idp _tech school

WEB SCRAPING:
QUASE UM DELIVERY DE DADOS

RODOLFO VIANA

AGOSTO DE 2023

sobre mim

Olá!

Sou engenheiro de dados, professor do MBA em Jornalismo de Dados do IDP e pós-graduando em Data Science e Analytics pela Esalq-USP.

Por mais de uma década fui repórter, produtor e editor em veículos como Rede Globo, Folha de S.Paulo e Editora Abril. Nesse tempo, me especializei em reportagens orientadas por dados.

Além disso, regularmente colaboro com projetos open-source, geralmente voltados ao ativismo social.

sobre esta aula

Hoje veremos:

- ➔ O que é *web scraping*
- ➔ Qual sua utilidade
- ➔ Como identificamos dados no código-fonte (HTML e CSS) de um site
- ➔ Como coletamos informações de maneira automatizada sem programação

Também faremos um exercício prático.

O que é e para que preciso disso?

Raspagem de dados, ou *web scraping*, é um método de captura de informações disponíveis on-line que se caracteriza pelo uso de automação.

De forma simples, você “terceiriza” o trabalho de milhões de Ctrl C + Ctrl V para um robô.

introdução

O que é e para que preciso disso?

Por exemplo, imagine ter de copiar a agenda do presidente, dia por dia, encontro por encontro [link]...

O que é e para que preciso disso?

...ou ter de percorrer página por página, copiando e colando todos os projetos de lei que tramitam na Câmara [link]...

Ementa: Torna-se obrigatório o uso da placa de recém habilitado durante o período de 4 meses.

17/03/2023 18:19

PL 1240/2023

Autor: Coronel Telhada - PP/SP

Ementa: Dispõe sobre a proibição do uso, fabricação, comercialização, distribuição, posse, depósito e importação de mistura de cola e vidro moído denominado "cerol", linha chilena, linha indonésia, ou de qualquer produto semelhante que possa ou não ser aplicado em linhas de papagaios, pipas, raias, pandorgas ou objetos similares, e dá outras providências.

20/03/2023 11:28

PL 1247/2023

Autor: Flávia Morais - PDT/GO

Ementa: Dispõe sobre a Semana Nacional de Prevenção e Conscientização da Obesidade em crianças e adolescentes.

20/03/2023 17:33

PL 1245/2023

Autor: Juninho do Pneu - UNIÃO/RJ

Ementa: Obriga os restaurantes, lanchonetes, bares e estabelecimentos congêneres a disponibilizar para os consumidores, nos atendimentos presenciais, cardápios impressos em formato físico e dá outras providências.

20/03/2023 17:23

O que é e para que preciso disso?

Para evitar esse trabalho repetitivo, podemos criar uma rotina, um “robô” que faça esse trabalho por nós.

A essa técnica damos do nome de *web scraping*.

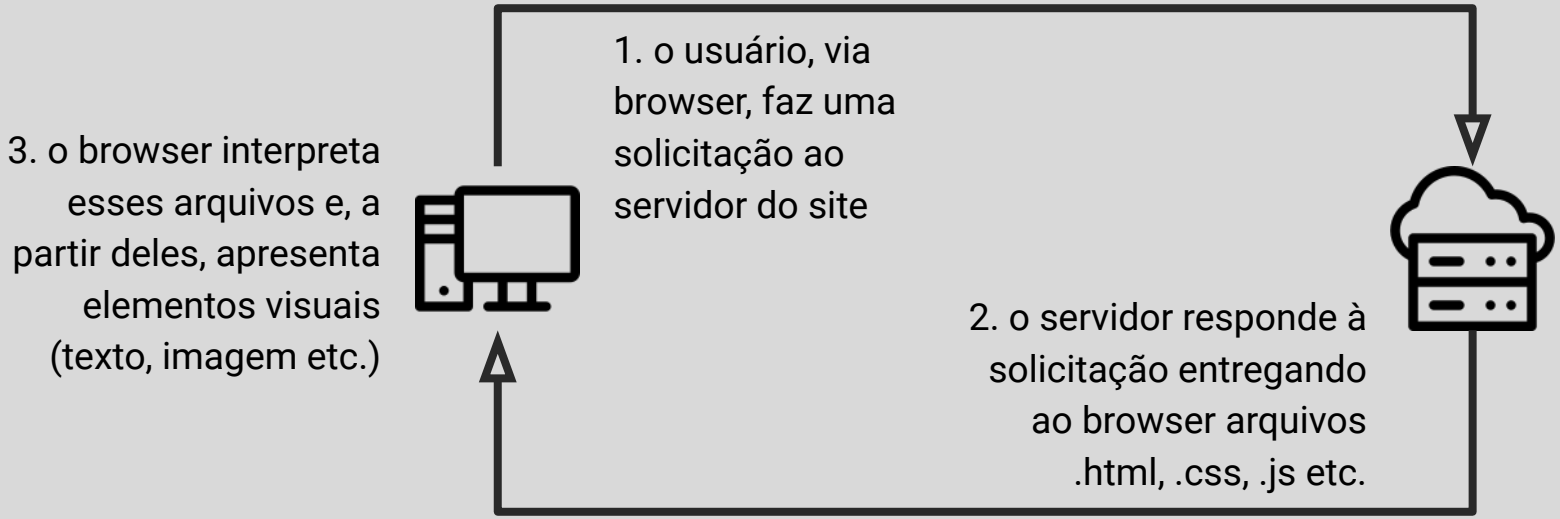
Para fazer *web scraping*, precisamos:

1. Entender minimamente o funcionamento de sites
2. Configurar alguma das ferramentas disponíveis

parte 1

Como funcionam os sites

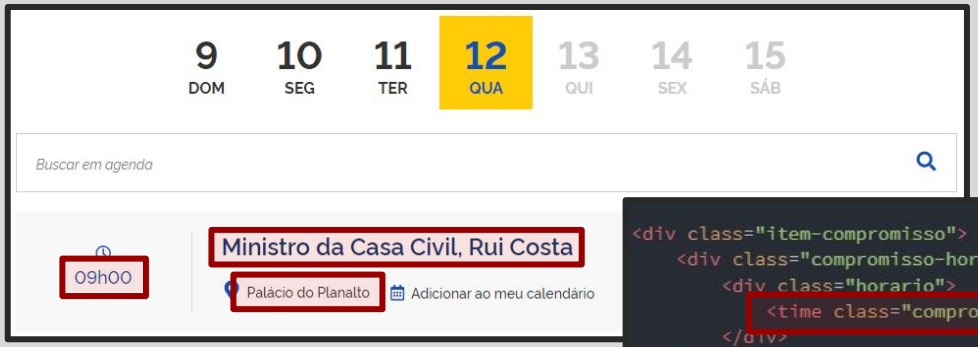
A primeira coisa que precisamos compreender é como funciona a internet:



parte 1

Como funcionam os sites

Exemplo:



visão interpretada pelo browser

visão no arquivo .html

```
<div class="item-compromisso">
    <div class="compromisso-horarios">
        <div class="horario">
            <time class="compromisso-inicio">09h00</time>
        </div>
    </div>
    <div class="compromisso-dados">
        <h2 class="compromisso-titulo">Ministro da Casa Civil, Rui
Costa</h2>
            <div class="compromisso-local">Palácio do Planalto</div>
        </div>
    </div>
</div>
```

parte 1

Um pouco de HTML

Ter o mínimo de entendimento sobre HTML é importante para raspar qualquer site. Exemplo:

Parágrafo
```
<p>...</p>
```

Cabeçalho
```
<h1>...</h1>
```

Link
```
<a>...</a>
```

Contêiner genérico
```
<div>...</div>
```

Lista não ordenada
```
<ul>
  <li>...</li>
  <li>...</li>
</ul>
```

Lista ordenada
```
<ol>
  <li>...</li>
  <li>...</li>
</ol>
```

```
<div>
  <h1>Lista de compras</h1>
  <p>Ver preços no
    <a href="mercado.com">site</a>
  </p>
  <ul>
    <li>Xampu</li>
    <li>Leite</li>
  </ul>
</div>
```

parte 1

Um pouco de HTML

No browser, o código anterior fica assim:

parte 1

Um pouco de CSS

Enquanto HTML cuida da estrutura, do “esqueleto” do site, atributos de CSS dão a aparência. Exemplo:

Estilo

```
<style>
    /* deixa o elemento azul */
    .txt-azul{color: blue;}
    /* usa fonte Arial no elemento */
    .txt-arial{font-family: ‘Arial’;}
    /* deixa o fundo do elemento amarelo */
    .bg-amarelo{background-color: yellow;}
</style>
```

```
<div class=”bg-amarelo”>
  <h1 class=”txt-arial”>Lista</h1>
  <p>Ver preços no
    <a href=“mercado.com”>site</a>
  </p>
  <ul class=”txt-azul”>
    <li>Xampu</li>
    <li>Leite</li>
  </ul>
</div>
```

parte 1

Um pouco de CSS

No browser, o código similar ao anterior fica assim:

Conclusão

Quando olhamos o código de um site, vemos que as informações que queremos estão dentro de tags de HTML, com atributos de CSS.

Isso significa que uma tag ou atributo funciona como localizador para os dados que queremos coletar. E não importa se há um ou 1 milhão de dados: o sistema vai retornar todos os valores que têm tal tag ou atributo.

Isso é raspagem de dados: encontrar padrões de HTML e CSS para capturar os valores neles contidos.

parte 1

À prática

1. Abra a agenda do presidente, disponível em https://bit.ly/agenda_lula.
2. Posicione o mouse sobre o horário de um evento e, com o botão direito, clique em *Inspecionar* ou *Inspect*.
3. Observe em quais tags e atributos a informação está contida.
4. Repita a operação, mas com o compromisso.

parte 2

Ferramentas

Há diversas formas de automatizar a captura de dados, e cada forma tem o seu grau de complexidade e custo.

Programação



Prós

- ➔ Baixo custo
- ➔ Personalizado

Contra

- ➔ Elevado grau técnico

Aplicações privadas

Pró

- ➔ Baixo grau técnico

Contra

- ➔ Custo elevado ou limitações

parte 2

Web Scraper

Uma ferramenta poderosa, gratuita e ubíqua é Web Scraper. Trata-se de uma extensão do navegador Google Chrome e Firefox. (A versão para Firefox, contudo, está desatualizada.)

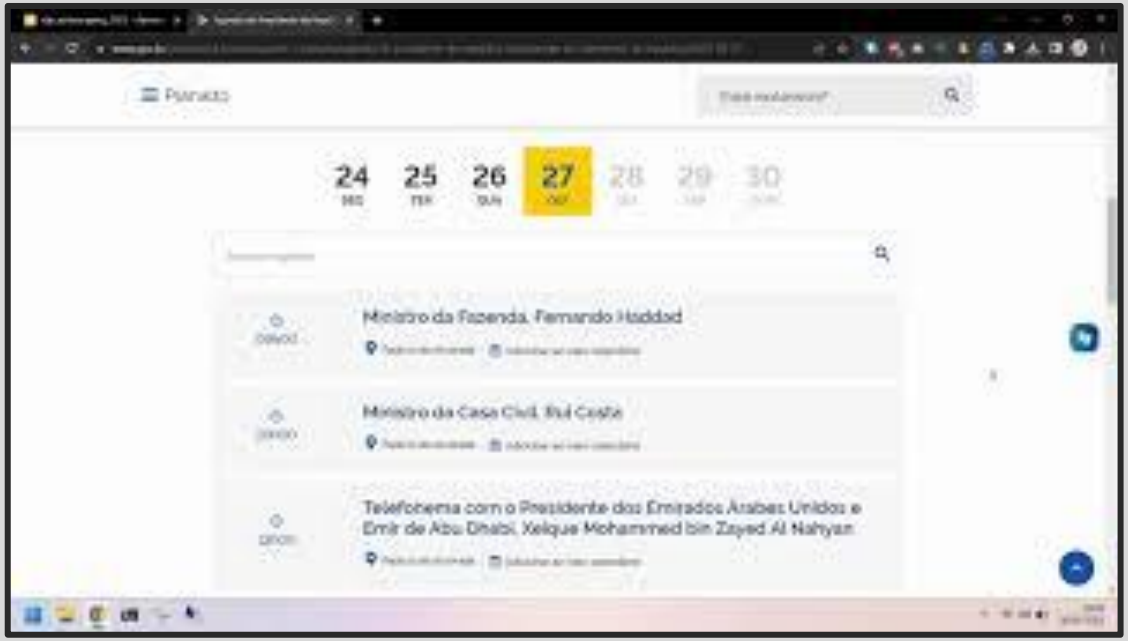
Como não é um programa instalado, funciona em Windows, MacOS e Linux.

Vamos trabalhar com ele.

parte 2

Web Scraper

Exemplo de como raspar um site em 2 minutos.



parte 2

À prática

1. Instale no Google Chrome o webscraper.io, disponível em https://webscraper.io.
2. Abra a agenda do presidente, disponível em https://bit.ly/agenda_lula e escolha um dia.
3. Posicione Ctrl + Shift + i (Windows e Linux) ou Cmd + Shift + i (MacOS).
4. Navegue até a aba *Web Scraper*.
5. Acompanhe o instrutor.

material extra

Continue o aprendizado

textos HTML Básico
https://developer.mozilla.org/pt-BR/docs/Learn/HTML

textos CSS Básico
https://developer.mozilla.org/pt-BR/docs/Learn/CSS

textos Documentação do webscraper.io
https://www.webscraper.io/documentation/

vídeos Tutorial do webscraper.io
https://webscraper.io/tutorials

Para continuar o papo...

- linkedin.com/in/rodolfoviana
- eu@rodolfoviana.com.br
- MBA em Jornalismo de Dados do IDP


idp
_tech
school