Entendendo o conteúdo web

HTML

HyperText Markup Language (HTML) é uma linguagem com que as páginas web são criadas.

Vamos fazer fazer um rápido tour pelo HTML para que possamos saber o suficiente para fazer o web scraping de maneira eficiente. O HTML é formado por elementos chamados tags. A tag mais básica é <HTML>. Essa tag diz ao navegador que tudo dentro dela é HTML. Podemos criar um documento HTML simples utilizando somente essa tag.

```
1 <html>
2 </html>
```

HTML

Dentro da tag html, nós colocamos outras duas tags, <head> e <body>. O conteúdo principal da página vai dentro da tag <body>. A tag <head> contém dados sobre o título da página e outras informações que normalmente não são úteis ao web scraping.

```
1 <html>
2     <head>
3     </head>
4     <body>
5     </body>
6 </html>
```

HTML

Em html, as tags são aninhadas e podem aparecer dentro de outras tags.

Agora, nós vamos adicionar nosso primeiro conteúdo à página, usando a tag “p”. A tag “p” define um parágrafo, e qualquer texto dentro dela é exibido em um parágrafo separado.

```
1  <html>
2      <head>
3      </head>
4      <body>
5          <p>
6              Here's a paragraph of text!
7          </p>
8          <p>
9              Here's a second paragraph of text!
10         </p>
11     </body>
12 </html>
```

HTML - Tags

As tags têm os nomes normalmente utilizados de acordo com sua posição em relação a outras tags.

- child (filha): Uma tag child é uma tag dentro de outra tag. Então as tags “p” acima são filhas da tag body.
- parent (pai): Uma tag parent é uma tag que tem outras tags dentro. Acima, a tag html é pai de head e body.
- sibling (irmãos): uma tag sibling é aquela que está aninhada dentro do mesmo pai que outra tag. Por exemplo: head e body são irmãs, pois ambas estão dentro da tag html. Ambas as tags “p” são irmãs, pois estão dentro de body.

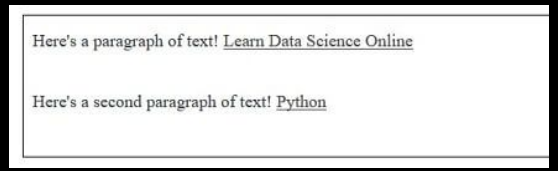
HTML - Propriedades Tags

Nós também podemos adicionar propriedades às tags HTML para mudar seus comportamentos:

```
1  <html>
2      <head>
3      </head>
4      <body>
5          <p>
6              Here's a paragraph of text!
7              <a href="https://www.dataquest.io">Learn Data Scie
8          </p>
9          <p>
10             Here's a second paragraph of text!
11             <a href="https://www.python.org">Python</a>
12         </p>
13     </body>
14 </html>
```

As tags a são links, e dizem ao navegador para renderizar uma outra página. A propriedade href determina para onde o link vai.

a e p são tags extremamente comuns.



HTML - Propriedades Tags

Aqui estão outras:

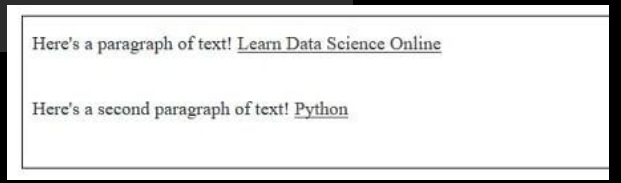
- div – indica uma divisão, uma área na página;
- b – deixa qualquer texto dentro dela em negrito;
- i – deixa em itálico;
- table – cria uma tabela;
- form – cria um formulário;

Class e ID

Cada elemento pode ter várias classes, e uma classe pode ser compartilhada entre elementos. Cada elemento pode ter apenas um id, e um id pode aparecer somente uma vez na página. Classes e ids são opcionais e nem todos os elementos as terão.

Nós podemos adicionar classes ao nosso exemplo:

```
<html>
    <head>
    </head>
    <body>
        <p class="bold-paragraph">
            Here's a paragraph of text!
            <a href="https://www.dataquest.io" id="learn-link"
        </p>
        <p class="bold-paragraph extra-large">
            Here's a second paragraph of text!
            <a href="https://www.python.org" class="extra-larg
        </p>
    </body>
</html>
```



Fonte

https://imasters.com.br/back-end/aprendendo-sobre-web-scraping-em-python-utilizando-beautifulsoup