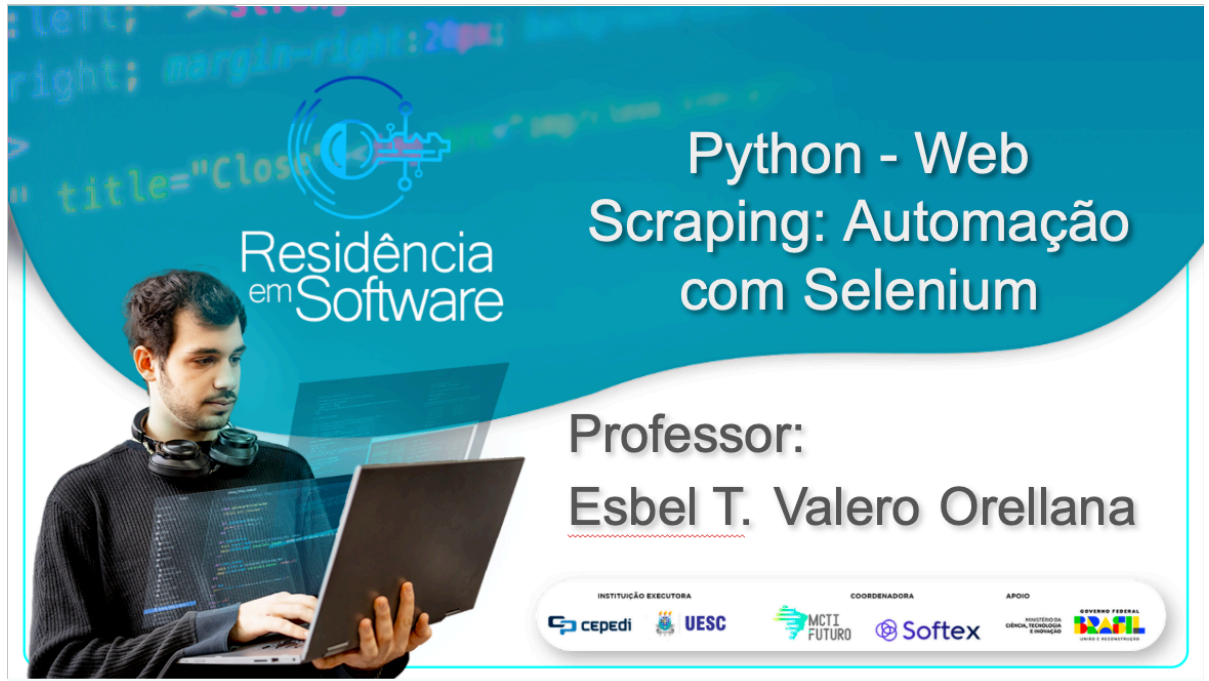


Módulo de Programação Python

Trilha Python - Aula 41/42: Web Scraping: Automação com Selenium



Residência em Software

Python - Web Scraping: Automação com Selenium

Professor:
Esbel T. Valero Orellana

INSTITUIÇÃO EXECUTORA: CEPEDI, UESC
COORDENADORA: MCTI FUTURO, Softex
APOIO: GOVERNO FEDERAL

Automatizando o processo de coleta de dados.

Na aula anterior vimos como usar o **Requests** e o **Beautiful Soup** para extrair informações de requisições feitas a servidores **WEB**.

Entretanto o uso destes recursos nem sempre é suficiente para coletar as informações que desejamos.

Vejamos o seguinte exemplo.

```
In [4]: 1 import requests
2 from bs4 import BeautifulSoup
3
4 url = 'https://www.google.com/search?q=cepedi+ilh%C3%A9us'
5 response = requests.get(url)
6 search = BeautifulSoup(response.text, 'html.parser')
7 print(search.prettify())
```

```
<!DOCTYPE html>
<html lang="pt-BR">
  <head>
    <meta charset="utf-8"/>
    <meta content="AmFMpMe9tdD7tYHZ5DTQG/aRAYYkWIQDI/+Hwz5Y70agTtD
B3LA01RMKtpQ8jCqukCmZ7HDuCpEwx7WwAifBywoAAABYeyJvcmlnaW4iOiJodHR
wczovL2dvd2dsZS5jb20uYnI6NDQzIiwZmVhdHVyZSI6IkxvbmdBbmltYXRpb25
GcmFtZVRpbWluZyIsImV4cGlyeSI6MTcwOTY4MzE5OSwiaXNTdWJkb21haW4iOnR
ydWV9" http-equiv="origin-trial"/>
    <meta content="/images/branding/google/1x/google_standard_co
lor_128dp.png" itemprop="image"/>
    <title>
      cepedi ilhéus - Pesquisa Google
    </title>
    <script nonce="GKtRPgGS0L-918t8fVfeoQ">
      (function(){
document.documentElement.addEventListener("submit",function(b){v
ar a;if(a=b.target){var c=a.getAttribute("data-submitfalse");a="
1"===c||"q"===c&&!a.elements.q.value?!0:!1}else a=!1;a&&(b.preve
ntDefault());b.stopImmediatePropagation();})});document.documentElement
```

Agora vamos procurar pelo painel onde estão os resultados da busca.

```
In [5]: 1 search = search.find('div', class_='g')
2 if search:
3     print(search.prettify())
4 else:
5     print('Nada encontrado')
```

Nada encontrado

A pergunta é: Por que um *tag* que existe na página não pode ser encontrado?

A resposta é relativamente simples: A página capturada pela requisição contém, fundamentalmente, scripts que serão executados do lado do cliente. Ou seja, em muitos casos a página é, de fato, renderizada no browser do cliente.

Vamos utilizar então uma ferramenta para automatizar o processo de coleta, controlando um browser e capturando a página após a renderização.

```
In [6]: 1 #pip install selenium  
        2
```

Aqui o [link \(https://selenium-python.readthedocs.io\)](https://selenium-python.readthedocs.io) para o **Selenium**

Aqui o [link \(https://sites.google.com/chromium.org/driver/\)](https://sites.google.com/chromium.org/driver/) para o Chrome Driver

Ideia por traz de usar esta ferramenta é controlar uma instância de um browser e pegar dela as páginas renderizadas.

Vejamos como usar esta ferramenta.

```
In [7]: 1 from selenium import webdriver
2 from time import sleep
3
4 url = 'https://www.google.com/search?q=cepedi+ilh%C3%A9us'
5
6 # Abre o navegador
7 navegador = webdriver.Chrome()
8
9 navegador.get(url)
10 sleep(10)
11
12 search = BeautifulSoup(navegador.page_source, 'html.parser')
13 #print(search.prettify())
14
15 search = search.find('div', class_='g')
16 if search:
17     print(search.prettify())
18 else:
19     print('Nada encontrado')
```

```
<div class="g" data-hveid="CA8QAA">
  <h2 class="bNg8Rb 0hScic zsYMMe BBwThe" style="clip:rect(1px,1p
x,1px,1px);height:1px;overflow:hidden;position:absolute;white-sp
ace:nowrap;width:1px;z-index:-1000;-webkit-user-select:none">
    Resultado da Web com links de sites
  </h2>
  <div class="BYM4Nd">
    <div class="eKjLze">
      <div class="g">
        <div data-hveid="CA4QAA" data-ved="2ahUKEwj12IHzl56EAXV1ppUC
HeIVBU8QFSgAegQIDhAA" lang="pt">
          <div class="tF2Cxc">
            <div class="yuRUbf">
              <div>
                <span jsaction="rcuQ6b:npT2md;PYDNKe:bLV6Bd;mLt3mc" jsco
ntroller="msmzHf">
                  <a data-ved="2ahUKEwj12IHzl56EAXV1ppUCHeIVBU8QFnoEAcQA
Q" href="https://cepedi.org.br/" jsname="UWckNb" ping="/url?sa=
t&source=web&rct=j&opi=89978449&url=https://cepe
di.org.br/6ahUKEwj12IHzl56EAXV1ppUCHeIVBU8QFSgAegQIDhAA"
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```

```
In [ ]: 1
```