

Web Scraping com Puppeteer

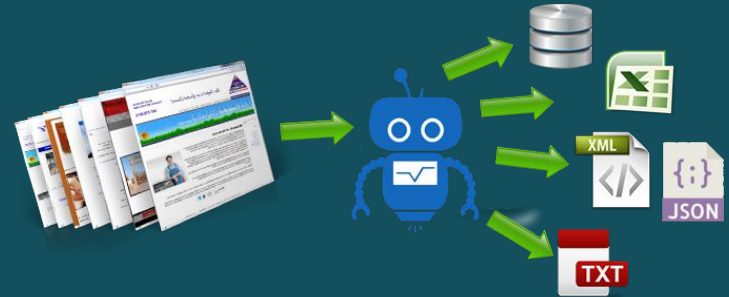
Consoma sites *client side* de forma simples

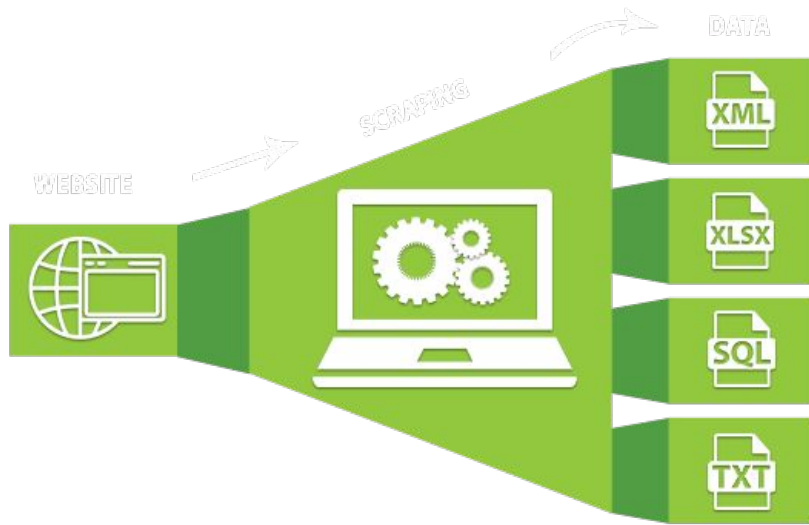
Mas....

- O que é WebScraping?
- O que são sites *client side*?
- O que é o Puppeteer?



WEB SCRAPING





**Técnica de
extração de
dados utilizada
para coletar
dados de sites**

“É possível fazer o mesmo processo manualmente, mas quando se fala de Web Scraping a ideia é automatizar o trabalho.”

[Daniel Moraes]

Formas de Uso



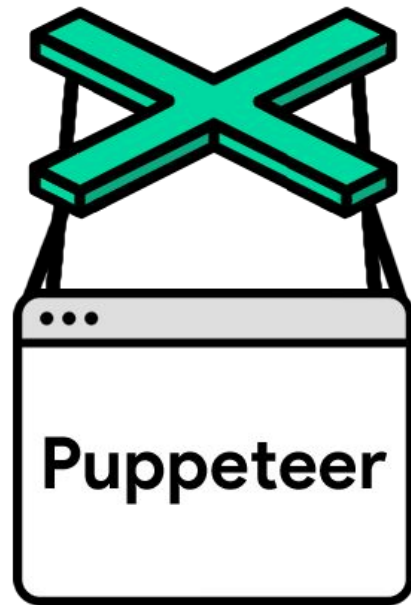
Store	Casa Behar Parque de Turismo e Comércio Ltda.	RECOMENDADA ONE CAMBIO e TURISMO	PANEMA EXCHANGE
05	R\$ 4,07 atualizado há 6 horas	R\$ 4,08 atualizado há 2 horas	R\$ 4,08 atualizado há 1 hora
IS	VEJA MAIS	VEJA MAIS	VEJA MAIS
ça	4 CANTOS SERVIÇOS DE TURISMO	ULTRAMAR SERVIÇOS • TURISMO • SEGUROS	RECOMENDADA Habitatar ADMINISTRAÇÃO
09	R\$ 4,10	R\$ 4,10	R\$ 4,10



Ferramentas de WebScraping



Colly



Scrapy

鋸

Nokogiri

Jaunt



IRON WEBSCRAPER

AIOHTTP

Sites

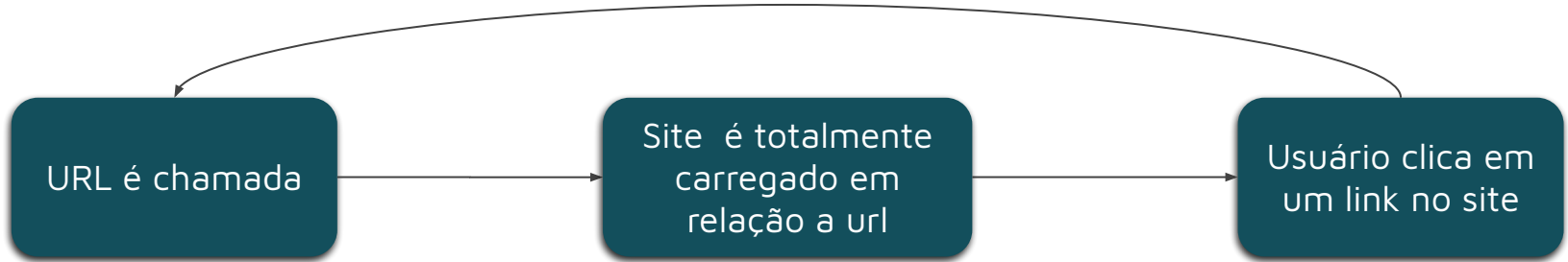
Client Side

Sites processados apenas e
diretamente pelo browser

The JavaScript logo, consisting of the letters 'JS' in a bold, black, sans-serif font, centered on a bright yellow square background.

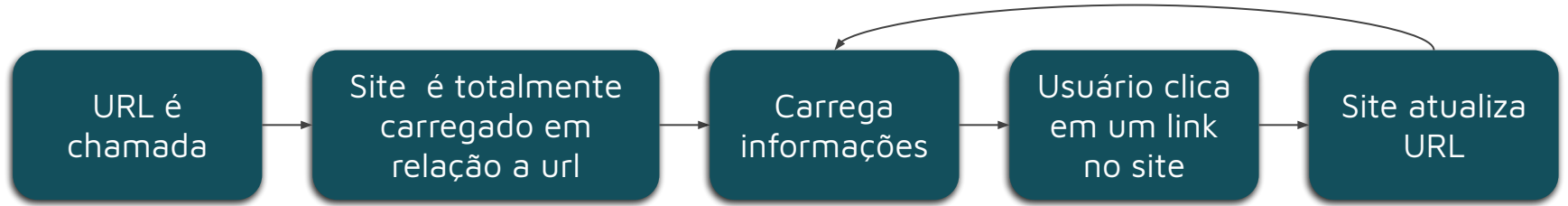
JS

Server Side



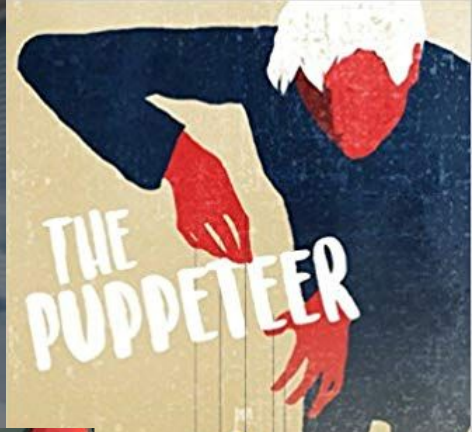
axios + cheerio

Client Side



PUPPETEER





MARIONETISTA



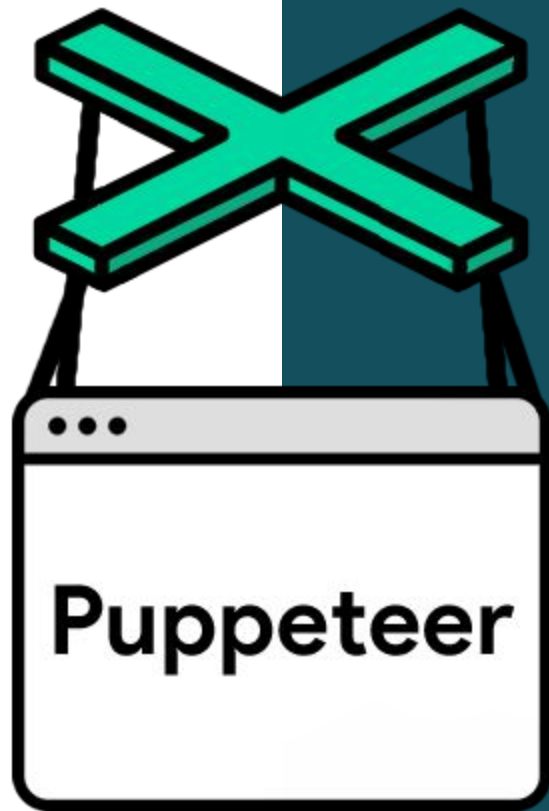
**Biblioteca de NodeJS que fornece uma
API de alto nível para controlar o Chrome
ou o Chromium através do protocolo
DevTools**

★ 50,297

👤 221

🕒 1,494

🔗 4,604



Quem mantém o Puppeteer?



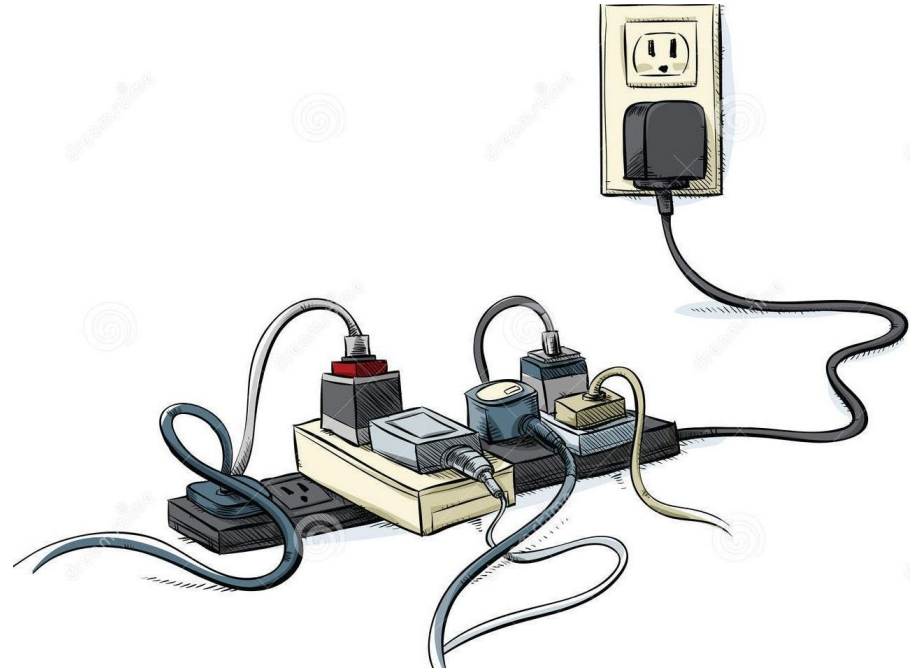
Chrome
Developer Tools

Vantagens do Puppeteer



Fornecer uma biblioteca canônica compacta que destaca os recursos do protocolo DevTools

Quase zero de
sobrecarga de
desempenho em uma
página automatizada





Não requer
configuração e vem
junto com a versão
do Chromium com a
qual ele funciona
melhor, facilitando
muito o início

Pode ser executado
ou não no formato
headless





COMO INSTALAR?



```
# Com download do chromium
```

```
npm i puppeteer
```

```
# ou
```

```
yarn add puppeteer
```



```
# Sem download do chromium
```

```
npm i puppeteer-core
```

```
# ou
```

```
yarn add puppeteer-core
```

COMO USAR?



Uso básico

```
const puppeteer = require('puppeteer');

(async () => {
  const browser = await puppeteer.launch();
  const page = await browser.newPage();

  await page.goto('https://google.com');
  await page.screenshot({path: 'example.png'});
  await browser.close();
})();
```



Try Puppeteer

Try Puppeteer v1.9.0

Run an example: `screenshot.js`



```
1 const browser = await puppeteer.launch();
2
3 const page = await browser.newPage();
4 await page.goto('https://example.com');
5
6 console.log(await page.content());
7 await page.screenshot({path: 'screenshot.png'});
8
9 await browser.close();
10
```

RUN IT

LOG

```
<!DOCTYPE html><html><head>
<title>Example Domain</title>

<meta charset="utf-8">
<meta http-equiv="Content-type" content="text/html; charset=utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<style type="text/css">
body {
  background-color: #f0f0f2;
```

RESULT

Example Domain

This domain is established to be used for illustrative examples in documents. You may use this domain in examples without prior coordination or asking for permission.

[More information](#)

Protip: you can use `async/await` directly in the editor.

LEGALIDADE DO WEB SCRAPING





WEB SCRAPING É LEGAL OU ILEGAL?

- Tem se tornado uma prática maliciosa utilizada por criminosos para roubar conteúdos protegidos e cometer fraudes;
- Muitas vezes é feito com total desconsideração das leis de direitos autorais e dos Termos de Serviço;
- Usado para contornar medidas de segurança;
- “Não há nada que proíba uma empresa de lhe processar”;



FREE SOCCER

API grátis com resultados de competições nacionais de futebol

- 22 campeonatos
- 7 países
- 6 portais consumidos

Ferramentas:

- NodeJS
- Mongoose
- Cheerio
- Puppeteer

Name	Country	Sex	Year	Results	Table	Statistics	Flags
Brasileirão Série A	Brazil	M	2012 - 2018	✓	✓	✗	✓
Brasileirão Série B	Brazil	M	2012 - 2018	✓	✓	✗	✓
Copa do Brasil	Brazil	M	2012 - 2018	✓	---	✗	✓
Copa do Brasil Sub-20	Brazil	M	2012 - 2018	✓	---	✗	✓
Copa do Brasil Sub-17	Brazil	M	2013 - 2018	✓	---	✗	✓
Copa Verde	Brazil	M	2014 - 2018	✓	---	✗	✓
La Liga	Spain	M	2013/2014 - 2018/2019	✓	✓	✗	✗
La Liga Segunda División	Spain	M	2013/2014 - 2018/2019	✓	✓	✗	✗
Primera División Femenina	Spain	F	2013/2014 - 2018/2019	✓	✓	✗	✗
Premier League	England	F	2000/2001 - 2018/2019	✓	✗	✗	✗
Ligue 1	France	M	2000/2001 - 2018/2019	✓	✓	✗	✓
Ligue 2	France	M	2000/2001 - 2018/2019	✓	✓	✗	✓
Coupe Ligue	France	M	2000/2001 - 2018/2019	✓	---	✗	✓
Serie A	Italy	M	2004/2005 - 2018/2019	✓	✓	✗	✓
Bundesliga	Germany	M	2000/2001 - 2018/2019	✓	✓	✗	✓
2 Bundesliga	Germany	M	2000/2001 - 2018/2019	✓	✓	✗	✓
3 Liga	Germany	M	2008/2009 - 2018/2019	✓	✓	✗	✓
Allianz Frauen-Bundesliga	Germany	F	2000/2001 - 2018/2019	✓	✓	✗	✓
2 Frauen-Bundesliga	Germany	F	2018/2019	✓	✓	✗	✓
Liga NOS	Portugal	M	2009/2010 - 2018/2019	✓	✓	✗	✗
Ledman LigaPro	Portugal	M	2009/2010 - 2018/2019	✓	✓	✗	✗



/andreImlins/freesoccer



ANDRÉ LINS

- Desenvolvedor FrontEnd ReactJS na Softplan
- Graduado em Ciência da Computação pela UFRPE
- Pós-Graduando em Engenharia de Software pela PUC Minas
- Viciado em programação
- Fundador do Projeto N.A.D.A.
- Tentando não ser evangelista Javascript



@andrelmlins

