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@rennerocha.com
- @rennerocha (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
<?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)

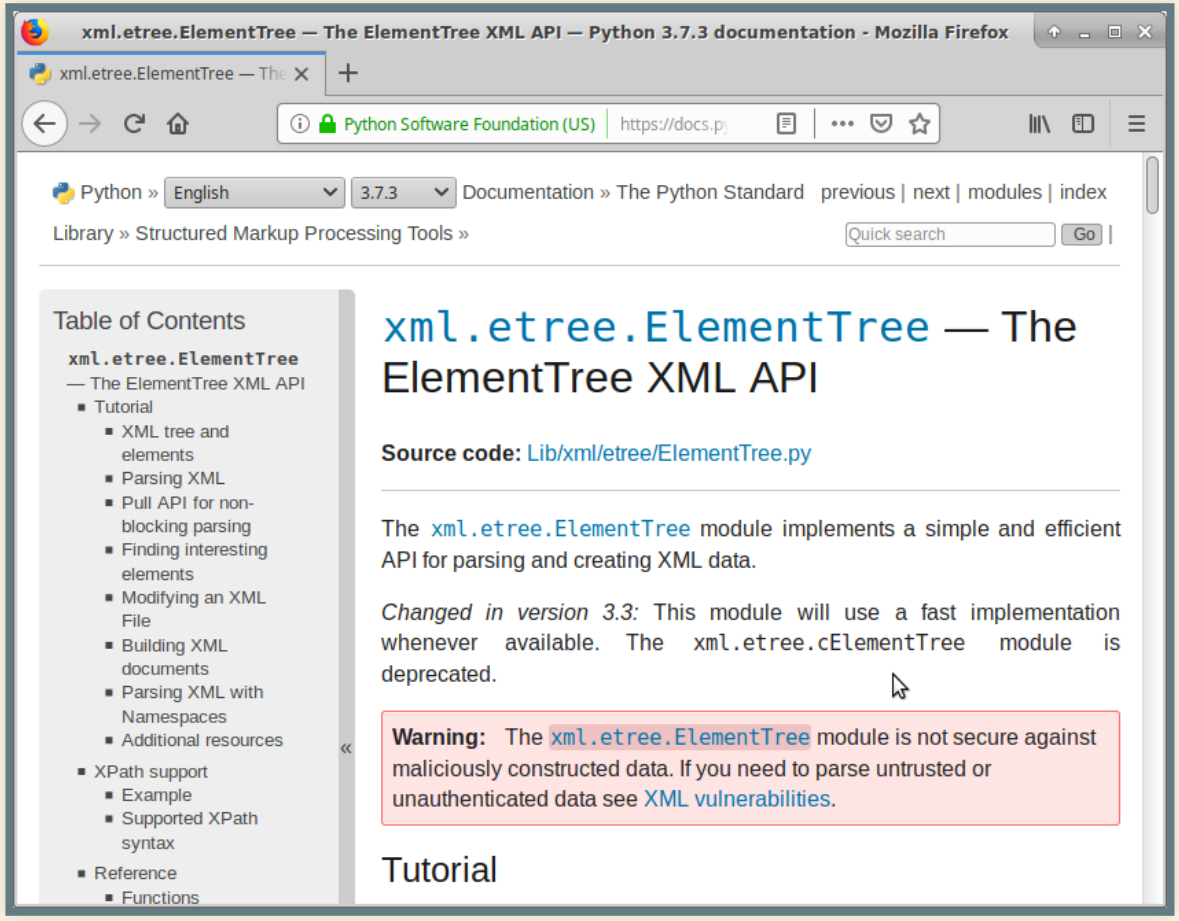
HTML

HTML

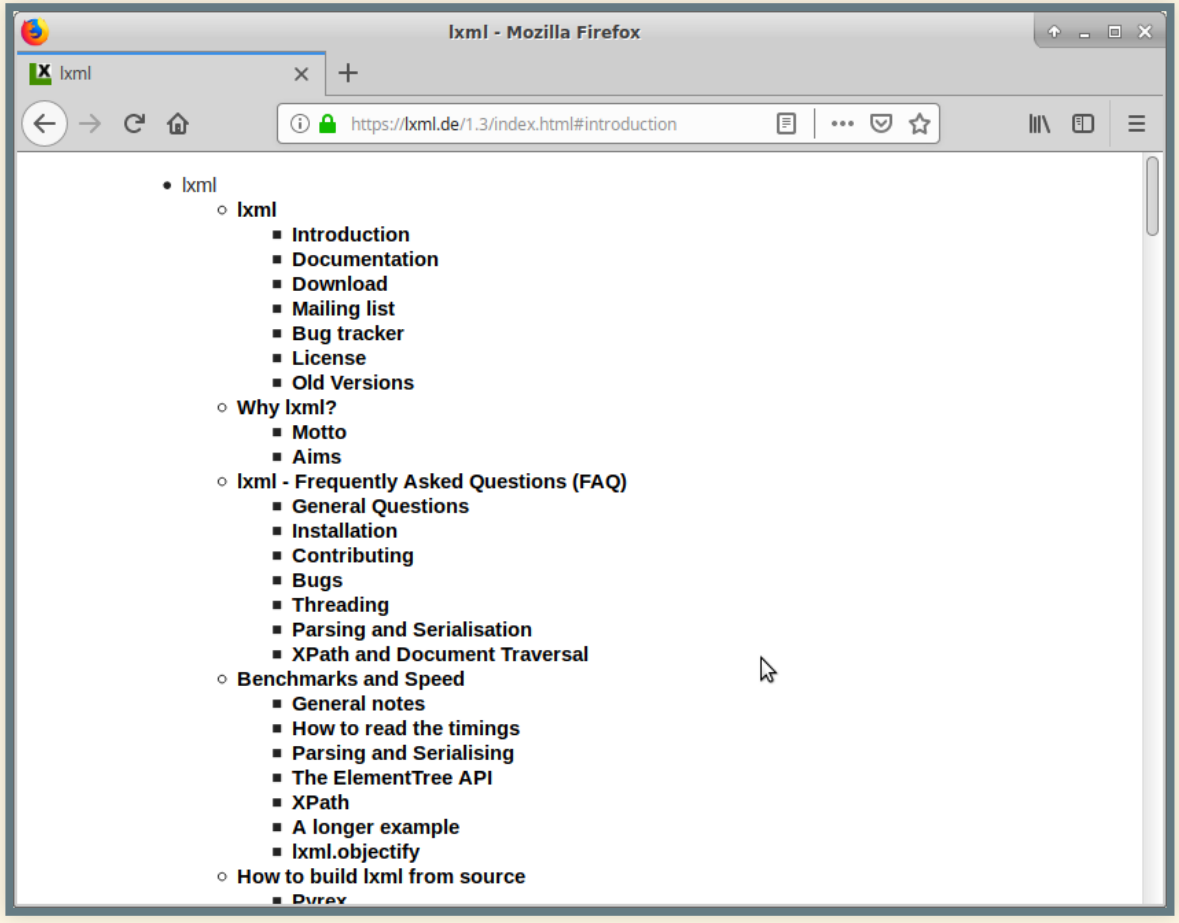
HYPERTEXT MARKUP LANGUAGE

```html
<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>
```

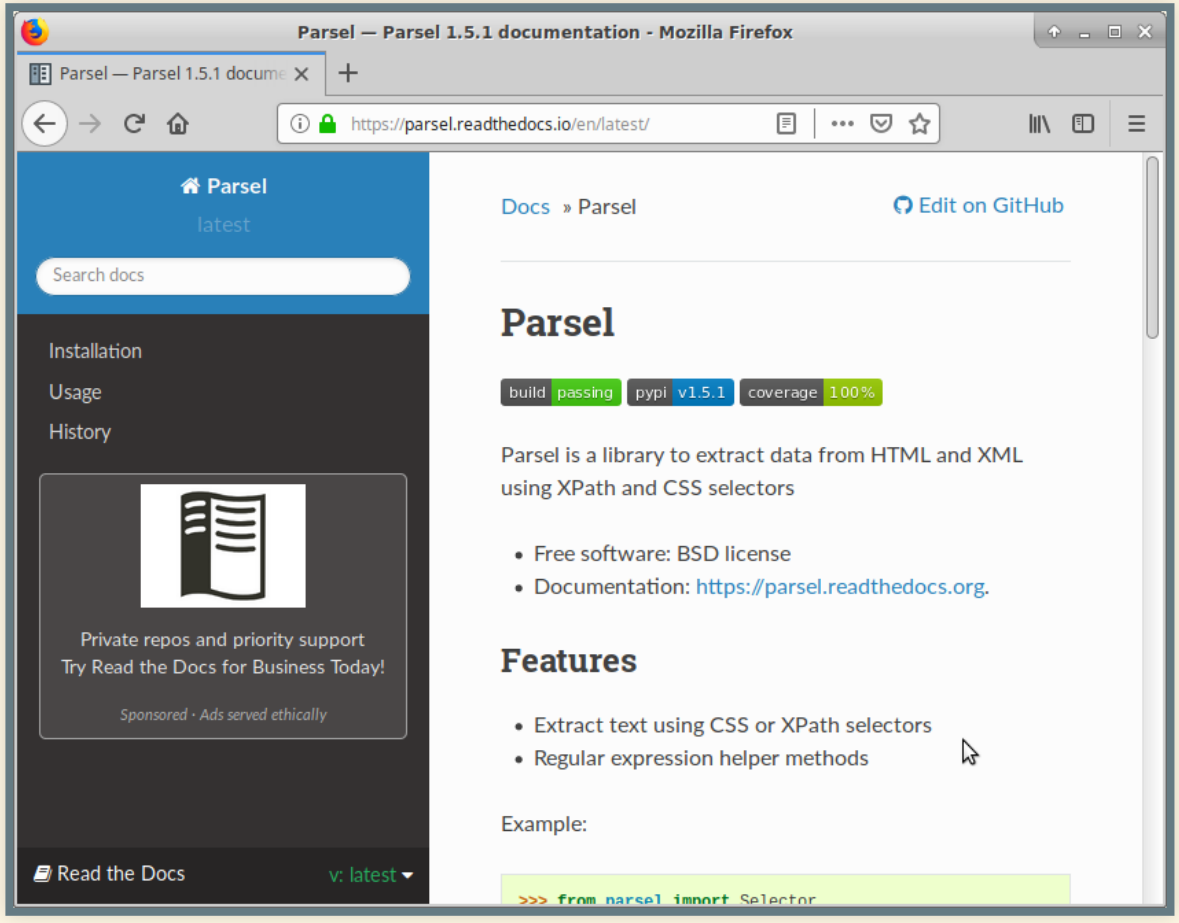
https://docs.python.org/3/library/xml.etree.elementtree.html/



https://lxml.de/



https://parsel.readthedocs.io/



```
from parsel import Selector

with open("brigadeiro.html", "r") as xml_file:
    content = xml_file.read()
    selector = Selector(text=content)
```

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

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

```html
<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 selector.xpath("//title")
```

```html
<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 selector.xpath("//h1")
```

```html
<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 selector.xpath("//ul/*")
```

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

```
<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>
```

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

```html
<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>
```

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

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

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

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

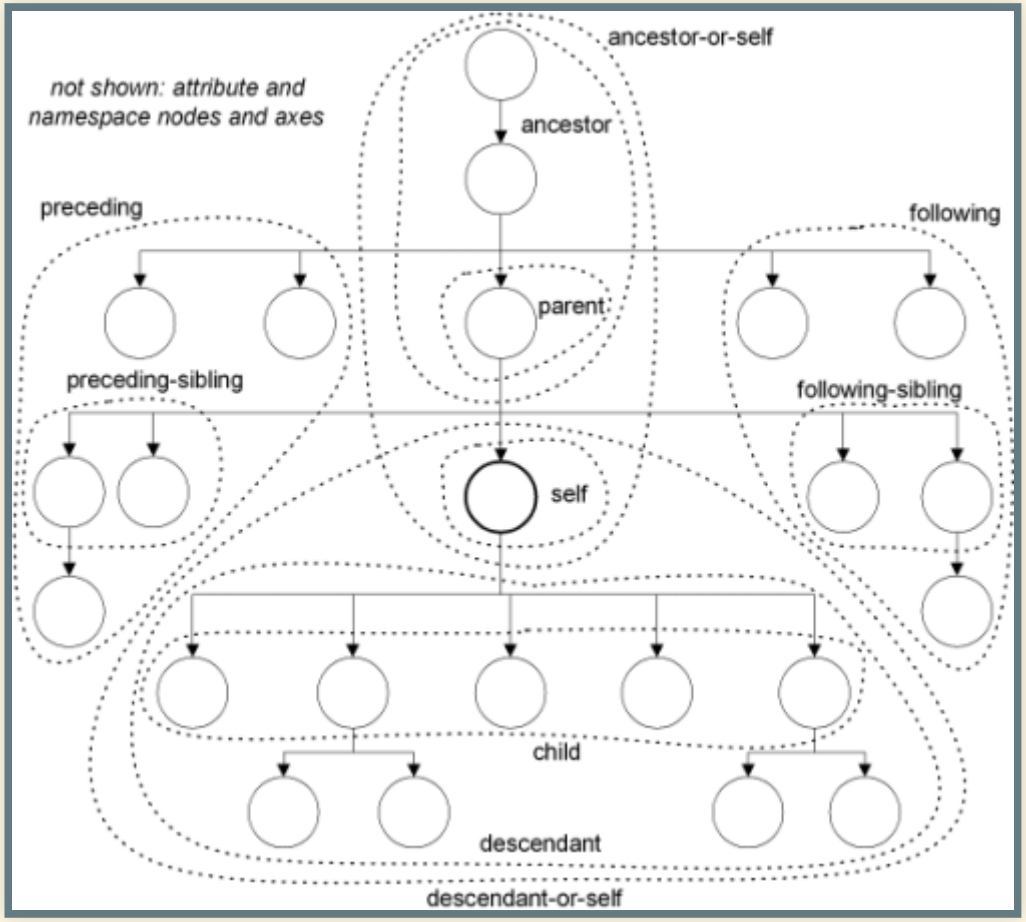
```html
<html>
<head>
  <title>Receita de Brigadeiro</title>
</head>
<body data-content="receita">
<h1>Ingredientes</h1>
<ul>
  <li data-qtd="1" data-unit="colher">Manteiga</li>
  <li data-qtd="1" data-unit="lata">Leite Condensado</li>
  <li data-qtd="4" data-unit="colher">Chocolate em Pó</li>
  <li 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 selector.xpath("//li[@data-qtd=1 and @data-unit='colher']")
```

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

```
<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>
```


ancestor-or-self
not shown: attribute and namespace nodes and axes
ancestor
preceding
following
parent
preceding-sibling
following-sibling
self
child
descendant
descendant-or-self

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

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

```html
<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>
```

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

OBRIGADO