



WEB SCRAPER

Raporu

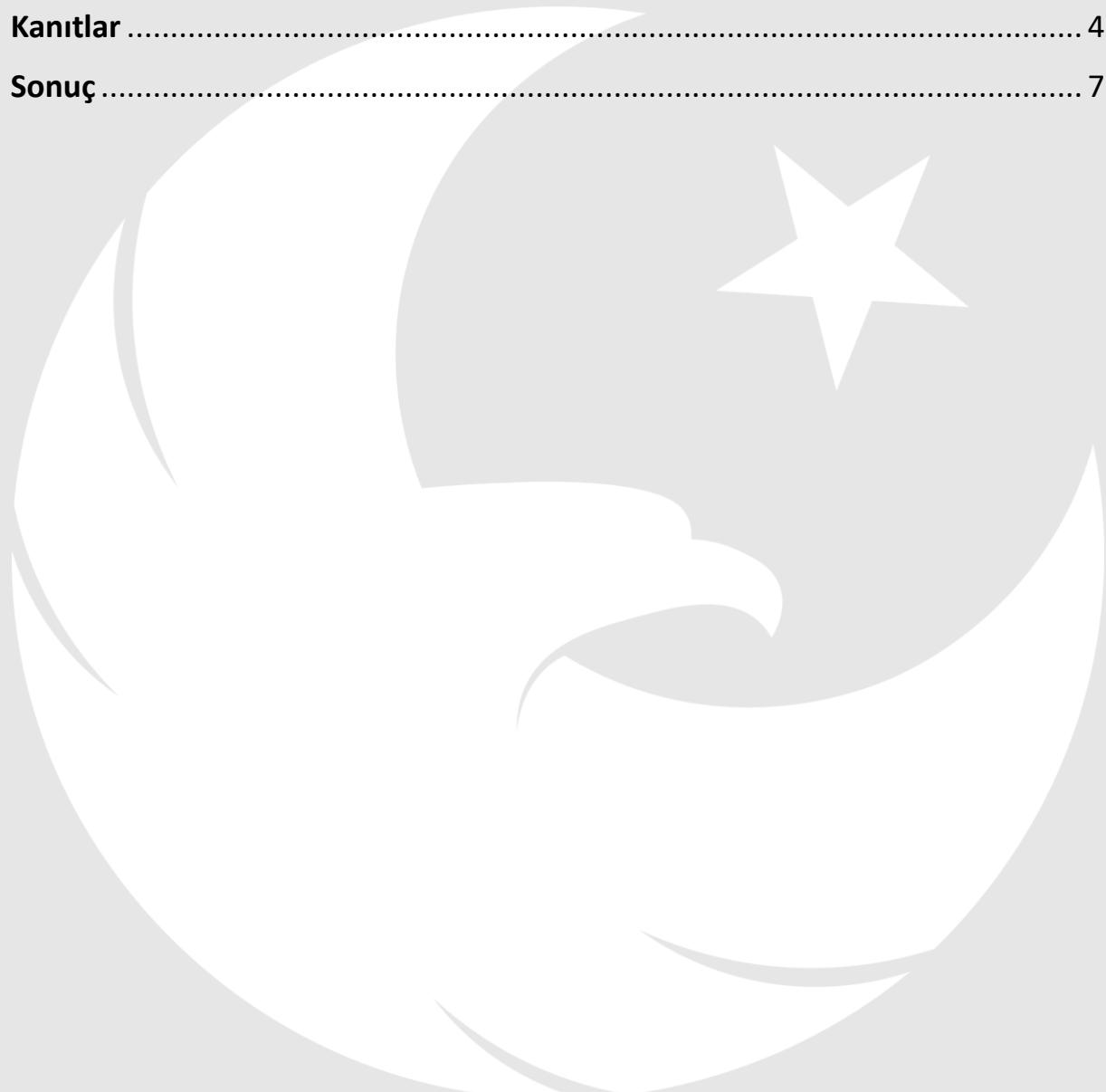
İsim: Mustafa Talha DOĞAN

Grup: Yıldız Cti-Grup2



İçindekiler

Giriş.....	2
Kullanılan Teknolojiler ve Teknik Yaklaşım	2
Test Süreci ve Bulgular	3
Kanıtlar	4
Sonuç	7



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 yüklenikten 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.

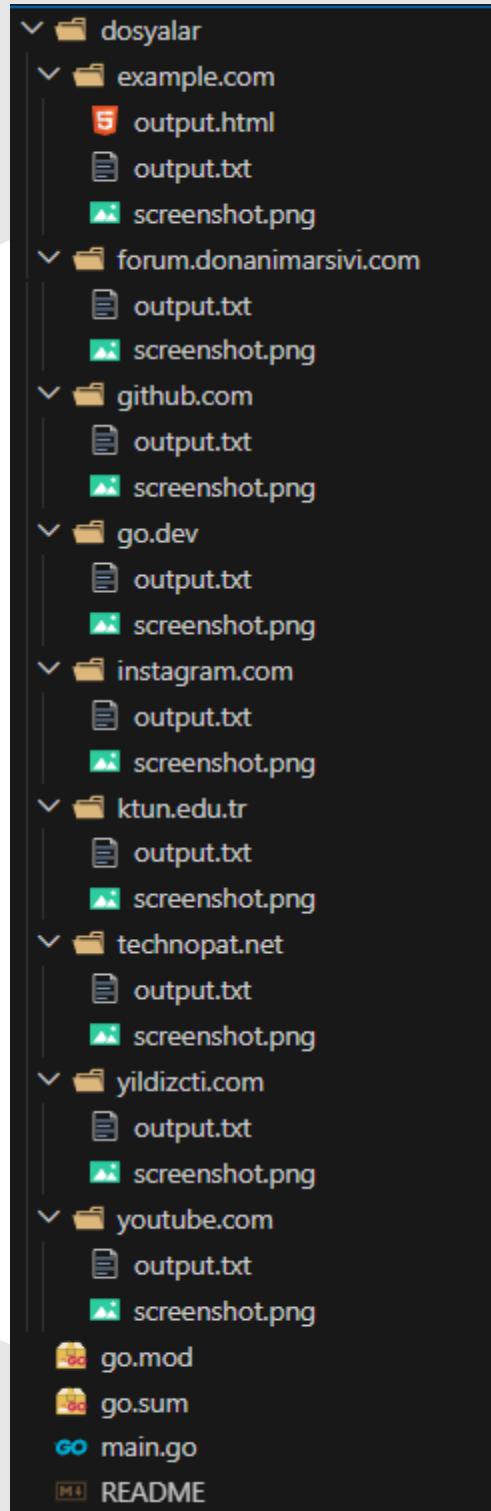
Test Süreci ve Bulgular

#	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/	Başarisız	404	yok
5	http://httpbin.org/status/500	Başarisız	500	yok
6	http://httpbin.org/status/403	Başarisız	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	Başarisız	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



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



Şekil-2. [Https://github.com/MustafaTalhadgn](https://github.com/MustafaTalhadgn) adresinden çekilen screenshot.png resim dosyası

```

dosyalar > github.com > output.txt
1 2
2 3  <!DOCTYPE html>
3 4  <html lang="en"
4 5    data-color-mode="auto" data-light-theme="light" data-dark-theme="dark"
5 6    data-ally-animated-images="system" data-ally-link-underlines="true"
6 7  >
7 8
8 9
9 10  >
10 11
11 12
12 13
13 14
14 15
15 16  <meta charset="utf-8">
16 17  <link rel="dns-prefetch" href="https://github.githubassets.com">
17 18  <link rel="dns-prefetch" href="https://avatars.githubusercontent.com">
18 19  <link rel="dns-prefetch" href="https://github-cloud.s3.amazonaws.com">
19 20  <link rel="dns-prefetch" href="https://user-images.githubusercontent.com/>
20 21  <link rel="preconnect" href="https://github.githubassets.com" crossorigin="anonymous">
21 22  <link rel="preconnect" href="https://avatars.githubusercontent.com" crossorigin="anonymous">
22 23
23 24
24 25
25 26
26 27
27 28
28 29
29 30
30 31
31 32
32 33
33 34
34 35
35 36
36 37
37 38
38 39
39 40
40 41
41 42
42 43
43 44

```

Şekil-3. [Https://github.com/MustafaTalhadgn](https://github.com/MustafaTalhadgn) adresinden alınan html kodları olan output.txt dosyası

EXPLORER ...

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

CTI

- dosalar
- example.com
 - output.html
 - output.txt
 - screenshot.png
- forum.donanimarsivi.com
 - output.txt
 - screenshot.png
- github.com
 - output.txt**
 - screenshot.png
- go.dev
 - output.txt
 - screenshot.png
- instagram.com
 - output.txt
 - screenshot.png
- ktun.edu.tr
 - output.txt
 - screenshot.png
- technopat.net
 - output.txt
 - screenshot.png
- yildizcti.com
 - output.txt
 - screenshot.png
- youtube.com
 - output.txt
 - screenshot.png
- go.mod
- go.sum
- main.go
- README

```
E:\cti>go run main.go http://httpbin.org/status/500
Hedef URL: http://httpbin.org/status/500
Hata! Sayfa durumu: 500 500 INTERNAL SERVER ERROR

E:\cti>http://httpbin.org/status/403
'http:' is not recognized as an internal or external command,
operable program or batch file.

E:\cti>go run main.go http://httpbin.org/status/403
Hedef URL: http://httpbin.org/status/403
Hata! Sayfa durumu: 403 403 FORBIDDEN

E:\cti>go run main.go https://yildizcti.com/
Hedef URL: https://yildizcti.com/
Kayıt Klasörü: dosyalar\yildizcti.com
Toplam 20868 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://www.youtube.com
Hedef URL: https://www.youtube.com
Kayıt Klasörü: dosyalar\youtube.com
Toplam 692603 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run https://go.dev
malformed import path "https://go.dev": invalid char ':'

E:\cti>go run main.go https://go.dev
Hedef URL: https://go.dev
Kayıt Klasörü: dosyalar\go.dev
Toplam 63009 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://www.ktun.edu.tr/
Hedef URL: https://www.ktun.edu.tr/
Kayıt Klasörü: dosyalar\ktun.edu.tr
Toplam 171411 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://forum.donanimarsivi.com/
Hedef URL: https://forum.donanimarsivi.com/
Kayıt Klasörü: dosyalar\forum.donanimarsivi.com
Toplam 229557 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://www.example.com/
Hedef URL: https://www.example.com/
Kayıt Klasörü: dosyalar\example.com
Toplam 513 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://www.example.com/
Hedef URL: https://www.example.com/
Kayıt Klasörü: dosyalar\example.com
Toplam 513 byte kaydedildi.
HTML içeriği kaydedildi.
Ekran görüntüsü kaydedildi.

E:\cti>go run main.go https://www.linkedin.com/in/mustafatalhadogan/
Hedef URL: https://www.linkedin.com/in/mustafatalhadogan/
Hata! Sayfa durumu: 999 999
```

> TIMELINE

> go

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 ağ entegrasyonu gibi özellikler bu aşamada kullanılmamıştır.