

Web Scraping

com Python



Introdução do HTML

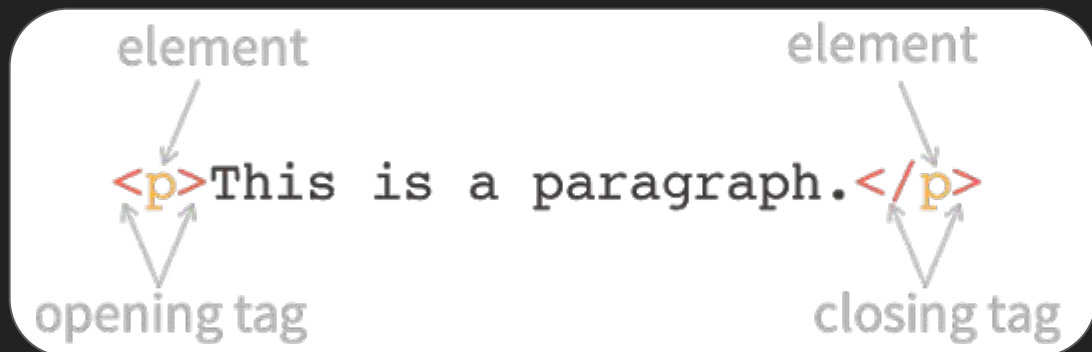
- **HTML:** *HyperText Markup Language*
- Marcação → tags
- Tags HTML:
 - Delimitadas por '<' e '>';
 - Usadas para descrever o elemento que será adicionado;
 - Exemplo de tags HTML:
 - <html>
 - <head>
 - Dentre outras



Introdução do HTML

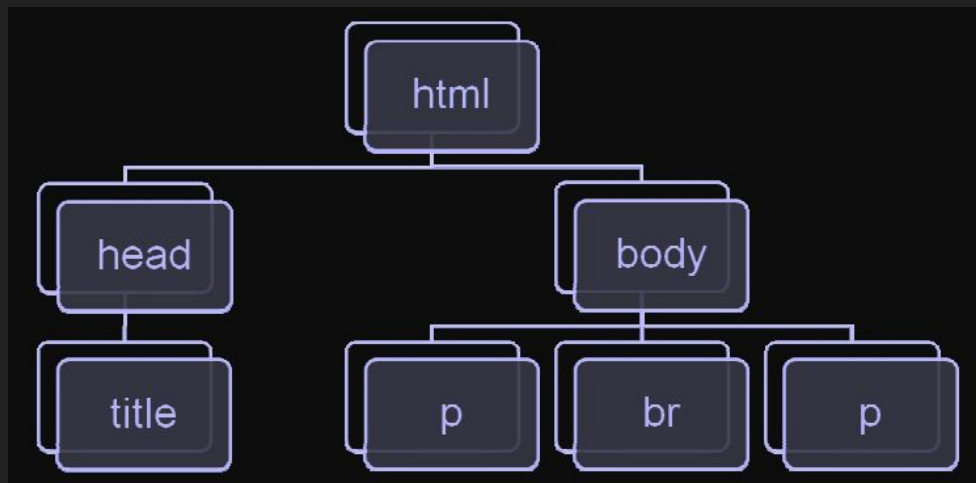
ELEMENTO HTML

- Geralmente contém três componentes:
 - Tag de abertura;
 - Conteúdo;
 - Tag de fechamento.



Introdução do HTML

Estrutura Básica de um Documento HTML



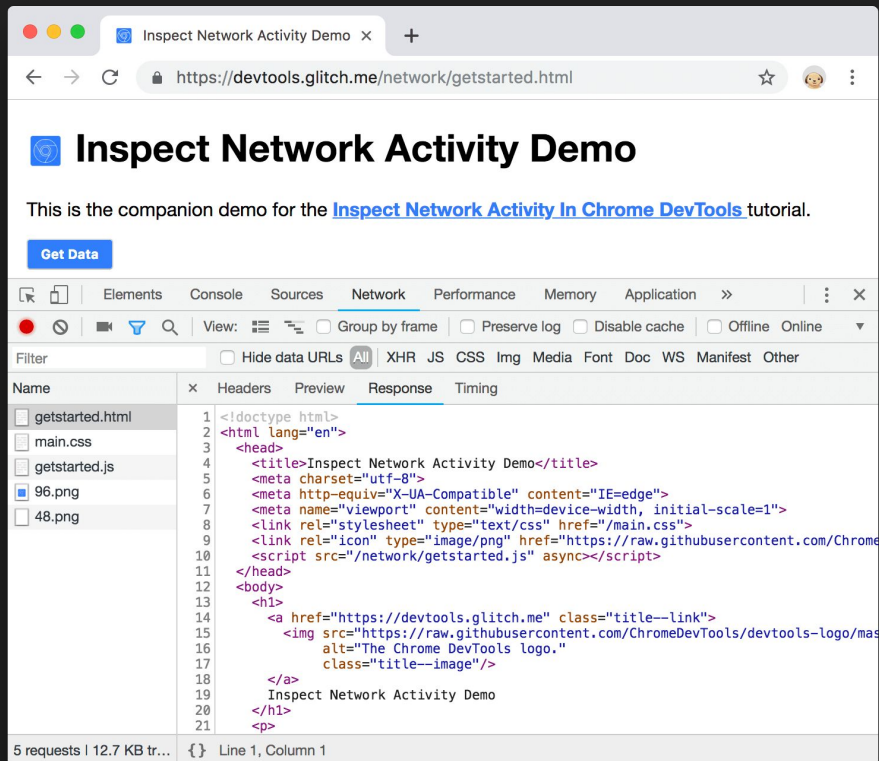
Principais Tags HTML

Tag	Descrição
<code><html> ... </html></code>	Conteúdo HTML
<code><head> ... </head></code>	Cabeçalho do documento
<code><title> ... </title></code>	Título da página HTML
<code><body> ... </body></code>	Corpo do documento (página)
<code><h1> ... </h1></code>	Cabeçalho de nível 1 (pode variar de 1 a 6)
<code><p> ... </p></code>	Parágrafo
<code><div> ... </div></code>	Conteúdo genérico

Principais Tags HTML

Tag	Descrição
<code><a> ... </code>	Link
<code> ... </code>	Conteúdo genérico em linha
<code><table> ... </table></code>	Tabela
<code> ... </code>	Lista não numerada
<code> ... </code>	Lista numerada
<code> ... </code>	Elemento da lista (ou)
<code></code>	Imagem

Inspeccionando o HTML de páginas web



The screenshot shows a web browser window with the address bar displaying `https://devtools.glitch.me/network/getstarted.html`. The page title is "Inspect Network Activity Demo". Below the title, there is a paragraph: "This is the companion demo for the [Inspect Network Activity In Chrome DevTools](#) tutorial." and a blue button labeled "Get Data".

The Chrome DevTools interface is open, with the "Network" tab selected. The left sidebar shows a list of requests: "getstarted.html", "main.css", "getstarted.js", "96.png", and "48.png". The "getstarted.html" request is selected, and the "Response" sub-tab is active. The response content is the HTML of the page, starting with `<!doctype html>` and `<html lang="en">`. The page content includes a title "Inspect Network Activity Demo", a meta charset of "utf-8", a viewport meta tag, and a link to the main CSS file. The body contains a link to the demo page and a Chrome DevTools logo.

```
1 <!doctype html>
2 <html lang="en">
3   <head>
4     <title>Inspect Network Activity Demo</title>
5     <meta charset="utf-8">
6     <meta http-equiv="X-UA-Compatible" content="IE=edge">
7     <meta name="viewport" content="width=device-width, initial-scale=1">
8     <link rel="stylesheet" type="text/css" href="/main.css">
9     <link rel="icon" type="image/png" href="https://raw.githubusercontent.com/ChromeDevTools/devtools-logs/master/logo.png" as="image">
10    <script src="/network/getstarted.js" async></script>
11  </head>
12  <body>
13    <h1>
14      <a href="https://devtools.glitch.me" class="title--link">
15        
16        Inspect Network Activity Demo
17      </a>
18    </h1>
19  </body>
20 </html>
```

At the bottom of the DevTools window, it shows "5 requests | 12.7 KB transferred" and "Line 1, Column 1".