

Web Scraping

HTML e XPath



Fluxo do web scraping

1. Imitar

- Na aba Network seu navegador, investigue as requisições.
- Tente imitá-las no R, copiando os caminhos e parâmetros utilizados.

2. Coletar

- Baixar todas as páginas HTML (ou outro formato).
- Boa prática: salvar arquivos brutos com `httr::write_disk()`.

3. Parsear

- Transformar os dados brutos em uma base de dados passível de análise.
- Utilizar pacotes `{xml2}`, `{jsonlite}`, `{pdftools}`, dependendo do arquivo.

Pacotes

- Utilizar `{httr}` para imitar/coletar.
- Utilizar `{xml2}` para parsear.
- Utilização massiva do `{tidyverse}`.

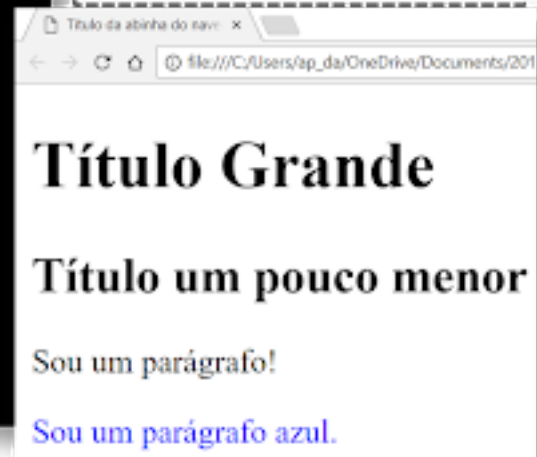
HTML

- HTML (Hypertext Markup Language) é uma linguagem de marcação cujo uso é a criação de páginas web.
- Por trás de todo site há pelo menos um arquivo .html.

Exemplo.html no editor de texto

```
1 <!DOCTYPE html>
2
3 <head>
4   <meta charset = latin1>
5   <title>Título da abinha do navegador</title>
6 </head>
7
8 <body>
9   <h1>Título Grande</h1>
10
11   <h2>Título um pouco menor</h2>
12
13   <p>Sou um parágrafo!</p>
14
15   <p style='color: blue;'>Sou um parágrafo azul.</p>
16
17 </body>
18 </html>
```

Exemplo.html no navegador



HTML

- Todo arquivo HTML pode ser dividido em seções que definirão diferentes aspectos da página.
- `<head>` descreve metadados, enquanto `<body>` é o corpo da página.

Exemplo.html no editor de texto

```
1 <!DOCTYPE html>
2
3 <head>
4   <meta charset = latin1>
5   <title>Título da abinha do navegador</title>
6 </head>
7
8 <body>
9   <h1>Título Grande</h1>
10
11   <h2>Título um pouco menor</h2>
12
13   <p>Sou um parágrafo!</p>
14
15   <p style='color: blue;'>Sou um parágrafo azul.</p>
16
17 </body>
18 </html>
```

Exemplo.html no navegador

file:///C:/Users/ap_da/OneDrive/Documents/201

Título Grande

Título um pouco menor

Sou um parágrafo!

Sou um parágrafo azul.

HTML

- Tags (head, body, h1, p, ...) demarcam as seções e sub-seções
- enquanto atributos (charset, style, ...) mudam como essas seções são renderizadas pelo navegador.

Exemplo.html no editor de texto

```
1 <!DOCTYPE html>
2
3 <head>
4   <meta charset = latin1>
5   <title>Título da abinha do navegador</title>
6 </head>
7
8 <body>
9   <h1>Título Grande</h1>
10
11  <h2>Título um pouco menor</h2>
12
13  <p>Sou um parágrafo!</p>
14
15  <p style='color: blue;'>Sou um parágrafo azul.</p>
16
17 </body>
18 </html>
```

Exemplo.html no navegador

Título da abinha do nave: x

file:///C:/Users/ap_da/OneDrive/Documents/201

Título Grande

Título um pouco menor

Sou um parágrafo!

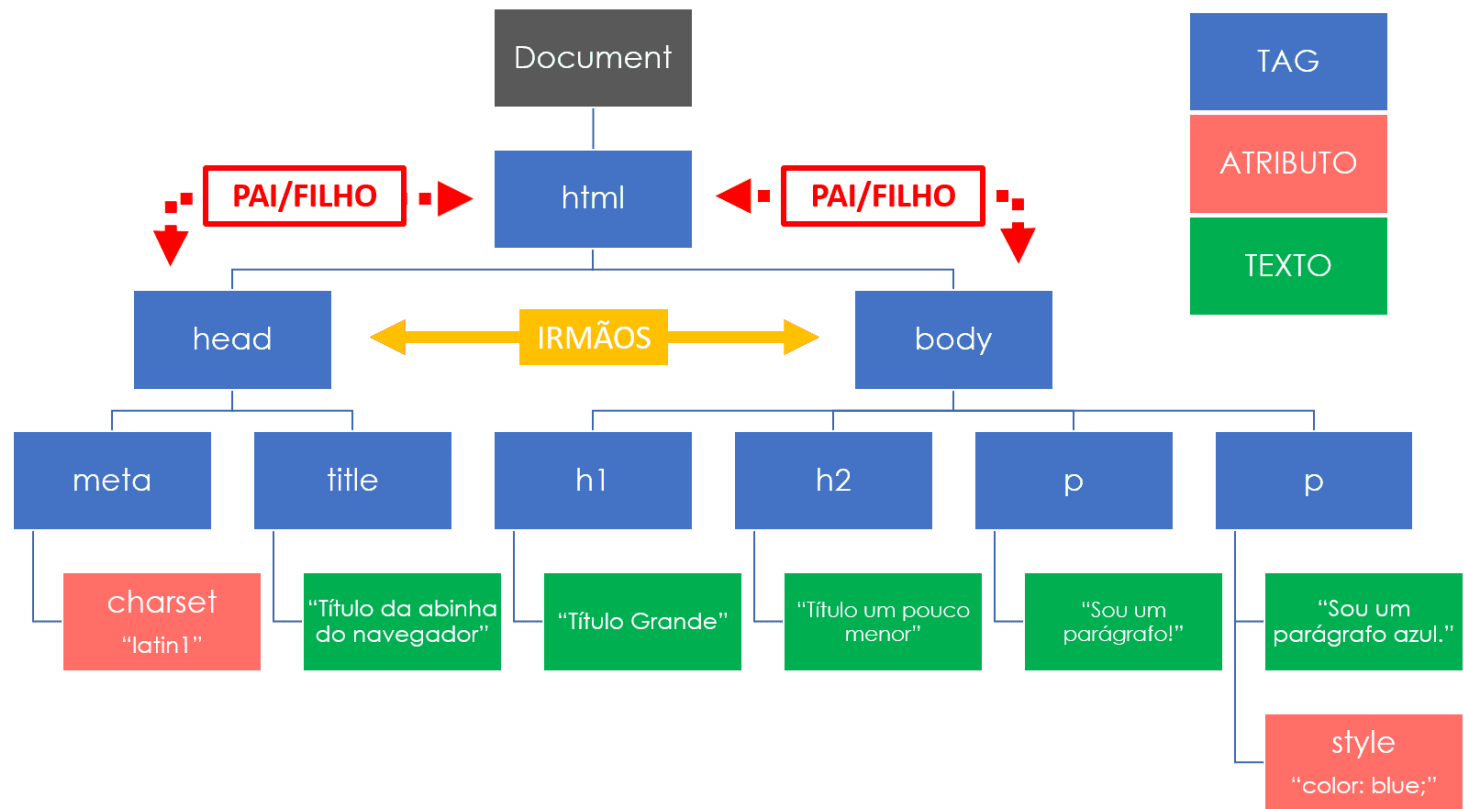
Sou um parágrafo azul.

Teoria

- 1) Todo HTML se transforma em um DOM (document object model) dentro do navegador.
- 2) Um DOM pode ser representado como uma árvore em que cada node é:
 - ou um atributo
 - ou um texto
 - ou uma tag
 - ou um comentário
- 3) Utiliza-se a relação de pai/filho/irmão entre os nós.
- 4) Para descrever a estrutura de um DOM, usamos uma linguagem de markup chamada XML (Extensible Markup Language) que define regras para a codificação de um documento.

HTML

O HTML do exemplo, na verdade, é isso aqui:



Vamos ao R!

