Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

올인원 패키지 Online.

안녕하세요 크롤링 강의 이은찬입니다.

PART1 I 크롤링 기초

크롤링이란 무엇이고 어떻게 하나요?

PART2 I 정적 크롤링

빠르고 간단한 크롤링

PART3 I 동적 크롤링

셀레니움을 활용한 크롤링

Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

[동적 크롤링]
자주 나는 오류 FAQs, 한계
stealth 셀레니움 아닌척 하기

Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

셀레니움의 한계

웹서버들이 자동화된 봇을 막을 수 있음

Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

Stealth 라이브러리

셀레니움 티 안내게 도와줌

Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

Stealth 라이브러리

https://intoli.com/blog/making-chrome-headless-undetectable/

```
 << Cannot establish TLS with client (sni: intoli.com): TlsException("(104, 'ECONNRESET')",)
127.0.0.1:59526: clientconnect
127.0.0.1:59524: clientdisconnect
Successfully injected the content.js script.
127.0.0.1:59526: GET https://intoli.com/blog/making-chrome-headless-undetectable/chrome-headless-test.html
             << 200 OK 1.12k
127.0.0.1:59528: clientconnect
127.0.0.1:59526: GET https://intoli.com/blog/making-chrome-headless-undetectable/modernizr.js
             << 200 OK 2.43k
127.0.0.1:59528: CONNECT intoli.com:443
 << Cannot establish TLS with client (sni: intoli.com): TlsException("(-1, 'Unexpected EOF')",)
127.0.0.1:59530: clientconnect
127.0.0.1:59528: clientdisconnect
127.0.0.1:59526: GET https://intoli.com/blog/making-chrome-headless-undetectable/chrome-headless-test.js
             << 200 OK 2.27k
127.0.0.1:59526: GET https://intoli.com/nonexistent-image.png
             << 404 Not Found 189b
```

There are also a number of errors, but these aren't anything to worry about because they are the result of the client needing to override the certificate errors. Ignoring those, it looks like we saw the expected requests and that the script tag was successfully injected before returning the `chrome-headless-test.html` response.

Finally, let's take a look at the generated `headless-results.png` to verify that we now pass all of the tests.

Test Name	Result
User Agent	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/60.0.3112.50 Safari/537.36
Plugins Length	5
Languages	en-US,en
WebGL Vendor	Intel Open Source Technology Center
WebGL Renderer	Mesa DRI Intel(R) Ivybridge Mobile
Hairline Feature	present
Broken Image Dimensions	20x20

We passed with flying colors!

Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

Stealth 라이브러리

https://pypi.org/project/selenium-stealth/

https://intoli.com/blog/not-possible-to-block-chrome-headless/chrome-headless-test.html

Usage

```
from selenium import webdriver
from selenium_stealth import stealth
import time

options = webdriver.ChromeOptions()
options.add_argument("start-maximized")

# options.add_argument("--headless")

options.add_experimental_option("excludeSwitches", ["enable-automation"])
options.add_experimental_option('useAutomationExtension', False)
driver = webdriver.Chrome(options=options, executable_path=r"C:\Users\DIPRAJ\Programming\adcli

stealth(driver,
        languages=["en-US", "en"],
        vendor="Google Inc.",
        platform="Win32",
        webgl_vendor="Intel Inc.",
        renderer="Intel Iris OpenGL Engine",
        fix_hairline=True,
        )

url = "https://bot.sannysoft.com/"
driver.get(url)
time.sleep(5)
driver.quit()
```