WEB SCRAPER

Raporu

İsim: Mustafa Talha DOĞAN

Grup: Yıldız Cti-Grup2

İçindekiler

Giriş

Bu çalışmanın amacı, Siber Tehdit İstihbaratı (CTI) süreçlerinin ilk adımı olan veri toplama işlemini otomatize etmektir. Geliştirmiş olduğum Go yazılımı, hedef web sitelerinden statik HTML içeriğini output.txt dosyasına kayıt etmekte ve sitenin o anki durumunu ekran görüntüsü almaktadır.

Projenin çalışması için terminale yazılması gereken kod : **go run main.go <URL>**

Kullanılan Teknolojiler ve Teknik Yaklaşım

Programlama Dili: Go ile yazılmıştır.

Kütüphaneler:

- **net/http:** Web sunucusuna http istekleri atmak ve dönen yanıtların(response) status kodlarını (200, 404 vb.) analiz etmek için kullanılmıştır.
- **net/url:** Kullanıcının girdiği uzun web adresinden sadece ana domain ismini ayıklayıp klasör adı olarak kullanmak için tercih edilmiştir.
- **chromedp:** Google Chrome'u arayüzsüz modda çalıştırıp sitelerin ekran görüntülerini almak için kullanılmıştır.
- **context**: Chromedp ile tarayıcıyı açarken bir zaman aşımı olup olmadığını kontrol etmek için kullanılmıştır.
- **errors:** HandleConnectionError fonksiyonunda, oluşan hatanın bir URL hatası mı yoksa başka bir hata mı olduğunu anlamak için kullanılmıştır.
- **fmt:** Konsola bilgi mesajları yazdırmak için kullanılmıştır.
- **io:** Çektiğimiz html verisini output.txt ye kopyalamak için kullanılmıştır.
- **log:** Kritik bir hata olduğunda hatayı ekrana basıp programı güvenli bir şekilde sonlandırmak için kullanılmıştır.
- **os:** Komut satırından girilen URL'yi almak , yeni klasör oluşturmak ve dosyaları diske kaydetmek için kullanıldı.
- **path/filepath:** Klasör ve dosya isimlerini düzgün birleştirmek için kullanılmıştır.
- **time:** Sayfa yüklendikten sonra tam render oluşması için 2 saniye beklemek için kullanılmıştır.

- **strings:** URL kısmındaki www. Kısmını dosya isminden kaldırmak için kullanılmıştır.

Teknik yaklaşım olarak; programın önce hedef URL'ye HTTP isteği atıp http durum kodunu (200 OK) doğruladığı, eğer site erişilebilir ise o domain adına özel bir klasör oluşturup html içeriğini ve ekran görüntüsünü klasöre kaydettiği, erişilebilir değil ise veya farklı bir hata var ise kullanıcıyı bilgilendirdiği bir yaklaşım izlenmiştir.

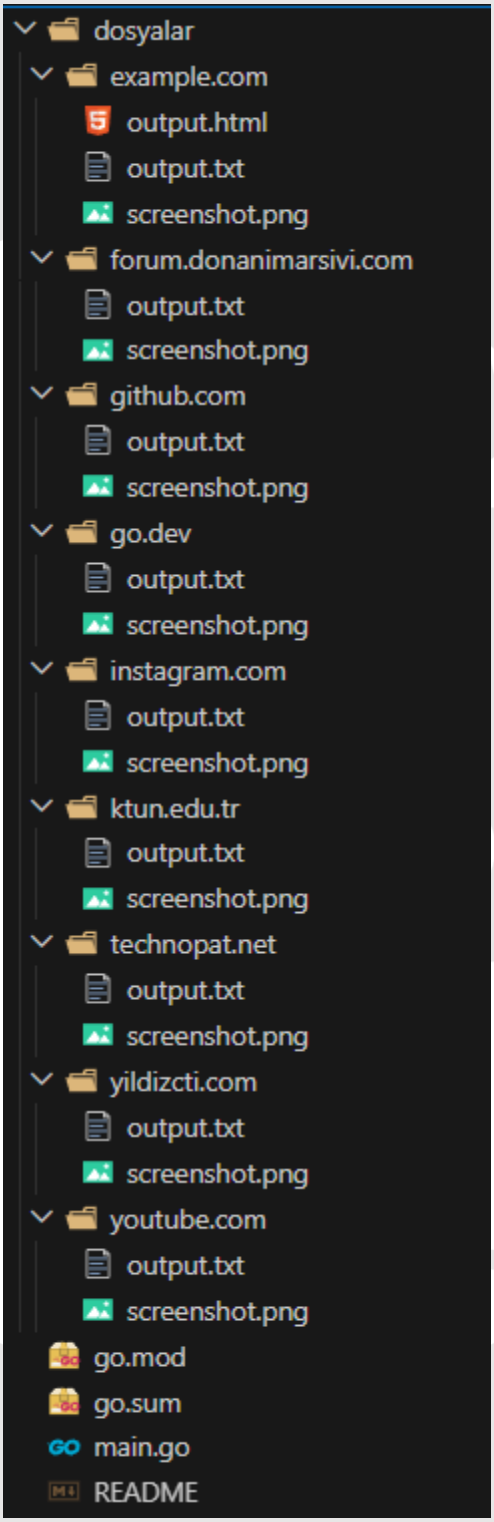
Test Süreci ve Bulgular

#	URL	Durum	HTTP Kodu	Sonuç Dosyaları
1	https://www.example.com/	Başarılı	200	Dosyalar/example.com
2	https://github.com/MustafaTalhadgn	Başarılı	200	Dosyalar/github.com
3	https://www.technopat.net/sosyal/	Başarılı	200	Dosyalar/technopat.net
4	https://busayfayok.com/	Başarısız	404	yok
5	http://httpbin.org/status/500	Başarısız	500	yok
6	http://httpbin.org/status/403	Başarısız	403	yok
7	https://yildizcti.com/	Başarılı	200	Dosyalar/yildizcti.com
8	https://www.youtube.com	Başarılı	200	Dosyalar/youtube.com
9	https://go.dev	Başarılı	200	Dosyalar/go.dev
10	https://www.ktun.edu.tr/	Başarılı	200	Dosyalar/ktun.edu.tr
11	https://forum.donanimarsivi.com/	Başarılı	200	Dosyalar/ forum.donanimarsivi.com
12	https://www.linkedin.com/in/ mustafatalhadogan/	Engellendi	999	yok
13	https://www.instagram.com/ mustafatalhadgn/	Başarılı	200	Dosyalar/instagram.com
14	http://10.255.255.1	Timeout	Timeout	yok
15	https://olmayansite.com	Başarısız	404	yok

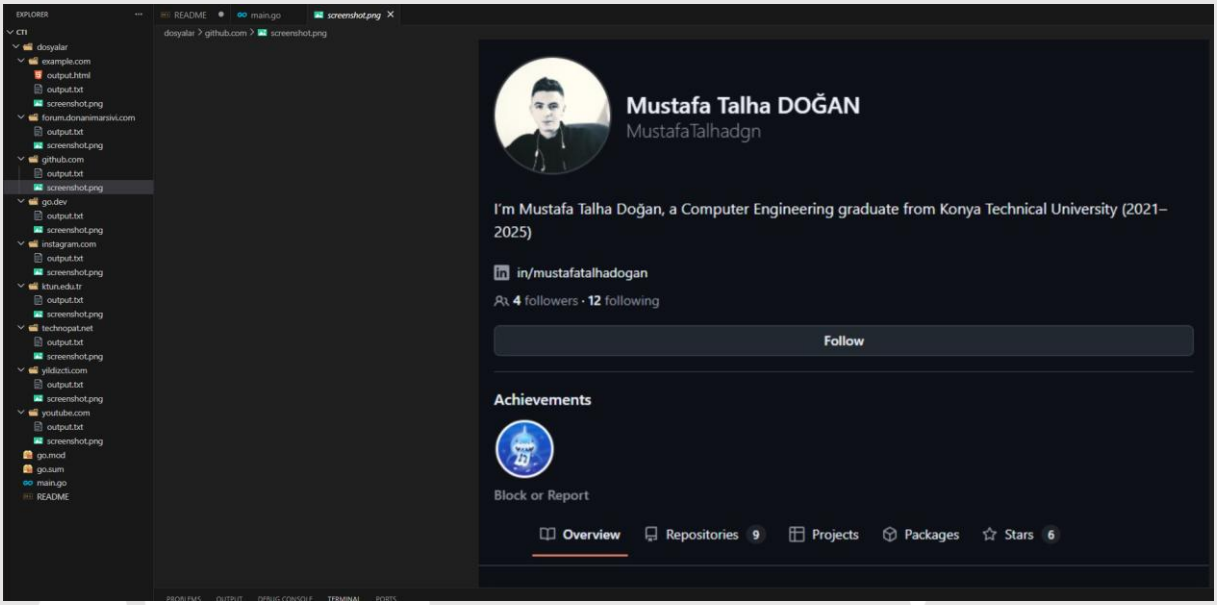
Tablo-1. Test Edilen Siteler ve Elde Edilen Sonuçlar

Not: Tabloda görülen 999 durum kodu, LinkedIn tarafından bot yazılımlara karşı uygulanan özel bir güvenlik önlemidir.

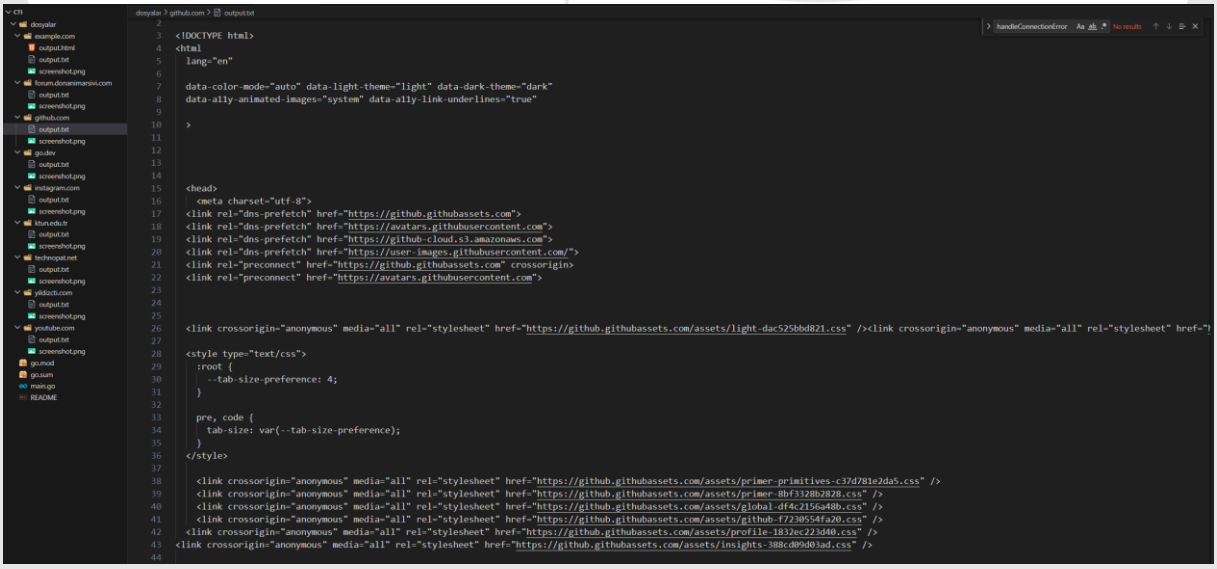
Kanıtlar



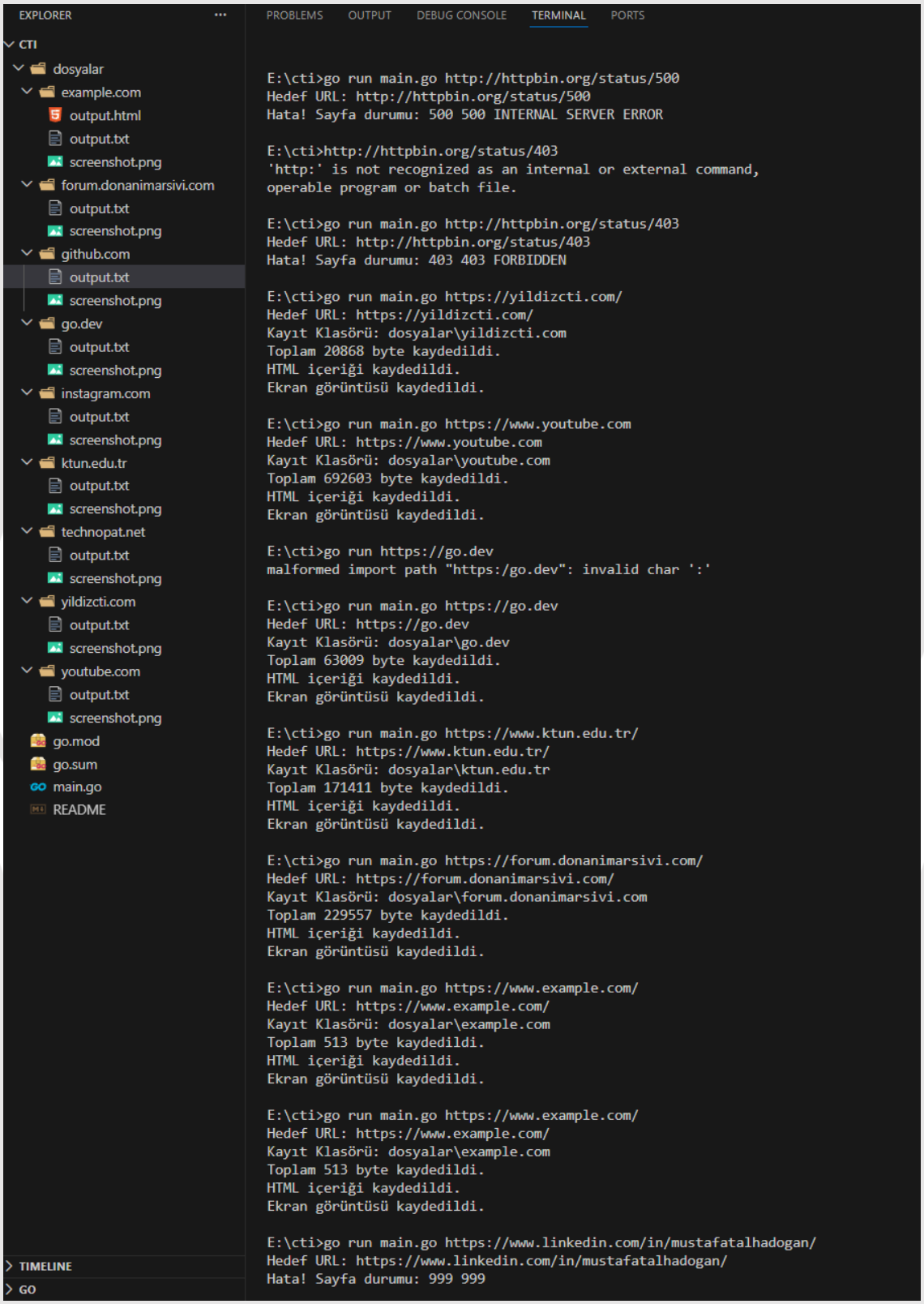
Şekil-1.Proje dizin yapısı, test sırasında oluşturulan dosyalar ve içindeki resim ve html kodları



Şekil-2. Https://github.com/MustafaTalhadgn adresinden çekilen screenshot.png resim dosyası



Şekil-3. Https://github.com/MustafaTalhadgn adresinden alınan html kodları olan output.txt dosyası



Şekil-4. Test sırasında denenen terminal kodları ve çıktılar

Sonuç

Bu proje kapsamında, Siber Tehdit İstihbaratı (CTI) süreçlerinde veri toplama aşamasını desteklemek amacıyla, Go programlama dili kullanılarak web scraper aracı geliştirilmiştir. Geliştirilen yazılım, hedef web sayfalarının HTML kaynak kodlarını eksiksiz çekebilmekte ve sayfanın durumunu belgelemek adına ekran görüntüsü alma işlemini yapabilmektedir.

Yazılımın test etmek amacıyla farklı HTTP durum kodları ve çeşitli hata senaryoları üzerinde testler yapılmıştır. Projenin mevcut kapsamı temel veri çekme işlevlerine odaklandığından; derinlemesine sayfa gezintisi , gelişmiş link filtreleme veya Tor ağı entegrasyonu gibi özellikler bu aşamada kullanılmamıştır.