Biblioteca LXML

Utilizando XPath

Biblioteca LXML

Utilizando XPath

O que é XPath?

XPath (XML Path Language) é uma linguagem para endereçar partes de um documento XML ou HTML.

https://www.w3.org/TR/2017/REC-xpath-31-20170321/

Biblioteca LXML

Utilizando XPath

O LXML possui o método xpath() para que possamos trabalhar com esta linguagem.

Vamos começar nosso exemplo acessando o seguinte endereço:

https://pythonparatodos.com.br/formulario.html

Biblioteca LXML

Utilizando XPath

Veja o HTML do formulário.

```html
<html>
    <head>
        <title>Aula Python Web Scraping</title>
    </head>
    <body>
        <h1>Curso Python Web Scraping</h1>

        <form action="formulario_pythonwebscraping1.php" method="POST">
            Informe os dados abaixo:<br><br>
            <table>
                <tr>
                    <td>Nome:</td>
                    <td><input type="text" name="nome" size="50" maxlength="100"></input></td>
                </tr>
                <tr>
                    <td>E-mail:</td>
                    <td><input type="text" name="email" size="50" maxlength="100"></input></td>
                </tr>
                <tr>
                    <td>Celular:</td>
                    <td><input type="text" name="celular" size="20" maxlength="14"></input></td>
                </tr>
                <tr>
                    <td><input type="submit" name="enviar" value="Enviar dados"></td>
                </tr>
            </table>
        </form>
    </body>
</html>

```

Biblioteca LXML

Utilizando XPath

Agora veja como navegamos na árvore HTML.

```
<html>
    <head>
        <title>Aula Python Web Scraping</title>
    </head>
    <body>
        <h1>Curso Python Web Scraping</h1>

        <form action="formulario_pythonwebscraping1.php" method="POST">
            Informe os dados abaixo:<br><br>
            <table>
                <tr>
                    <td>Nome:</td>
                    <td><input type="text" name="nome" size="50" maxlength="100"></input></td>
                </tr>
                <tr>
                    <td>E-mail:</td>
                    <td><input type="text" name="email" size="50" maxlength="100"></input></td>
                </tr>
                <tr>
                    <td>Celular:</td>
                    <td><input type="text" name="celular" size="20" maxlength="14"></input></td>
                </tr>
                <tr>
                    <td><input type="submit" name="enviar" value="Enviar dados"></td>
                </tr>
            </table>
        </form>
    </body>
</html>
```

Biblioteca LXML

Utilizando XPath

Algumas tags HTML podem possuir atributos, como a tag input que tem o tipo, nome, tamanho e quantidade máxima de caracteres.

Um atributo consiste em:

nome_atributo=valor_atributo

```
<input type="text" name="nome" size="50" maxlength="100">
```

Por exemplo, o valor do atributo type acima é text e do tamanho é 50.

Biblioteca LXML

Utilizando XPath

Veja uma forma simples de utilizar o XPath:

xpath = ‘/html/body’

Usando uma barra simples, avançamos 1 elemento. Após a barra, colocamos o nome da tag que indica a direção para cada elemento.

No exemplo acima, estamos navegando até a tag body, que fica dentro da tag html.

Biblioteca LXML

Utilizando XPath

Veja o primeiro exemplo.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
body = tree.xpath('/html/body')
print(body)
for elemento in body:
    print(etree.tostring(elemento))
```

Biblioteca LXML

Utilizando XPath

O html importado de lxml é para converter o HTML à partir de uma string.

Estamos usando o método xpath passando '/html/body'.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
body = tree.xpath('/html/body')
print(body)
for elemento in body:
    print(etree.tostring(elemento))
```

Biblioteca LXML

Utilizando XPath

A variável body receberá uma lista de objetos Element de LXML. Ao percorrer a variável body, estamos transformando esse elemento em string para realizar a impressão.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
body = tree.xpath('/html/body')
print(body)
for elemento in body:
    print(etree.tostring(elemento))
```

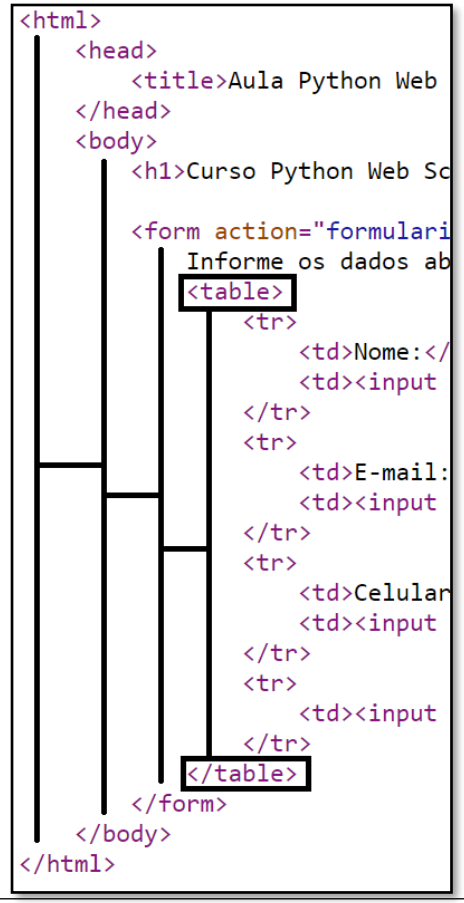
Biblioteca LXML

Utilizando XPath

Vamos agora para o segundo exemplo. Nesse exemplo, vamos acessar a tabela html percorrendo todo caminho desde a tag html.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
table = tree.xpath('/html/body/form/table')
print(table)
for elemento in table:
    print(etree.tostring(elemento))
```



Biblioteca LXML

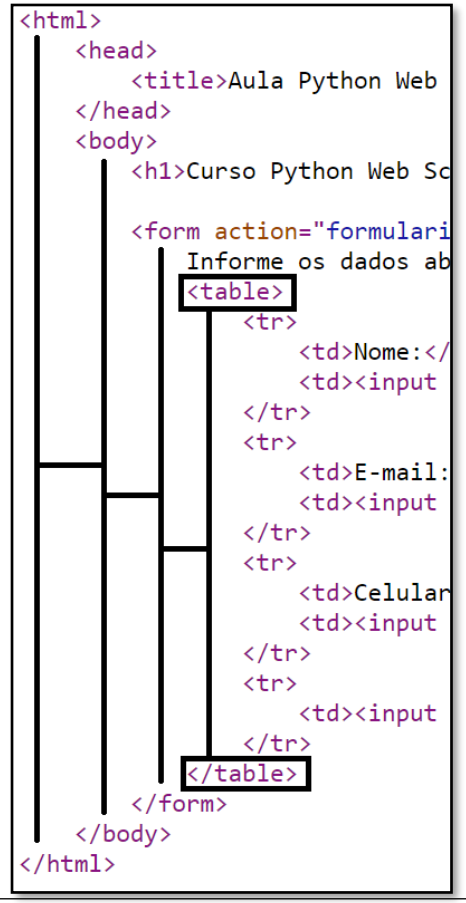
Utilizando XPath

No terceiro exemplo vamos utilizar duas barras “//”.

O “//table” vai direto para todos elementos table do documento (na nossa página temos apenas 1 tag table).

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
table = tree.xpath('//table')
print(table)
for elemento in table:
    print(etree.tostring(elemento))
```



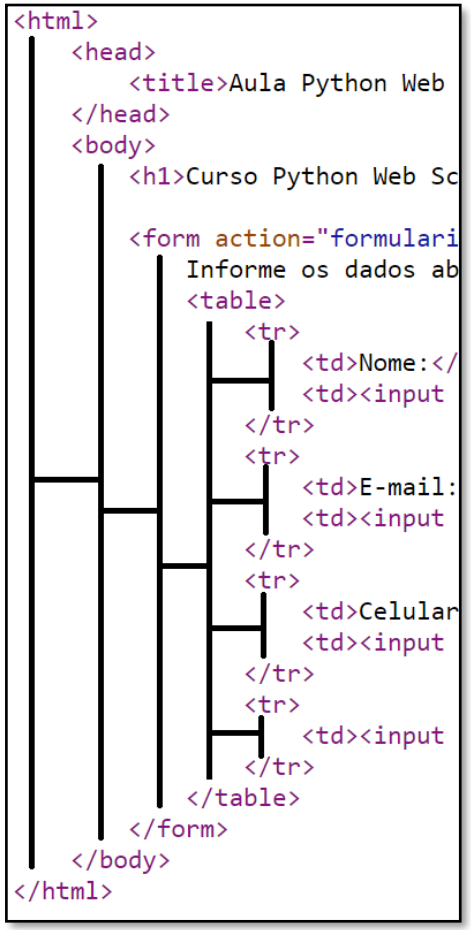
Biblioteca LXML

Utilizando XPath

No quarto exemplo, vamos acessar todos elementos tr do documento.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
tr = tree.xpath('//tr')
print(tr)
for elemento in tr:
    print(etree.tostring(elemento))
```



Biblioteca LXML

Utilizando XPath

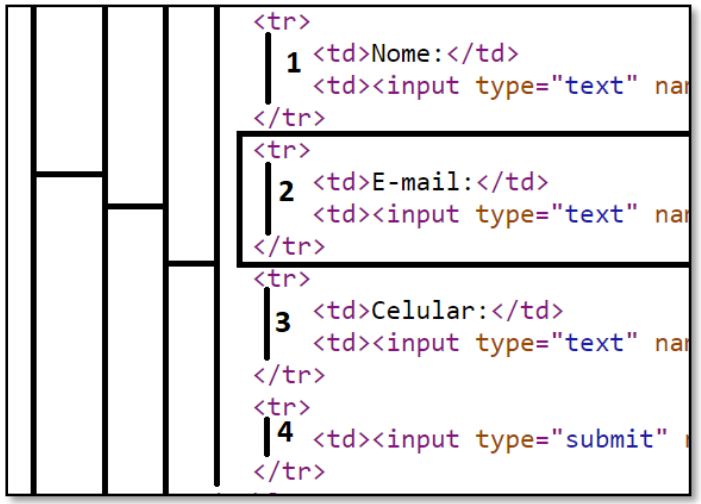
No quinto exemplo, vamos acessar o segundo elemento tr (E-mail).

Colchetes em uma tag informa qual dos elementos vamos selecionar.

Temos que passar o índice do elemento, que começa em 1.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
tr = tree.xpath('//tr[2]')
print(tr)
for elemento in tr:
    print(etree.tostring(elemento))
```



Biblioteca LXML

Utilizando XPath

No sexto exemplo, vamos acessar o segundo elemento td do segundo elemento tr.

```
<tr>
  <td>E-mail:</td>
  <td><input type="text" name="email" size="50" maxlength="100"></input></td>
</tr>
```

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
td = tree.xpath('//tr[2]/td[2]')
print(td)
for elemento in td:
    print(etree.tostring(elemento))
```

Biblioteca LXML

Utilizando XPath

Podemos utilizar também o asterisco como caracter coringa.

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
table = tree.xpath('//table/*')
print(table)
for elemento in table :
    print(etree.tostring(elemento))
```

Biblioteca LXML

Utilizando XPath

Podemos também utilizar atributos das tags HTML com XPath.

Para isso utilizamos o caracter “@”.

Por exemplo, na página utilizada nesta aula temos várias tags input. Como faríamos para retornar apenas a tag cujo atributo name seja celular?

Poderíamos utilizar o seguinte XPath:

```
input = tree.xpath('//input[@name="celular"]')
```

Biblioteca LXML

Utilizando XPath

Assim será o oitavo exemplo:

```
from lxml import html, etree
from urllib.request import urlopen

pagina = urlopen("https://www.pythonparatodos.com.br/formulario.html")
tree = html.fromstring(pagina.read())
input = tree.xpath('//input[@name="celular"]')
print(input)
for elemento in input :
    print(etree.tostring(elemento))
```

Biblioteca LXML

Utilizando XPath

Você verá mais sobre XPath nas aulas sobre o framework Scrapy.

FIM