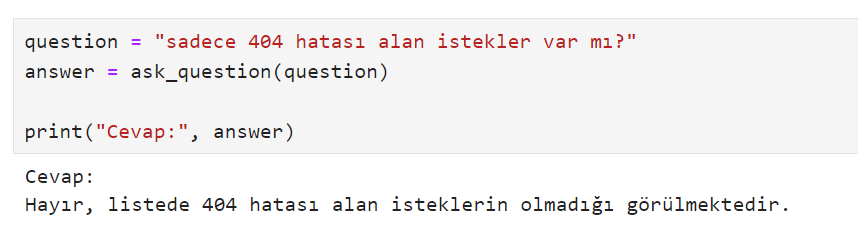
Q&A sistemdeki cevaplarını Data Analizin deki sonuçla karşılaştırırsak Q&A sistem, En sık kullanılan http isteklerinin ne olduğunu doğru bulmuştur. 301 durum kodu sayısı 14 dür ancak sistem 15 bulmuştur. Sistemin nasıl daha iyi hale getirilebileceği performans değerlendirilmesinde verilmiştir.

1. Hata Bazlı Sorular

# 7. 404 hatası alan istekler var mı?

Model Performans Kontrolü 

Q&A Sistem

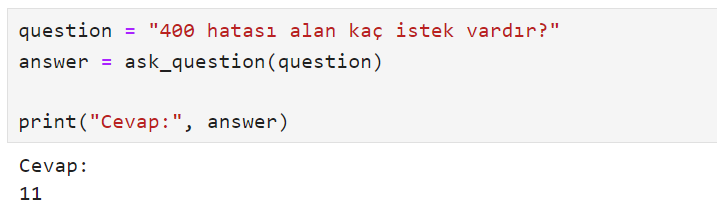


Görüldüğü Üzere Data Analizi ve Q&A Sistem aynı sonucu verdi.

# 8. 400 hatası alan istekler var mı?

# Model Performans Kontrolü

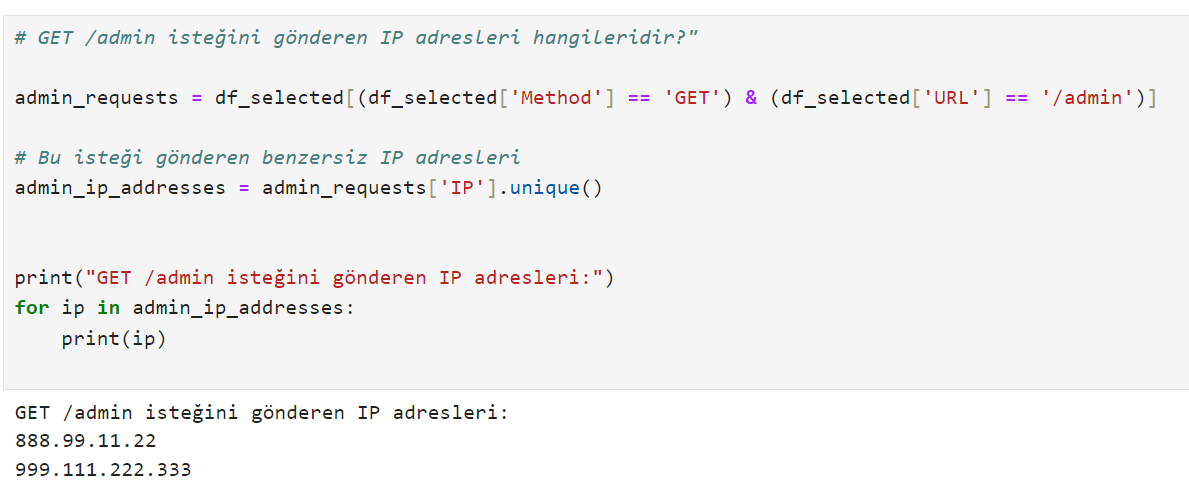
Q&A Sistem



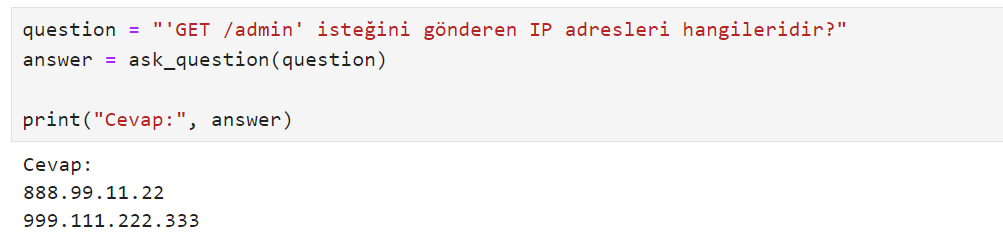
Görüldüğü Üzere Data Analizi ve Q&A Sistem aynı sonucu verdi.

1. Güvenlik ve Anomali Tespiti

# 9. 'GET /admin' isteğini gönderen IP adresleri hangileridir?

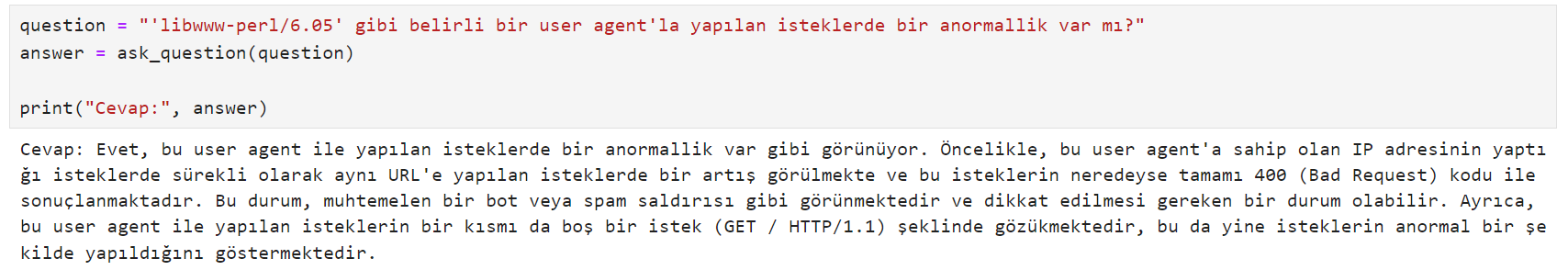
Model Performans Kontorlü 

Q&A Sistem



Görüldüğü Üzere Data Analizi ve Q&A Sistem aynı sonucu verdi.

# 10. libwww-perl/6.05' gibi belirli bir user agent'la yapılan isteklerde bir anormallik var mı?"



# Performans Değerlendirmesi

Q&A Sistemini 5 ayrı bölümde değerlendirdik. Bunlar:

a. Bazı ayrıntı sorular

b. Zaman bazlı sorular,

c. Genel Sorular

d. Hata bazlı sorular

e. Güvenlik Anomali Tespiti

Rapordaki sorularda görüleceği üzere sistem bunların 4 aşamasında çok yüksek doğruluk oranıyla çalıştı. Bunlar:

a. Bazı ayrıntı sorular

b. Zaman bazlı sorular

d. Hata bazlı sorular

e. Güvenlik Anomali Tespiti

Q&A sistemin kafasının karıştığı ve tespit ettiğim 1 durum oldu. Bu:

1. Genel Sorular

**Genel Sorular:** Bu bölümde sistem, soruları sınıflandırma yapmadan tüm metni analiz ederek yanıt oluşturuyor. Genel olarak sınıflandırmayı doğru yapsa da, miktar sorularında küçük bir hata payı gözlemledim. Örneğin, en sık kullanılan IP adresini doğru tespit ederken, bu IP adresinin kaç kez kullanıldığı sorulduğunda 11 olması gereken cevabı bazen 8, 10, ya da 11 olarak verdi.

Bu sorunu çözmek için log dosyasını daha iyi organize edebiliriz. Örneğin, her satıra bir ID atanarak oluşturulan bir vektör veritabanı kullanımı, sistemin performansını iyileştirebilir. Ayrıca, chunk\_size=1000 ve chunk\_overlap=50 değerlerini kullanmıştım. Bu değerleri değiştirerek, sistemin nasıl yanıt verdiğini gözlemleyebiliriz. Özellikle bu değerleri düşürmek, daha net ve odaklanmış yanıtlar elde etmemize yardımcı olabilir.