

INTRODUÇÃO AO XPATH:

NAVEGANDO (E EXTRAINDO DADOS) DE DOCUMENTOS HTML

RENNE ROCHA

- Python Developer na **Scrapinghub**
- Laboratório Hacker de Campinas (<https://lhc.net.br>)
- Grupy-Campinas | Grupy-Jundiaí
- renne@rennerochoa.com
- [@rennerochoa](#) (Github, Twitter, Telegram, etc)

XPath (XML Path Language) é uma linguagem de consulta para selecionar e navegar por nós de um documento XML.

XPath (XML Path Language) é uma **linguagem de consulta** para selecionar e navegar por nós de um documento XML.

XPath (XML Path Language) é uma linguagem de consulta para selecionar e navegar por nós de um documento XML.

XPath (XML Path Language) é uma linguagem de consulta para selecionar e navegar por nós de um documento XML.

XPath (XML Path Language) é uma linguagem de consulta para selecionar e navegar por nós de um documento XML.

XML (Extensible Markup Language) é um formato baseado em texto para representação de informações estruturadas, como por exemplo documentos, configurações, livros, transações, etc.

XML (Extensible Markup Language) é um formato baseado em texto para **representação de informações estruturadas**, como por exemplo documentos, configurações, livros, transações, etc.



```
<?xml version="1.0" encoding="ISO-8859-1"?>
<receita titulo="brigadeiro">
  <ingredientes>
    <item qtd="1" unit="colher">Manteiga</item>
    <item qtd="1" unit="lata">Leite Condensado</item>
    <item qtd="4" unit="colher">Chocolate em Pó</item>
    <item qtd="1" unit="pacote">Chocolate Granulado</item>
  </ingredientes>
  <instrucoes>
    <passo>Aqueça a panela em fogo médio.</passo>
    <passo>Acrescente 1 colher de sopa de manteiga.</passo>
    <passo>Acrescente o Leite Condensado</passo>
    <passo>Acrescente o Chocolate em Pó</passo>
    <passo>Mexa sem parar até desgrudar da panela.</passo>
  </instrucoes>
</receita>
```

CASOS DE USO

- Arquivos de Configuração
- Troca de mensagens entre aplicações
- Padrão para documentos (Open Document Format)

HTML

HTML

HYPertext MARKUP LANGUAGE

```
<html>
  <title>Receita de Brigadeiro</title>
  <body data-content="receita">
    <h1>Ingredientes</h1>
    <ul>
      <li data-qtd="1" data-unit="colher">Manteiga</li>
      <li data-qtd="1" data-unit="lata">Leite Condensado</li>
      <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
      <li data-qtd="1" data-unit="pacote">Chocolate Granulado</li>
    </ul>
    <h1>Instruções de Preparo</h1>
    <ol>
      <li>Aqueça a panela em fogo médio.</li>
      <li>Acrescente 1 colher de sopa de manteiga.</li>
      <li>Acrescente o Leite Condensado</li>
      <li>Acrescente o Chocolate em Pó</li>
      <li>Mexa sem parar até desgrudar da panela.</li>
    </ol>
  </body>
</html>
```

<https://docs.python.org/3/library/xml.etree.elementtree.html/>

The screenshot shows a web browser window displaying the Python documentation for the `xml.etree.ElementTree` module. The browser's address bar shows the URL `https://docs.python.org/3/library/xml.etree.elementtree.html/`. The page title is `xml.etree.ElementTree — The ElementTree XML API — Python 3.7.3 documentation`. The page layout includes a sidebar with a "Table of Contents" and a main content area. The sidebar lists the following items: `xml.etree.ElementTree`, Tutorial, XML tree and elements, Parsing XML, Pull API for non-blocking parsing, Finding interesting elements, Modifying an XML File, Building XML documents, Parsing XML with Namespaces, Additional resources, XPath support, Example, Supported XPath syntax, Reference, and Functions. The main content area features the title `xml.etree.ElementTree — The ElementTree XML API`, a source code link (`Lib/xml/etree/ElementTree.py`), a description of the module, a version change note for 3.3, a warning box, and a "Tutorial" section.

Table of Contents

- `xml.etree.ElementTree`
 - The ElementTree XML API
 - Tutorial
 - XML tree and elements
 - Parsing XML
 - Pull API for non-blocking parsing
 - Finding interesting elements
 - Modifying an XML File
 - Building XML documents
 - Parsing XML with Namespaces
 - Additional resources
 - XPath support
 - Example
 - Supported XPath syntax
 - Reference
 - Functions

`xml.etree.ElementTree` — The ElementTree XML API

Source code: [Lib/xml/etree/ElementTree.py](#)

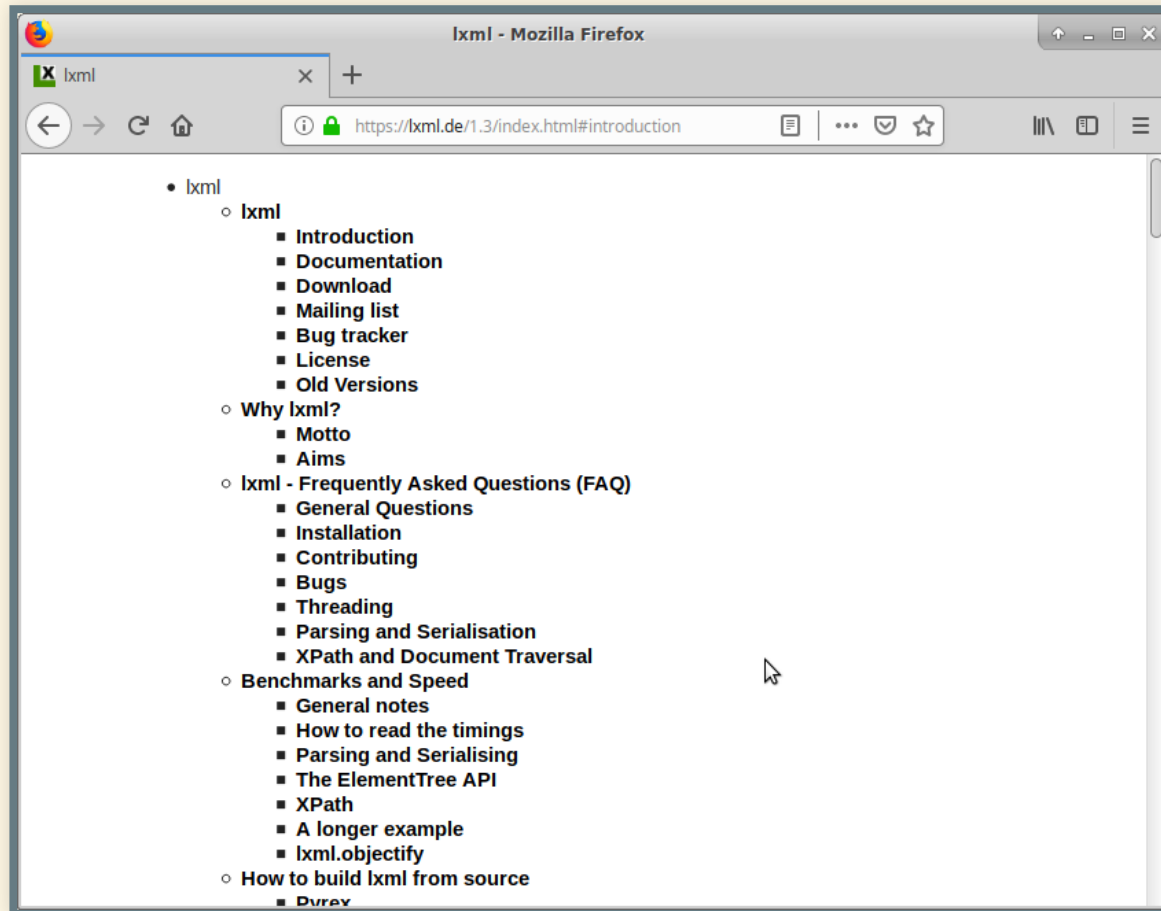
The `xml.etree.ElementTree` module implements a simple and efficient API for parsing and creating XML data.

Changed in version 3.3: This module will use a fast implementation whenever available. The `xml.etree.cElementTree` module is deprecated.

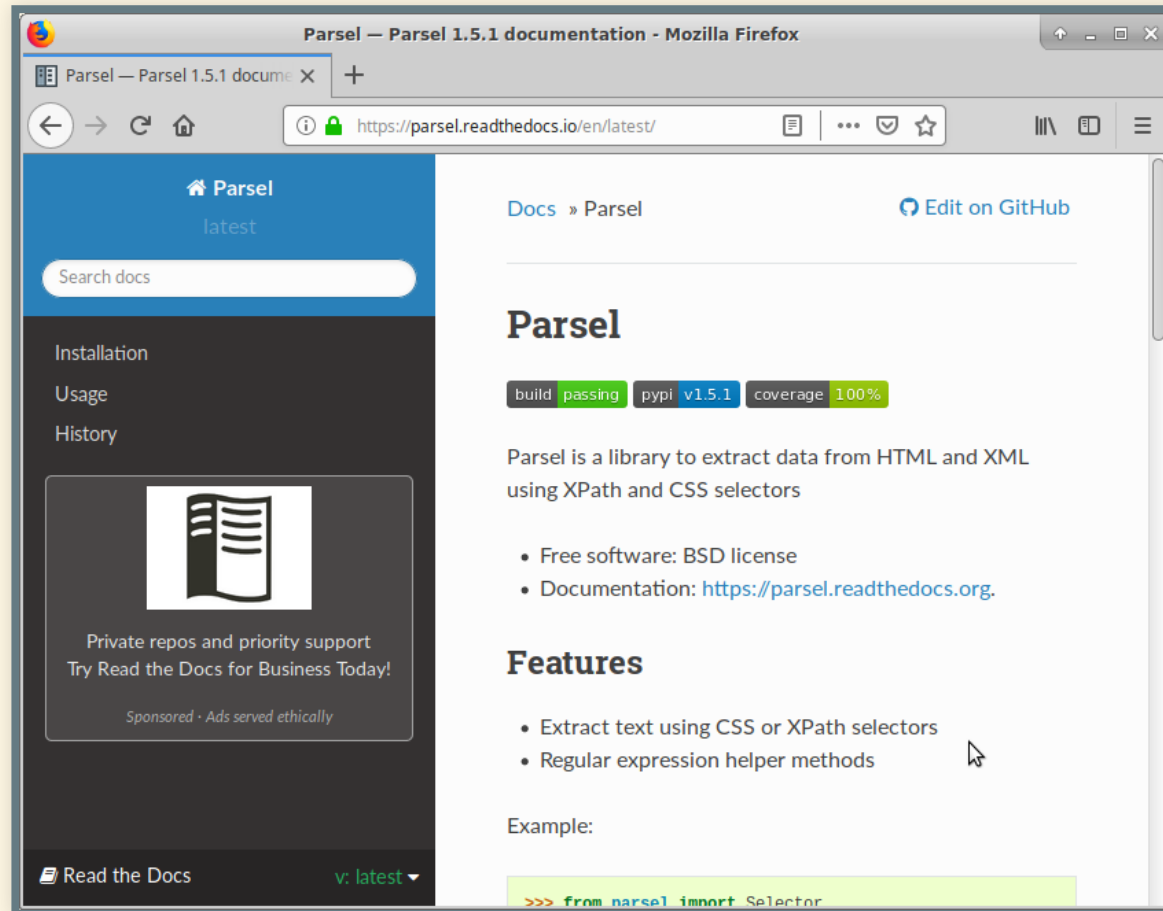
Warning: The `xml.etree.ElementTree` module is not secure against maliciously constructed data. If you need to parse untrusted or unauthenticated data see [XML vulnerabilities](#).

Tutorial

<https://lxml.de/>



<https://parsel.readthedocs.io/>



```
1 from parsel import Selector
2
3 with open("brigueiro.html", "r") as xml_file:
4     content = xml_file.read()
5     selector = Selector(text=content)
```

```
1 selector.xpath("/title")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6   <h1>Ingredientes</h1>
7   <ul>
8     <li data-qtd="1" data-unit="colher">Manteiga</li>
9     <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10    <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11    <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12  </ul>
13  <h1>Instruções de Preparo</h1>
14  <ol>
15    <li>Aqueça a panela em fogo médio.</li>
16    <li>Acrescente 1 colher de sopa de manteiga.</li>
17    <li>Acrescente o Leite Condensado</li>
18    <li>Acrescente o Chocolate em Pó</li>
19    <li>Mexa sem parar até desgrudar da panela.</li>
20  </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("/html/head/title")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6   <h1>Ingredientes</h1>
7   <ul>
8     <li data-qtd="1" data-unit="colher">Manteiga</li>
9     <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10    <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11    <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12  </ul>
13  <h1>Instruções de Preparo</h1>
14  <ol>
15    <li>Aqueça a panela em fogo médio.</li>
16    <li>Acrescente 1 colher de sopa de manteiga.</li>
17    <li>Acrescente o Leite Condensado</li>
18    <li>Acrescente o Chocolate em Pó</li>
19    <li>Mexa sem parar até desgrudar da panela.</li>
20  </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//title")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6   <h1>Ingredientes</h1>
7   <ul>
8     <li data-qtd="1" data-unit="colher">Manteiga</li>
9     <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10    <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11    <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12  </ul>
13  <h1>Instruções de Preparo</h1>
14  <ol>
15    <li>Aqueça a panela em fogo médio.</li>
16    <li>Acrescente 1 colher de sopa de manteiga.</li>
17    <li>Acrescente o Leite Condensado</li>
18    <li>Acrescente o Chocolate em Pó</li>
19    <li>Mexa sem parar até desgrudar da panela.</li>
20  </ol>
21 </body>
22 </html>
```



```
1 selector.xpath("//h1")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//ul/*")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//ul/li[2]")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 In [1]: selector.xpath("//ul/li/text()").getall()
2 Out[1]: [
3     'Manteiga',
4     'Leite Condensado',
5     'Chocolate em Pó',
6     'Chocolate Granulado'
7 ]
```

```
<html>
<head>
  <title>Receita de Brigadeiro</title>
</head>
<body data-content="receita">
<h1>Ingredientes</h1>
<ul>
  <li data-qtd="1" data-unit="colher">Manteiga</li>
  <li data-qtd="1" data-unit="lata">Leite Condensado</li>
  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
</ul>
<h1>Instruções de Preparo</h1>
<ol>
  <li>Aqueça a panela em fogo médio.</li>
  <li>Acrescente 1 colher de sopa de manteiga.</li>
  <li>Acrescente o Leite Condensado</li>
  <li>Acrescente o Chocolate em Pó</li>
  <li>Mexa sem parar até desgrudar da panela.</li>
</ol>
</body>
</html>
```



```
1 In [1]: selector.xpath("//ul/li/@data-qtd").getall()
2 Out[1]: ['1', '1', '4', '1']
```

```
1 selector.xpath("//li[contains(., 'Chocolate')]")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//ul/li[not(contains(., 'Chocolate'))])")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//li[@data-unit='colher']")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

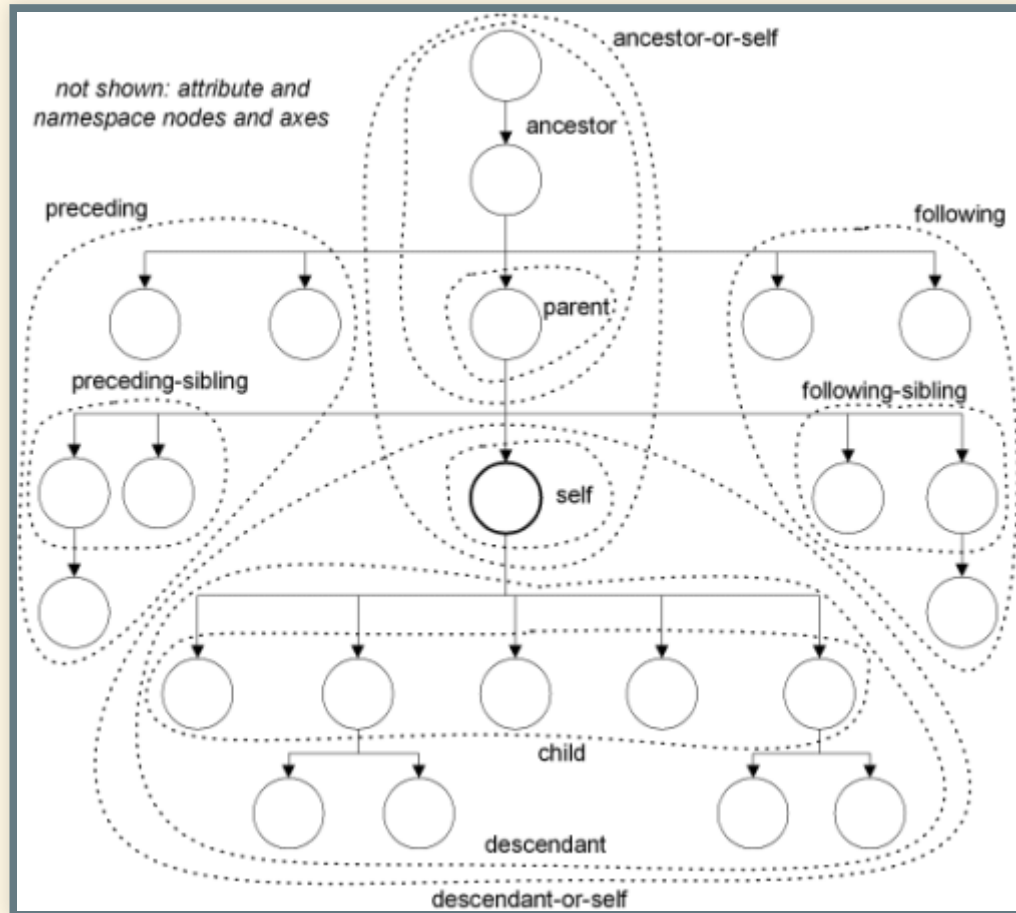
```
1 selector.xpath("//li[@data-qtd=1 and @data-unit='colher']")
```



```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//li[@data-qtd>2]")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```



```
1 selector.xpath(  
2     "//h1[contains(., 'Instruções')]/preceding-sibling::ul"  
3 )
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtyd="1" data-unit="colher">Manteiga</li>
9   <li data-qtyd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtyd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtyd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```

```
1 selector.xpath("//title/ancestor::html//h1")
```

```
1 <html>
2 <head>
3   <title>Receita de Brigadeiro</title>
4 </head>
5 <body data-content="receita">
6 <h1>Ingredientes</h1>
7 <ul>
8   <li data-qtd="1" data-unit="colher">Manteiga</li>
9   <li data-qtd="1" data-unit="lata">Leite Condensado</li>
10  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
11  <li data-qtd="1" data-unit="pacote">Choc. Granulado</li>
12 </ul>
13 <h1>Instruções de Preparo</h1>
14 <ol>
15   <li>Aqueça a panela em fogo médio.</li>
16   <li>Acrescente 1 colher de sopa de manteiga.</li>
17   <li>Acrescente o Leite Condensado</li>
18   <li>Acrescente o Chocolate em Pó</li>
19   <li>Mexa sem parar até desgrudar da panela.</li>
20 </ol>
21 </body>
22 </html>
```


- <https://developer.mozilla.org/pt-PT/docs/Web/XPath>
- http://www.dicas-l.com.br/arquivo/tutorial_xpath.php

OBRIGADO