



SMI 2018 :: web scraping

parte I

Augusto Fadel

6 de novembro de 2018

parte I

- R e RStudio
- Web data e web scraping
- Coleta de dados na web
 - download
 - API
 - web scraping

sugestões

- Repositório no github: [web-scraping-smi2018](#)
- IDE: [RStudio](#)
- Pacotes: [tidyverse](#) | [rvest](#) | [httr](#) | [xml2](#) | [jsonlite](#) | [DBI](#)
- Outros recursos:
 - [R for Data Science](#)
 - [Happy Git and GitHub for the useR](#)
 - [rstudio::conf 2018](#) | [webinars](#) | [cheat sheets](#)
 - [DataCamp](#) | [edX](#) | [Cousera](#) | [Udacity](#)
 - [Stackoverflow](#)

Web Scraping

web scraping



web data

- Download
- APIs (Application Programming Interfaces)
- Web scraping
- Crawler

	t	iso3_j	iso3_i	hs92	v	gdp_i
3174	1995	AGO	BRA	847982	2024	768951320576
3254	1995	AGO	BRA	401199	4183	768951320576
3258	1995	AGO	BRA	691010	8777	768951320576
3287	1995	AGO	BRA	441211	2303	768951320576
3305	1995	AGO	BRA	841850	2849	768951320576
3306	1995	AGO	BRA	846592	10312	768951320576
3322	1995	AGO	BRA	842630	3353	768951320576
3333	1995	AGO	BRA	420330	2270	768951320576
3350	1995	AGO	BRA	392620	3779	768951320576
3360	1995	AGO	BRA	902780	1877	768951320576
3369	1995	AGO	BRA	721690	2230	768951320576

R for Data Science (Inglês) Capa Comum – 20 jan 2017

por [Garrett Wickham](#) (Autor), [Garrett Grolemund](#) (Autor)

★★★★★ 7 avaliações de clientes

> [Ver todos os 2 formatos e edições](#)

eBook Kindle
R\$ 82,99

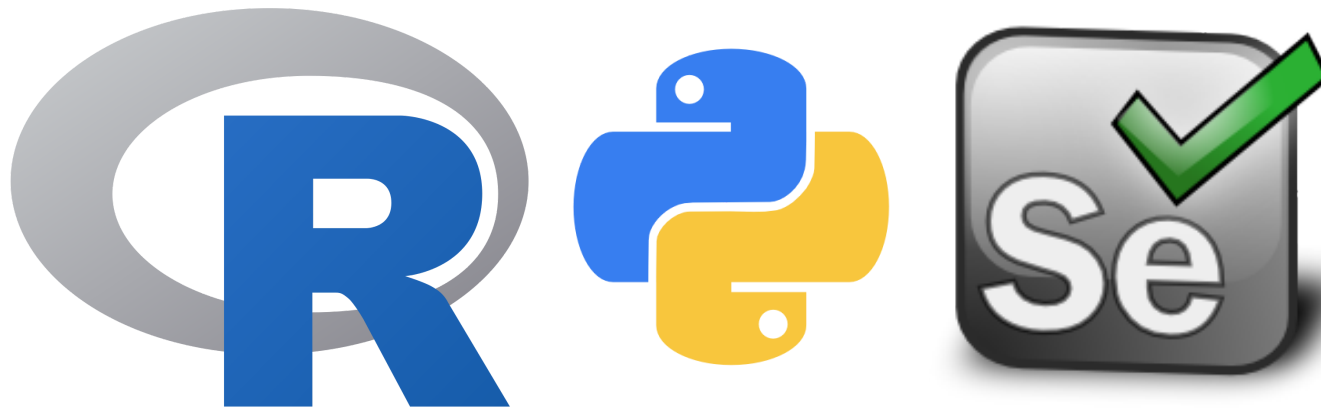
Capa Comum
R\$ 142,21

[Leia com nossos apps gratuitos](#)

1 Novo(s) a partir de R\$ 142,21

Em até 4x R\$ 35,56 sem juros [Calculadora de prestações](#)

web data



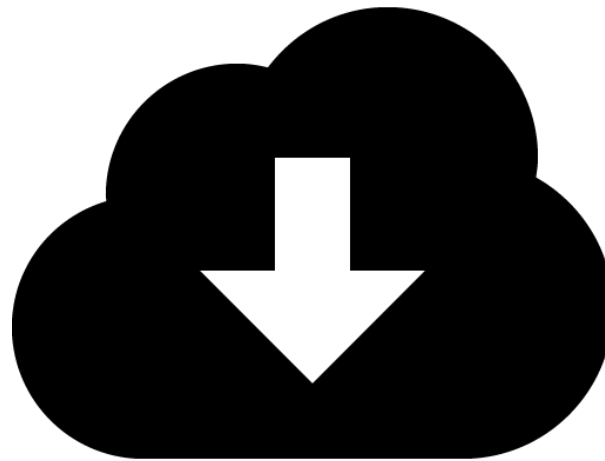
Download

tipos de arquivos

tipos de arquivos

- csv (comma-separated values)
- tsv (tab-separated values)
- MS Excel (xls, xlsx)
- zip
- [JSON](#) (JavaScript Object Notation)
- [XML](#) (Extensible Markup Language)

download



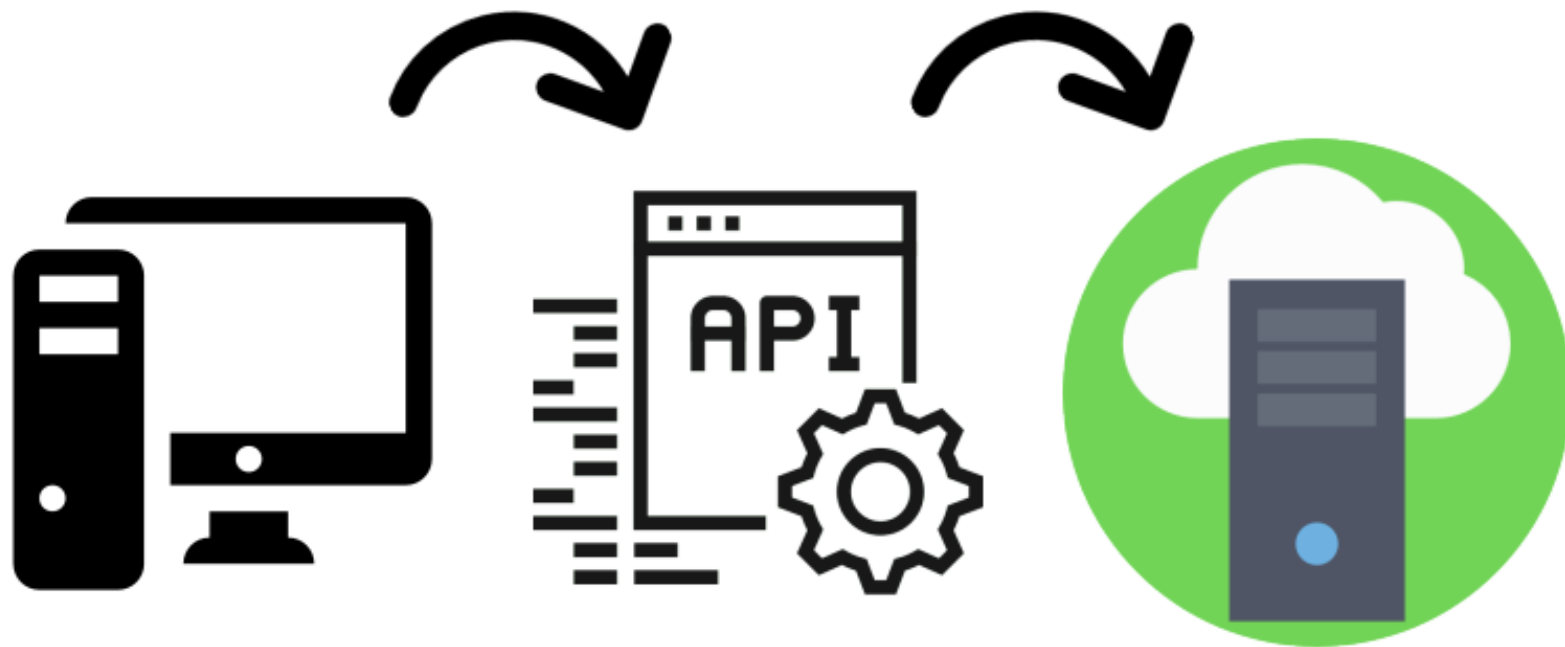
Portal Brasileiro de Dados Abertos: dados.gov.br

Preços de comercialização de combustíveis: anp.gov.br

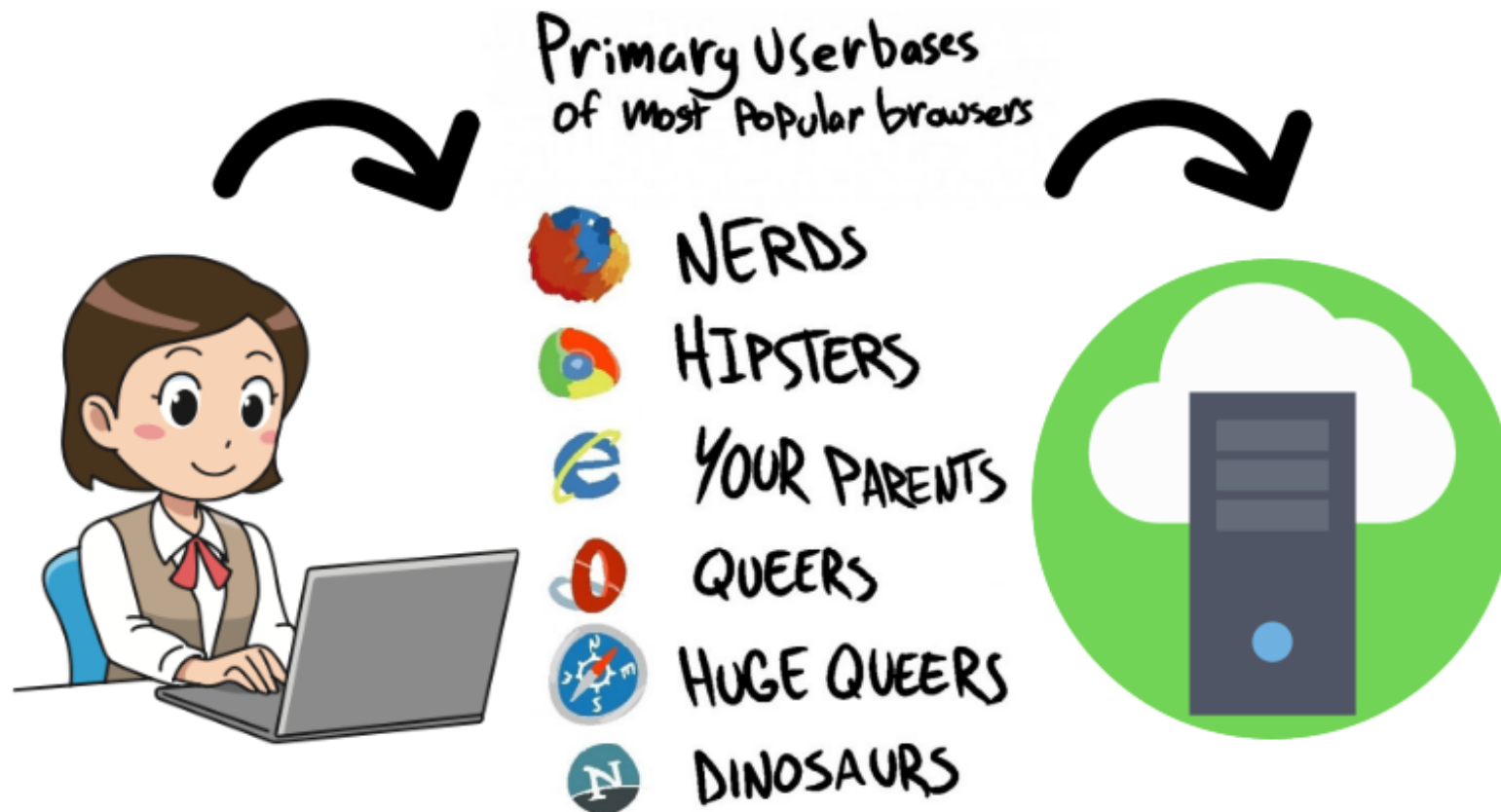
IPCA: ftp.ibge.gov.br

APIs

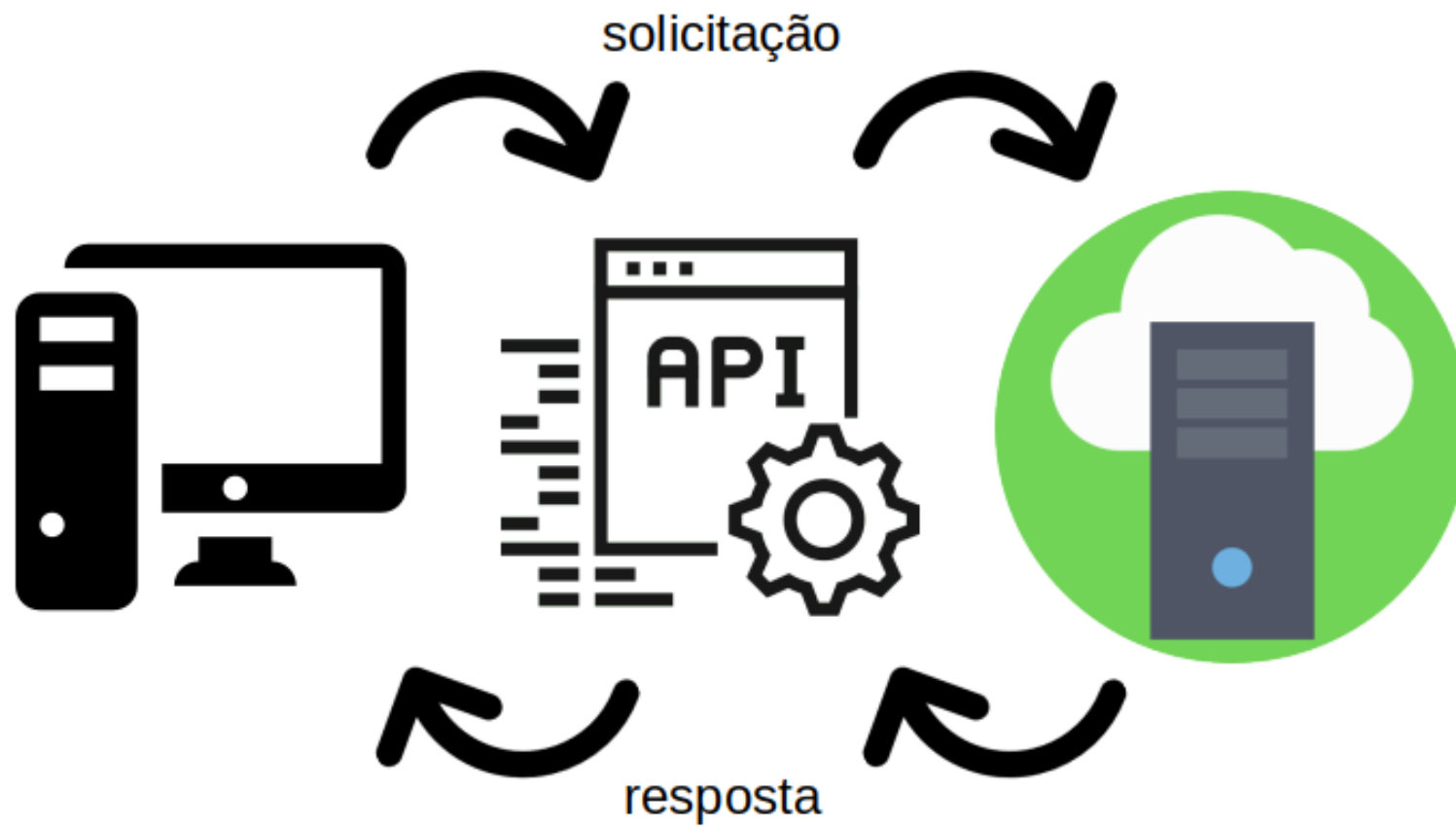
APIs



APIs



APIs



APIs

HTTP requests

- GET
- POST
- DELETE
- HEAD
- [outros](#)

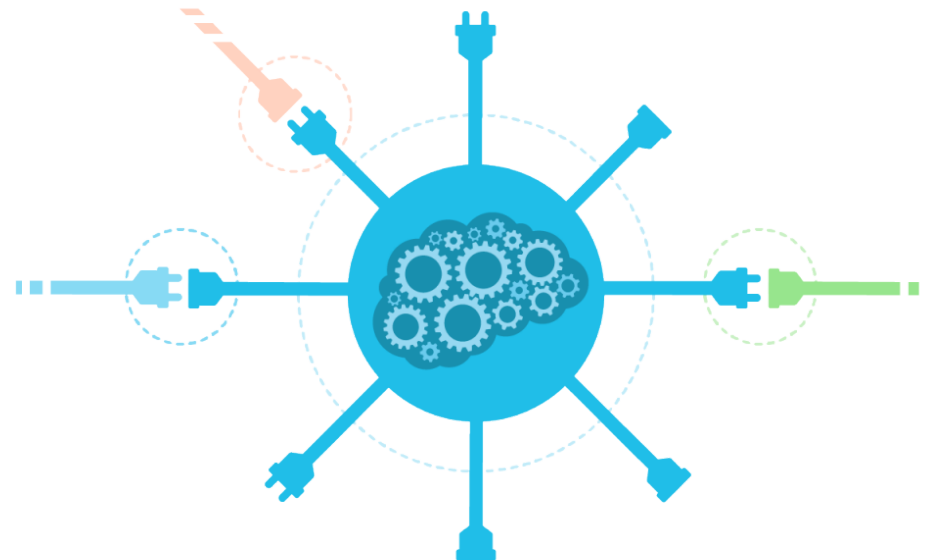
APIs

Respostas

- **200:** sucesso!
- **300:** redirecionamento
- **400:** erro de cliente
- **500:** erro de servidor

[Códigos de resposta HTTP](#)

APIs



The Star Wars API: swapi.co (pacote [rwars](#))

Sistema IBGE de Recuperação Automática: [SIDRA](#)

Banco Central do Brasil: [Portal de Dados Abertos](#)

Web Scraping

web scraping

Estrutura HTML ([tags](#))

- título: `<title> ... </title>`
- parágrafo de texto: `<p> ... </p>`
- blocos: `<div> ... </div>`
- tabela: `<table> ... </table>`
- hiperlink (âncora): `<a> ... `

`IBGE`

web scraping

Extrair com rvest:

- `html_name()`
- `html_attr()`
- `html_text()`
- `html_table()`

```
<a href="http://www.ibge.gov.br/">IBGE</a>
```

web scraping



Amazon: amazon.com.br

web scraping

Selenium

- Selenium 2: [WebDriver](#)
- [Standalone server](#) ([v3.9.1](#))
- Pacote [RSelenium](#)

Executar o servidor

- Usando [Docker](#)
- Usando `RSelenium::rsDriver()`
- Manualmente

web scraping

Funções [RSelenium](#):

- `navigate()`
- `goBack`
- `goForward()`
- `refresh()`
- `findElement()`
- `highlightElement()`
- `clickElement()`
- `mouseMoveToLocation()`
- `click()`
- `sendKeysToElement()`

web scraping



GOL Linhas Aéreas: voegol.com.br

Boas práticas


boas práticas

- Verificar e **respeitar** o `robot.txt`.
- Identificação.
- Usar `http::user_agent()` com e-mail de contato.
- Respeitar o limite de solicitações (rate-limiting).
- Se não houver limite explícito, usar o bom senso.
- Regra geral: intervalo de um segundo entre solicitações, usar `Sys.sleep(1)`.
- Priorizar horários de menor tráfego.

Obrigado!

obrigado

Augusto Fadel
DPE/CEEC/GCAD

 21 2142-0452

 augusto.fadel@ibge.gov.br

 augustofadel