

Headless Chrome ile Web Scraping'e Hızlı Bir Bakış

İsmail Batuhan NARCI
Prisync

Ben kimim?

- ▶ İsmail Batuhan NARCI
- ▶ Marmara Üniversitesi, Bilgisayar Mühendisliği
- ▶ Software Developer - Prisync



ismail-batuhan-narci



batuhannarci



Headless browser nedir?

- Headless browser, arayüzü olmadan, komut satırı veya ağ iletişimi kullanılarak bir web sayfasının diğer web tarayıcılarına benzer bir ortamda otomatik olarak kontrol edilmesini sağlar. Bir web sayfasını normal bir tarayıcı gibi anlayıp oluşturabildikleri, Javascript ve AJAX çalıştırabildikleri için web uygulamalarının test edilmesinde kullanılırlar. - Wikipedia

Nerelerde kullanılır?

- ▶ Web uygulamalarını test etmek
- ▶ Ekran görüntüsü yakalamak
- ▶ Web'den bilgi toplamak

Ayrıca,

- ▶ DDOS atak gerçekleştirmek
- ▶ Reklam gösterimlerini arttırma
- ▶ Web sayfalarını istenmeyen şekilde otomatikleştirmek. (Credential stuffing)

Bazı headless browserlar

PhantomJS

- DOM manipulasyonu, CSS selettörleri ve javascript desteği.

Splash

- Birden fazla web sayfasını aynı anda işleyebilme.

Zombie JS

- Node JS ile kullanılıyor.

HtmlUnit

- Java programları için yazılmış.

Headless Chrome

Headless Firefox

Headless Chrome vs PhantomJS

- ▶ ” I think people will switch to it, eventually. Chrome is faster and more stable than PhantomJS. And it doesn't eat memory like crazy. ” - Vitaly Slobodin
- ▶ Daha stabil
- ▶ Daha az kaynak harcıyor
- ▶ PhantomJS eskimiş bir tarayıcı. (Safari 7, macOS 10.9)*
- ▶ PhantomJS de karşılaşılan Javascript hataları yok
- ▶ Normal bir kullanıcı deneyimine daha yakın

*<https://about.gitlab.com/2017/12/19/moving-to-headless-chrome/>

GitLab

Projects

Groups

Activity

Milestones

Snippets

This project

Search

Go

Update all

Cancel

project210

Overview

Repository

Issues

Merge Requests

CI / CD

Wiki

John Doe756 , project210 , Merge Requests

Open 1

Merged 0

Closed 0

All 1

Edit merge requests

New merge request

☒

Search or filter results...

Created date

☒ My title 103

!1 · opened about a minute ago by John Doe757 · feature

updated about a minute ago · 0

Status

Select status

Assignee

Unassigned

Milestone

Select milestone

Labels

Select labels



GitLab

Projects

Groups

Activity

Milestones

Snippets

This project

Search

Go

Update all

Cancel

project1

Overview

Repository

Issues

Merge Requests

CI / CD

Wiki

Snippets

Settings

<< Collapse sidebar

John Doe2 > project1 > Merge Requests

Open 1

Merged 0

Closed 0

All 1

Edit merge requests

New merge request

☐

Search or filter results...

Created date

☐ My title 1

!1 · opened less than a minute ago by John Doe3 · feature

updated less than a minute ago · 0

Status

Select status

Assignee

Select assignee

Milestone

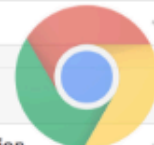
Select milestone

Labels

Select labels

Subscriptions

Select subscription



Test ortamı

- ▶ Ubuntu 16.04
- ▶ Python 3.5.2
- ▶ Google Chrome 65.0.3325.181
- ▶ Selenium 3.8.0 - Browser Automation Tool
pip3 install selenium==3.8.0
- ▶ ChromeDriver 2.37
<https://sites.google.com/a/chromium.org/chromedriver/>

Örnekler

Bir web sayfasının ekran görüntüsünü almak

Bir ürün sayfasını ziyaret edip, ürün ismi ve fiyat bilgisi almak

Bir web sayfasında script çalıştırmak

<https://github.com/batuhannarci/headless-chrome-examples>

Teşekkürler

Sorular?

Çalışma arkadaşları arıyoruz...

► careers@prisync.com