RASPAGEM DE DADOS

Rodolfo Viana

introdução

O que é e para que preciso disso?

Raspagem de dados, ou *webscraping*, é um método de captura de informações caracterizado pelo uso de automação.

De forma simples, podemos dizer que, em vez de dar milhões de Ctrl C + Ctrl V em um site, você pode “terceirizar” o serviço para um robô.

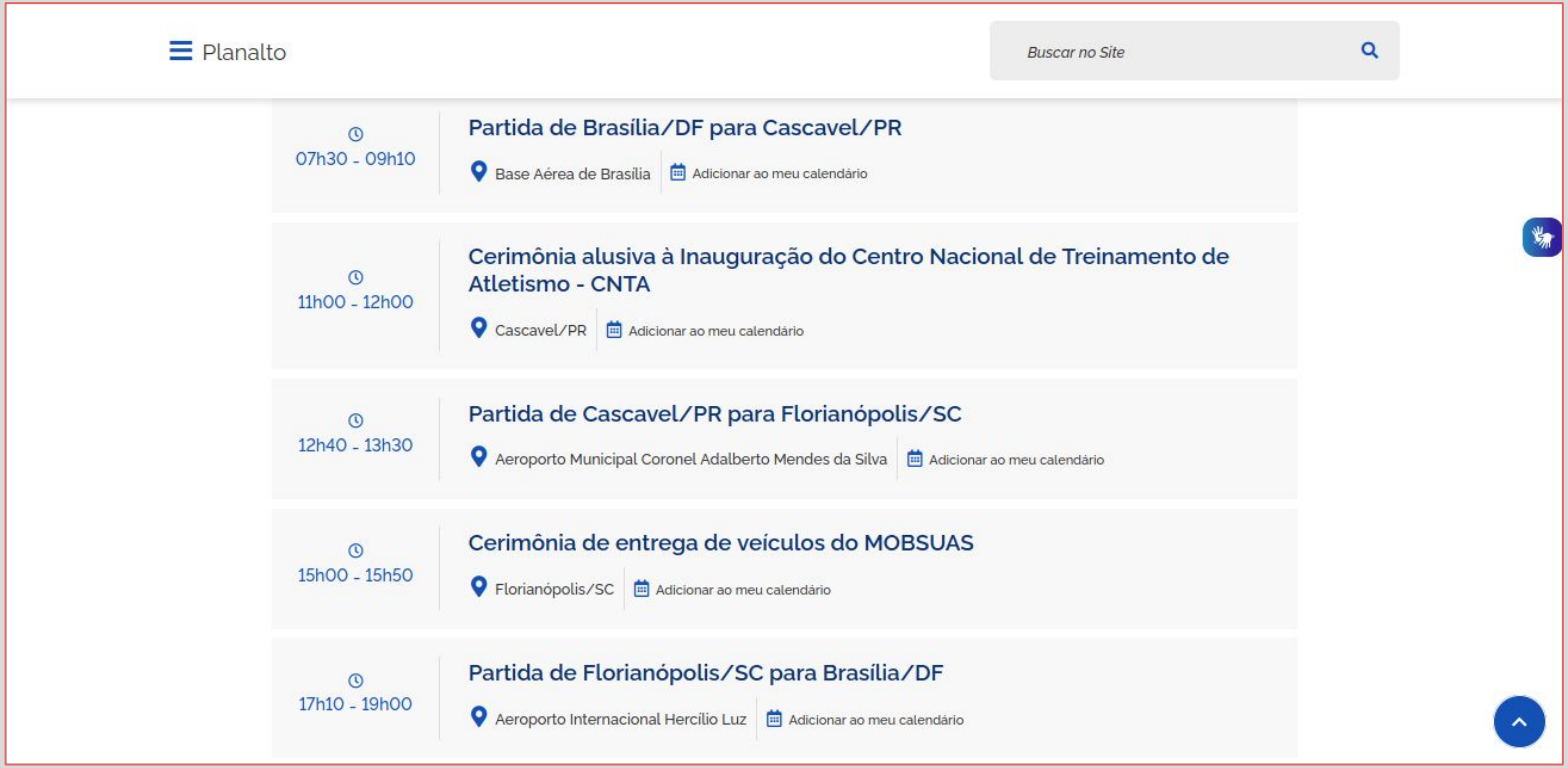
introdução

O que é e para que preciso disso?

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

introdução

O que é e para que preciso disso?



introdução

O que é e para que preciso disso?

...ou ter de clicar hospital por hospital para saber quantos dos estabelecimentos com UTI para covid-19 são administradas por prefeituras, quanto por estados, quantos pelo governo federal...

introdução

O que é e para que preciso disso?

Indicadores - Leitos
Estado - Todos
Município - Todos
Tipo Leito - Complementar - SUPORTE VENTILATÓRIO PULMONAR - COVID-19

CNES	Estabelecimento	Existentes	SUS
2208857	AISI HOSPITAL DE CLINICAS DE ITAJUBA	30	30
7607547	ASSOCIACAO BENEFICENTE OSWALDO CRUZ DE HORIZONTINA	4	0
2200457	ASSOCIACAO MARIO PENNA	6	0
2023016	CASA DE CARIDADE SAO VICENTE DE PAULO CAJURU	2	0
0211125	CENTRO DE SAUDE DE XIQUE XIQUE	1	0
0102652	CENTRO INTENSIVO DE COMBATE AO CORONAVIRUS CICC	16	0
0118095	CENTRO INTERMEDIARIO DE ENFRENTAMENTO AO CORONAVIRUS CIEC	12	0
2399717	COMPLEXO DE DOENCAS INFECTO CONTAGIOSAS CLEMENTINO FRAGA	21	0
2716097	COMPLEXO HOSPITALAR IRMA DULCE O S S	10	0
0026840	COMPLEXO HOSPITALAR SAO FRANCISCO	10	0
0105058	COVID 19 CENTRO DE COMBATE AO CORONAVIRUS CCC JANDIRA	10	0
2205998	FHAHC	5	0
2232146	FUNDACAO DE SAUDE PUBLICA DE NOVO HAMBURGO FSNH	7	0
2232030	FUNDACAO DE SAUDE PUBLICA SAO CAMILO DE ESTEIO	4	0
7424981	FUNDACAO HARRY GUIDO GREIPEL	3	0
2139049	HEFA	5	0
2449641	HGM HOSPITAL GERAL MUNICIPAL DR MARCOLINO JR	10	0
2324172	HOSP EST DE CANTO DO BURITI	1	1
2324261	HOSP EST JOSE FURT DE MENDONCA	1	0
2323583	HOSP EST JULIO HARTMAN	3	0
2324334	HOSP LOCAL DE DEMERVAL LOBAO	2	0
2324288	HOSP LOCAL DE LUZILANDIA	2	0
2777746	HOSP REG CHAGAS RODRIGUES	2	2

O que é necessário para fazer isso?

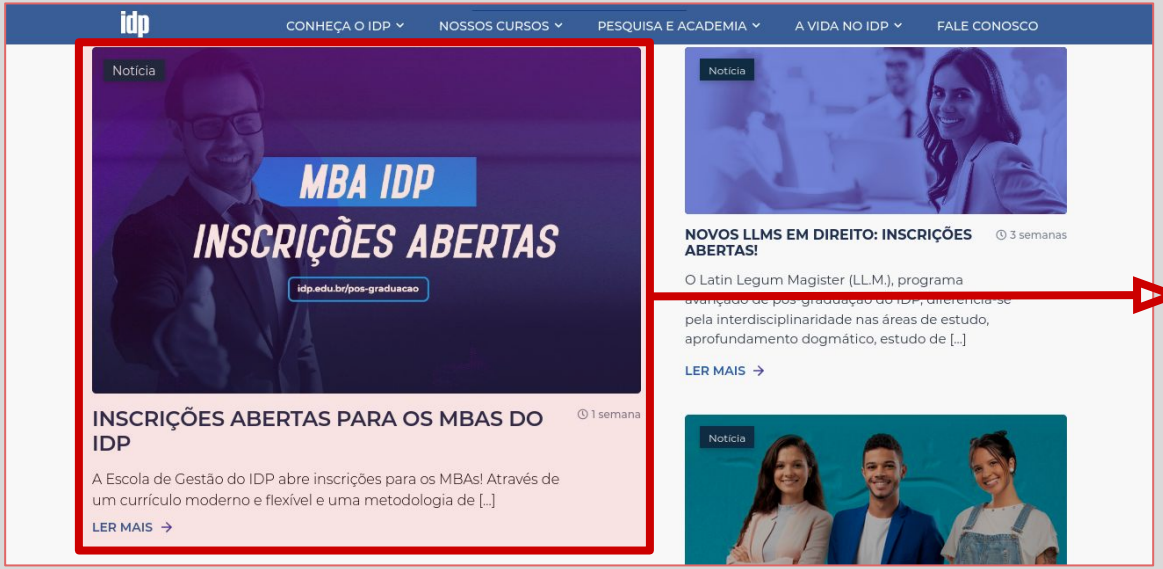
Duas coisas:

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

parte 1

O esqueleto de um site

Quando você visita um site, vê a parte “interpretada” de códigos HTML, o “esqueleto”, e CSS, a “aparência” do site.




```
<a href="https://www.idp.edu.br/inscricoes-abertas-para-os-mbas-do-idp/">
  <figure class="imgPrimary">
    <img width="1000" height="630" src="https://wpcdn.idp.edu.br
/idpsiteportal/2020/11/934fa94c-1401-491c-b37c-96036339a295.jpg"
class="mainImg img-fluid wp-post-image"/>
    <div class="blogCategoria">
      <span>Notícia</span>
    </div>
  </figure>
  <div class="blogInfoPri">
    <div class="postTime">
      <p><i class="fa fa-clock-o" aria-hidden="true"></i>1 semana</p>
    </div>
    <h2>Inscrições abertas para os MBAs do IDP</h2>
    A Escola de Gestão do IDP abre inscrições para os MBAs! Através de um
currículo moderno e flexível e uma metodologia de [...]
    <a class="ver-mais-blog" href="https://www.idp.edu.br/inscricoes-
abertas-para-os-mbas-do-idp/">LER MAIS
      <img src="https://www.idp.edu.br/wp-content/themes
/idp_principal_2020/img/arrowBlog.png"/></a>
  </div>
</a>
```

parte 1

Um pouco de HTML

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

`<p>...</p>`
parágrafo

```
<p>Texto qualquer</p>
```

`<h1>...</h1>`
cabeçalho

```
<h1>Título do texto</h1>
```

```
<ul>
  <li>...</li>
</ul>
```
lista

```
<ul>
  <li>Item 1</li>
  <li>Item 2</li>
</ul>
```

`<div>...</div>`
caixa de conteúdo

```
<div>
  <h1>Título</h1>
  <p>Parágrafo simples</p>
  <ul>
    <li>Item 1 da lista</li>
    <li>Item 2 da lista</li>
  </ul>
</div>
```

parte 1

Um pouco de CSS

Enquanto HTML cuida da estrutura, CSS dá a aparência do site. Exemplos:

`styles.css`
folha de estilos

```
.txt-red {color: red;}

#comic-blue {
  color: blue;
  font-family: ‘Comic Sans’, display;
}
```

`<p class=“txt-red”>...</p>`
parágrafo com texto em vermelho

```
<p class=“txt-red”>Texto qualquer</p>
```

`<p id=“comic-blue”>...</p>`
parágrafo com texto em azul e Comic Sans

```
<p id=“comic-blue”>Texto qualquer</p>
```

parte 1

Conclusão

Quando olhamos o código de um site, vemos que as informações que queremos estão em tags 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.

parte 1

À prática

1. Abra a agenda do presidente, disponível em https://bit.ly/3b04wdF.
2. Posicione o mouse sobre o horário e, com o botão direito, clique em *Inspecionar* ou *Inspect*.
3. Observe quais tags e atributos “rodeiam” a informação.
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 ou 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.

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

Vamos trabalhar com ele.

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/3b04wdF.
3. Posicione Ctrl + Shift + i (Windows e Linux) ou Cmd + Shift + i (MacOS).
4. Navegue até a aba *Web Scraper*.
5. Acompanhe com 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

vídeos Curso de HTML e CSS
https://www.youtube.com/playlist?list=PLwgL9IEA0PxUjbhob9UMdpVq12sGrjgU6

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

Para continuar o papo...

- linkedin.com/in/rodolfoviana
- twitter.com/rodolfoviana
- eu@rodolfoviana.com.br