

WORKSHOP WEB SCRAPING

OPENDATA DAY - 07/03/26



O QUE É O PYLADIES?

- Grupo internacional sem fins lucrativos: <https://www.pyladies.com/>
- Focado em aumentar a atividade e a liderança das mulheres na comunidade Python
- Missão: promover, educar e desenvolver uma comunidade de Python plural por meio de divulgação, educação, conferências, eventos e reuniões sociais.

CAPÍTULO SÃO PAULO

- Criado em 7 de setembro de 2015
- Missão: incentivar mulheres a aprender, ensinar e motivar outras a conhecerem Python

VAMOS NOS CONECTAR?



<https://github.com/pyladiessp>



www.linkedin.com/company/pyladiessp



@pyladies.saopaulo



"O POUCO QUE VOCÊ SABE PODE SER
MUITO PARA quem NÃO SABE
NADA!"



E QUEM É VOCÊ?

MINI BIO

- Olá, sou a Carol Cortez! Mãe de dois mimis.
- Sou jornalista por formação e programadora por profissão.
- Trabalhei por mais de dez anos na cobertura de economia em veículos como El País Brasil, Valor Econômico e Infomoney.
- Fiz transição de carreira na pandemia para a área de tecnologia e hoje trabalho como desenvolvedora backend no CPB (Comitê Paralímpico Brasileiro).
- Tenho 40 anos e muito a aprender e conquistar!



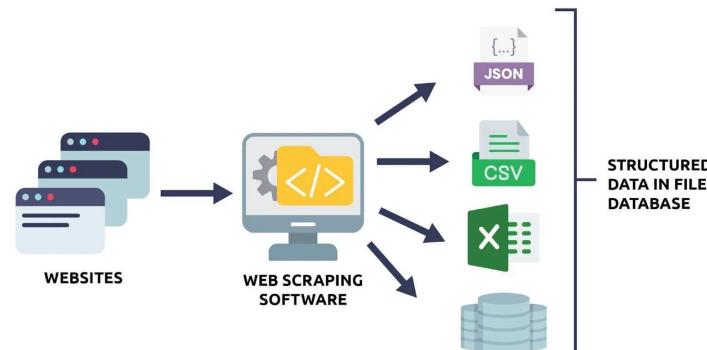
linkedin.com/in/ana-c-cortez



O QUE É WEBSRAPING?

WEB SCRAPING

O webscraping é um termo em inglês que se refere à **raspagem estratégica de dados na internet**, por meio de **extração automatizada** de informações usando bots (softwares) para **coletar dados não estruturados (HTML)** e **convertê-los em formatos estruturados (CSV, JSON, bancos de dados)**.



O DOM DA INTERNET

- Para extrair dados de qualquer site na grande world wide web, precisamos antes entender como funciona a estrutura das páginas da internet.
- As páginas são estruturadas no que chamamos de DOM: Modelo de Objeto de Documento.
- Ele fornece uma representação estruturada do documento como uma árvore e define os métodos que permitem acesso a ela, para que possam alterar a estrutura, estilo e conteúdo do documento.
- Vamos, para resumir didaticamente, chamar o DOM de uma árvore de elementos, que geralmente vêm em HTML.

INSPECIONAR ELEMENTOS



☰ Imprensa Nacional

The screenshot shows the homepage of the Diário Oficial da União. At the top, there's a navigation bar with links to 'Órgãos do Governo', 'Acesso à Informação', 'Legislação', 'Acessibilidade', and 'Acesso GOV.BR'. Below this is a search bar with a placeholder 'PESQUISA' and a button 'PESQUISA AVANÇADA'. To the right of the search bar is a link 'Verificação de autenticidade'. The main title 'Diário Oficial da União' is centered above a grid of three sections: 'SEÇÃO 1: ATOS NORMATIVOS', 'SEÇÃO 2: ATOS DE PESSOAL', and 'SEÇÃO 3: CONTRATOS, EDITAIS E AVISOS'. At the bottom, a dark banner features the text 'DIÁRIO COMPLETO: VERSÃO CERTIFICADA'.

Para acessar o DOM da página, precisamos clicar na tecla F12 (ou clicar no botão direito do mouse sobre a página e em “Inspecionar”)

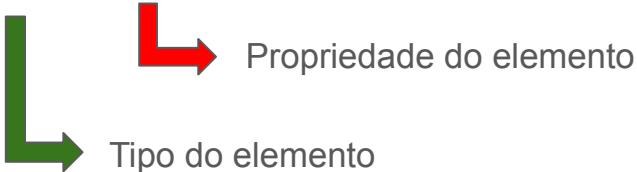
```
<!DOCTYPE html>
<html class="ltr yui3-js-enabled gecko js firefox firefox147 firefox147-1
dir="ltr" lang="pt-BR"> [event] deslocamento overflow)
<div id="yui3-css-stamp" class="" style="position: absolute !important;
!important"></div>
> <head>[...]</head>
<body id="senna_surface1" class=" controls-visible default yui3-skin-sa
signed-out public-page site"> [event]
  <div id="senna_surface1-default" class="flipped" style="display: bloc
  <noscript>[...]</noscript>
    <!--Atalhos para conteúdo-->
    <ul id="quick-access-nav" class="hide-accessible">[...]</ul>
  <div id="wrapper" class="container-fluid">
    <header id="banner" class="banner has-bg" role="banner">
      <div class="bg-primary position-relative">
        <div class="header-gov">
          <div class="bg-white pt-3">[...]</div>
        <nav class="navigation-wrapper">[...]</nav>
      <div class="with-bg" style="background: url(/documents/20181/
home.png/13c3fbff-1d35-11c9-b74f-2a4abf8dc1217t=155598788443
<div id="p_p_id_br_com_seatecnologia_in_buscadou_BuscaDouPo
class="portlet-boundary portlet-boundary_br_com_seatecnolog
portlet-static portlet-static-end portlet-decorate ">
  <span id="p_br_com_seatecnologia_in_buscadou_BuscaDouPort
<section id="portlet_br_com_seatecnologia_in_buscadou_Bus
class="portlet">
  <div class="portlet-content">
    <div class="autofit-float autofit-row portlet-header">
      <div class="portlet-content-container">
        <div class="portlet-body">
          <div class="clearfix journal-content-article " data
```

SELECIONAR ELEMENTOS

- Cada elemento HTML da página ocupa um lugar no DOM. Para selecionar e interagir com esse elemento, precisamos encontrar o caminho dele nessa grande árvore.
- Muitos elementos têm características únicas que nos permite encontrá-los com mais facilidade, como um “id”.
- Se não tiver, localizamos pelo caminho parcial, com a ajuda dos elementos próximos.

LOCALIZADORES

//input[@id='search-bar']



```
#Busca simples  
input_pesquisa = driver.find_element("id",  
"search-bar")  
input_pesquisa.send_keys("material didático")
```

```
<div id="div-search-bar" class="header">  
  <div class="form-group">  
    <div class="input-group"> (flex)  
      <label class="d-none" for="search-bar" style="flex: 1;">Busca</label>  
      <input id="search-bar" class="form-control" value="" type="text" name="search-bar" placeholder="Informe o termo que deseja pesquisar nas Edições do Diário Oficial" style="flex: 1; border-radius: 0; border: none; padding: 0; margin: 0; font-size: 1em; font-weight: bold; font-family: inherit; color: inherit; background-color: transparent; border: none; outline: none; transition: none; width: 100%; height: 100%;"/>  
      <div class="input-group-addon" style="flex: 1; border: none; border-radius: 0; background-color: transparent; width: 100%; height: 100%; position: relative; display: flex; align-items: center; justify-content: center; gap: 10px; padding: 0; margin: 0; font-size: 1em; font-weight: bold; font-family: inherit; color: inherit; background-color: transparent; border: none; outline: none; transition: none; width: 100%; height: 100%;">  
        <span style="font-size: 1.5em; font-weight: bold; font-family: inherit; color: inherit; background-color: transparent; border: none; outline: none; transition: none; width: 100%; height: 100%; position: relative; display: flex; align-items: center; justify-content: center; gap: 10px; padding: 0; margin: 0; font-size: 1em; font-weight: bold; font-family: inherit; color: inherit; background-color: transparent; border: none; outline: none; transition: none; width: 100%; height: 100%;"></span>  
      </div>  
    </div>  
  </div>  
</div>
```

Elementos comuns: div, a, input, button, span, form, tb, td

Propriedades comuns: id, class, text, name, type, placeholder



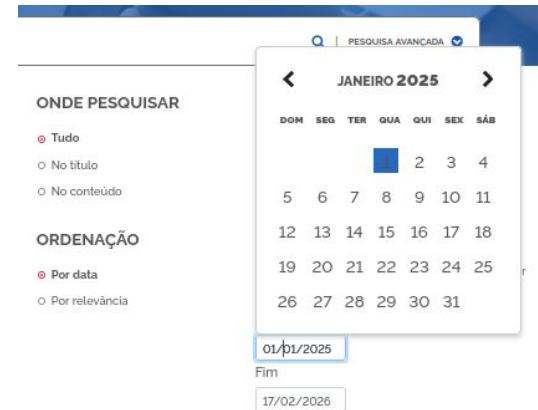
ABRE CAMINHOS - XPATH

CAMINHOS (XPATH)

- Infelizmente, nem tudo são flores: grande parte dos sites não são tão “semânticos”, principalmente com o crescimento de frameworks de front-end, como React e Angular.
- Nesses casos, precisaremos chegar aos elementos pelo “endereço” deles na árvore. O “xpath”.

```
//a[@class='ui-state-default' and normalize-space(text())='1']
```

```
<div class="lfr-spa-loading-bar"></div>
▶ <div class="tooltip fade clay-tooltip-top" role="tooltip" style="display: flex; align-items: center; justify-content: space-between; width: fit-content; margin-left: auto; margin-right: auto; position: absolute; left: 50%; top: 50%; transform: translate(-50%, -50%);">
  <div id="ui-datepicker-div" class="ui-datepicker ui-widget ui-widget-content ui-corner-all search-datepicker" style="position: absolute; left: 660.983px; z-index: 1001; display: block; ">
    <div class="ui-datepicker-header ui-widget-header ui-helper-clearfix ui-corner-all" style="background-color: #f0f0f0; border-bottom: 1px solid #ccc; padding: 5px; font-size: 1em; font-weight: bold; margin-bottom: 5px; ">
      <span>Janeiro 2025</span>
      <span>< </span>
      <span>< </span>
    </div>
    <table border="1" style="width: 100%; border-collapse: collapse; border: none; text-align: center; font-size: 0.8em; ">
      <thead>
        <tr>
          <th>DOM</th>
          <th>SEG</th>
          <th>TER</th>
          <th>QUA</th>
          <th>QUI</th>
          <th>SEX</th>
          <th>SÁB</th>
        </tr>
      </thead>
      <tbody>
        <tr>
          <td>1</td>
          <td>2</td>
          <td>3</td>
          <td>4</td>
          <td>5</td>
          <td>6</td>
          <td>7</td>
        </tr>
        <tr>
          <td>8</td>
          <td>9</td>
          <td>10</td>
          <td>11</td>
          <td>12</td>
          <td>13</td>
          <td>14</td>
        </tr>
        <tr>
          <td>15</td>
          <td>16</td>
          <td>17</td>
          <td>18</td>
          <td>19</td>
          <td>20</td>
          <td>21</td>
        </tr>
        <tr>
          <td>22</td>
          <td>23</td>
          <td>24</td>
          <td>25</td>
          <td>26</td>
          <td>27</td>
          <td>28</td>
        </tr>
        <tr>
          <td>29</td>
          <td>30</td>
          <td>31</td>
          <td></td>
          <td></td>
          <td></td>
          <td></td>
        </tr>
      </tbody>
    </table>
    <div style="text-align: right; margin-top: 5px; ">
      <span>01/01/2025</span>
      <span>Fim</span>
      <span>17/02/2026</span>
    </div>
  </div>
</div>
```



COLINHA DE XPATH MAIS COMUNS

- Por atributo:
HTML: <input name="senha">
XPATH: `//input[@name='senha']`
- Pelo texto visível:
HTML <button>Entrar</button>
XPATH: `//button[contains(text(),'Entrar')]`
- Pelo class, quando não há id:
HTML: <div class="btn primary">
XPATH: `//div[contains(@class,'btn')]`
- Pelos parentes próximos:
HTML:

```
<div class="form">
  <input name="email">
</div>
XPATH: //div[@class='form']//input[@name='email']
```

OBS: O “//” significa que o elemento deve ser encontrado independente da posição dele na árvore. Para respeitar um caminho sequencial, utiliza-se somente “/”

Ex: `//div[contains(text(), 'Resultados')]/p`

Aqui, busca-se um parágrafo (p) que esteja logo abaixo, na árvore do HTML, de uma div que tem o texto “Resultados” dentro dela.

Para buscar por textos, às vezes é melhor ignorar sua formatação. Para isso, tem a função normalize-space:

- Remove espaços no começo e/ou no final
 - Troca vários espaços internos por apenas um
- `//button[normalize-space()='PESQUISAR']`

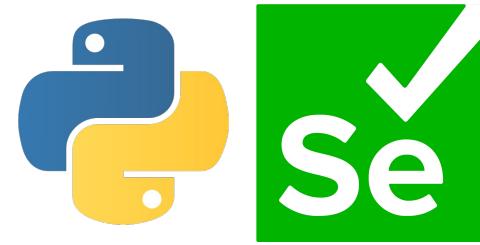
Isso funciona em casos em que o texto dentro do elemento está assim: "
 PESQUISAR
 "



NOSSAS FERRAMENTAS

PYTHON + SELENIUM

- Para este workshop, vamos utilizar a linguagem de programação **Python**, pois é rica em bibliotecas de coleta e análise de dados.
- O **Selenium WebDriver** é uma ferramenta que permite controlar um navegador automaticamente, como se fosse uma pessoa clicando, digitando e navegando.
- Ele abre um navegador real (Chrome, Firefox, Edge...) e executa ações na página, permitindo interações completas.



O Selenium permite:

- Abrir páginas
- Clicar em botões
- Preencher formulários
- Coletar dados da tela
- Testar sistemas



MÃO NA MASSA!!!!!!

BORA PARA O COLAB

https://colab.research.google.com/drive/1Fo2G-E9i8mcHFxXvU-uBN_DNRobHaUCy

CÓDIGO NO GITHUB

No repositório, você vai encontrar uma forma de rodar o código e visualizá-lo interagindo com o browser, pois utilizará uma máquina virtual com auxílio do Docker ;)

<https://github.com/anacarolcortez/workshop-webscraping-docker>



OBRIGADA :)