Dinamik Web Scraping ve Sayfa Analiz Aracı

HAZIRLAYAN: GAZALİ KEPENÇ

TARİH: 20.12.2025

İÇİNDEKİLER

1.GİRİŞ

Bu rapor, Siber Tehdit İstihbaratı (CTI) faaliyetlerinde temel bir yetenek olan **web tabanlı veri toplama ve sayfa analizinin** otomatikleştirilmesini amaçlayan *Dinamik Web Scraping ve Sayfa Analiz Aracı* çalışmasının sonuçlarını sunmaktadır.

Modern tehdit aktörleri; forumlar, haber siteleri, açık kaynak platformlar ve kamuya açık web sayfaları üzerinden iz bırakmaktadır. Bu nedenle bir CTI analistinin, hedef web kaynaklarını **güvenilir**, **kanıt üretir** ve **tekrar edilebilir** şekilde analiz edebilmesi kritik önemdedir.

Geliştirilen araç;

- Hedef web sayfasının ham HTML içeriğini,
- Sayfanın tarama anındaki görsel durumunu (ekran görüntüsü),
- Sayfa üzerindeki harici bağlantı yapısını (link haritası)

Otomatik olarak elde ederek CTI analiz süreçleri için temel veri setini oluşturmaktadır.

2. TEKNİK METODOLOJİ

Araç, modern web uygulamalarında yaygın olarak kullanılan JavaScript tabanlı dinamik içerik üretimini doğru şekilde analiz edebilmek amacıyla Headless Browser yaklaşımı ile tasarlanmıştır.

Kullanılan Teknolojiler

- **Programlama Dili:** Go (Golang)
- **Tarayıcı Motoru:** chromedp (Chrome DevTools Protocol)

Chromedp tercih edilmesinin temel nedeni; yalnızca statik HTML çekimi değil, aynı zamanda **tarayıcı seviyesinde render edilmiş DOM yapısına** erişim sağlamasıdır. Bu yaklaşım, klasik HTTP tabanlı scraper'ların kaçırdığı dinamik içeriklerin analiz edilmesine olanak tanır.

Temel Yetenekler

- Hedef URL'ye otomatik navigasyon
- Render edilmiş DOM ağacının tamamının çıkarılması
- HTTP durum kodunun tespiti
- Mükerrer bağlantıların elenerek sayfa link haritasının oluşturulması
- Tam sayfa ekran görüntüsü alınarak görsel kanıt üretilmesi

Bu metodoloji, CTI süreçlerinde kanıta dayalı analiz yapılmasını desteklemektedir.

3. ANALİZ TABLOSU

Bu bölümde, geliştirilen aracın 15 farklı karakteristikteki web sitesi üzerindeki performansı ve elde edilen istihbari veriler tablolaştırılmıştır.

No	Web Sitesi	Platform Türü	İçerik Kapsamı	Erişim Durumu	Teknik Not
1	btk.gov.tr	Resmî Kurum	Mevzuat, duyuru	200	Statik + yarı dinamik
2	archive.org	Dijital Arşiv	Açık veri, arşiv	200	Yüksek veri erişilebilirliği
3	youtube.com	Sosyal Medya	Video, medya	200	JS ağırlıklı
4	google.com	Arama Motoru	Genel bilgi	200	Bot algılama mevcut
5	figma.com	Tasarım Platformu	UI/UX içerik	200	Dinamik SPA
6	telegram.org	İletişim Platformu	Mesajlaşma	200	Hafif bot koruması
7	stackoverflow.com	Teknik Forum	Yazılım, Q&A	403	İçerik yine de alındı
8	yildizcti.com	CTI Platformu	Siber tehdit raporları	200	CTI odaklı
9	imdb.com	Medya Veritabanı	Film, dizi	403	Anti-bot aktif
10	shodan.io	Siber Güvenlik	Açık servis tarama	200	Güçlü CTI kaynağı
11	github.com	Kod Deposu	Açık kaynak	200	Dinamik içerik
12	sans.org	Güvenlik Kurumu	Eğitim, rapor	200	Kurumsal yapı
13	linkedin.com	Sosyal Ağ	Profesyonel veri	200	Oturum bağımlı
14	ktun.edu.tr	Akademik Kurum	Duyuru, akademik	200	Statik yapı
15	sciencedirect.com	Akademik Yayın	Bilimsel makale	403	Yayıncı koruması

4. TEKNİK ÇIKTI ÖRNEKLERİ

Sistemin çalışma prensibi ve ürettiği veriler aşağıda örnek çıktılar üzerinden gösterilmiştir. Bu bölümde, farklı HTTP durum kodlarına sahip hedefler üzerinde aracın davranışı ve ürettiği teknik çıktılar karşılaştırmalı olarak sunulmaktadır.

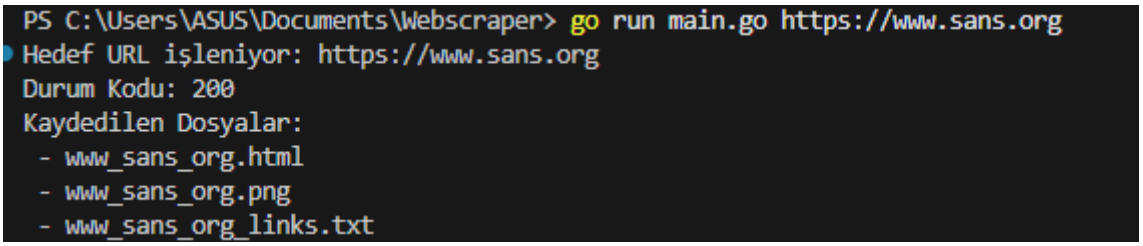
4.1 Terminal Üzerinden Çalıştırma ve Log Kayıtları (HTTP 200 – Başarılı Senaryo)

Uygulama, komut satırı üzerinden hedef URL argümanı alarak çalıştırılmıştır. Navigasyon, HTTP durum kodu tespiti ve dosya üretim süreci terminal üzerinden anlık olarak izlenmiştir.

Çalıştırılan Komut: ***go run main.go https://www.sans.org***

Terminal Çıktısı Analizi:

- Hedef URL başarıyla yüklenmiştir.
- HTTP durum kodu 200 OK olarak tespit edilmiştir.
- HTML içeriği, tam sayfa ekran görüntüsü ve bağlantı listesi hatasız şekilde oluşturulmuştur.



Görsel 1: go run main.go https://www.sans.org komutunun çalıştırılması, HTTP 200 durum kodunun alınması ve çıktı dosyalarının başarıyla oluşturulduğunu gösteren terminal çıktısı.

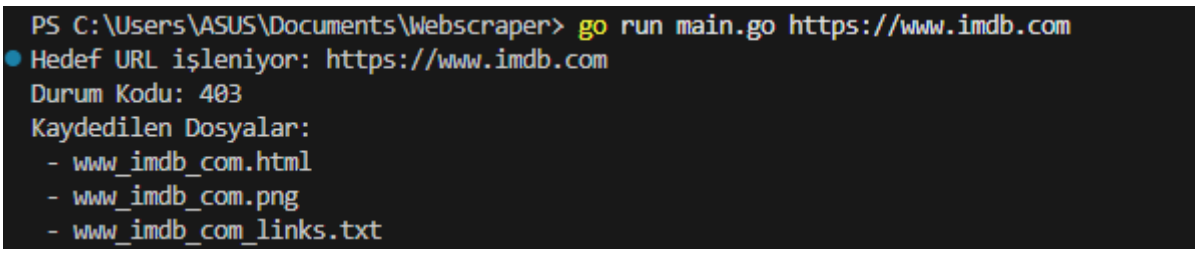
4.2 Terminal Üzerinden Çalıştırma ve Log Kayıtları (HTTP 403 – Erişim Kısıtlı Senaryo)

Bu senaryoda, bot algılama ve erişim kısıtlaması uygulayan bir platform üzerinde aracın davranışı gözlemlenmiştir.

Çalıştırılan Komut: ***go run main.go https://www.imdb.com***

Terminal Çıktısı Analizi:

- Hedef URL'ye bağlantı kurulmuştur.
- HTTP durum kodu 403 Forbidden olarak alınmıştır.
- Erişim kısıtına rağmen sayfa içeriği render edilmiş ve teknik çıktılar üretilmiştir.



Görsel 2: go run main.go https://www.imdb.com komutunun çalıştırılması sonucunda HTTP 403 erişim kısıtı alınmasına rağmen dosya üretiminin devam ettiğini gösteren terminal çıktısı.

4.3 Dosya Sistemi ve Çıktı Yönetimi

Her bir tarama işlemi sonucunda oluşturulan dosyalar, hedef siteye özgü şekilde isimlendirilerek proje dizinine kaydedilmektedir. Bu yaklaşım, veri bütünlüğünü korumakta ve CTI analiz süreçlerinde geriye dönük incelemeyi kolaylaştırmaktadır.

Üretilen Dosya Türleri:

- .html → Ham sayfa içeriği
- .png → Tam sayfa ekran görüntüsü
- _links.txt → Ayrıştırılmış bağlantı listesi

www_sans_org.html	20.12.2025 13:02	Chrome HTML Do...	439 KB
www_sans_org.png	20.12.2025 13:02	PNG Dosyası	141 KB
www_sans_org_links.txt	20.12.2025 13:02	Metin Belgesi	1 KB
www_imdb_com.html	20.12.2025 13:00	Chrome HTML Do...	1 KB
www_imdb_com.png	20.12.2025 13:00	PNG Dosyası	7 KB
www_imdb_com_links.txt	20.12.2025 13:00	Metin Belgesi	0 KB

Görsel 3: Farklı hedef siteler için oluşturulmuş HTML, PNG ve bağlantı listesi dosyalarının site bazlı isimlendirme ve dizin yapısını gösteren dosya sistemi görünümü.

4.4 Görsel Veri Analizi

Aşağıdaki görsel, chromedp motoru kullanılarak alınan tam sayfa ekran görüntüsünü göstermektedir. Bu çıktı, web sitesinin tarama anındaki gerçek kullanıcı görünümünü ve DOM yapısının başarıyla render edildiğini doğrulamaktadır.

Görsel 4: HTTP 200 durum kodu ile erişilen bir hedef sitenin, araç tarafından otomatik olarak oluşturulmuş tam sayfa ekran görüntüsü.

4.5 403 Erişim Kısıtı Altındaki Görsel Çıktı Değerlendirmesi

Bot algılama mekanizması bulunan platformlarda HTTP 403 yanıtı alınmasına rağmen, sayfa içeriğinin belirli bir seviyede render edilebildiği ve görsel çıktının oluşturulabildiği gözlemlenmiştir.

Görsel 5: HTTP 403 erişim kısıtına sahip bir platformdan elde edilen tam sayfa ekran görüntüsü. Bu çıktı, pasif istihbarat toplama açısından görsel kanıt niteliği taşımaktadır.

5.SONUÇ VE DEĞERLENDİRME

Geliştirilen araç, test edilen 15 farklı web kaynağının büyük çoğunluğunda HTML içeriği, görsel kanıt ve link haritası üretme görevlerini başarıyla yerine getirmiştir. Elde edilen sonuçlar, aracın CTI ön keşif (initial reconnaissance) aşamasında etkin bir şekilde kullanılabileceğini göstermektedir.

Analiz sürecinde özellikle arama motorları ve büyük platformlarda **403 erişim engelleri** ile karşılaşılmıştır. Bu durum, modern web altyapılarında bot tespit mekanizmalarının yaygın olarak kullanıldığını ortaya koymaktadır.

Bir sonraki geliştirme aşamasında aşağıdaki iyileştirmeler kritik önemdedir:

1. **User-Agent rotasyonu** ile tarayıcı parmak izinin çeşitlendirilmesi
2. **Headless tarayıcı tespitini azaltmaya yönelik önlemler**
3. Elde edilen linklerin ikinci seviye analiz (deep crawl) için kullanılması

Bu geliştirmeler ile araç, operasyonel CTI çalışmalarında daha etkin bir keşif platformuna dönüştürülebilir.