Pràctica 8.2: Web Scraping (XPath)

Lliuraments

Els resultats d'aquesta part de la pràctica s'hauran d'entregar en format PDF i l'entrega pot ser a través de GIT* o el moodle.

*** S'ha d'entregar l'enllaç del GIT al moodle.**

Guió

Amb l'ajuda de l'inspector d'elements del navegador, investiga com està formatada la pàgina https://scrapepark.org/ . Aquesta pàgina està preparada per fer *web scraping,* de manera que les rutes per arribar als diferents elements no són trivials.

Exercici 1

Per començar, clona el repositori de GIT que es troba en aquesta ubicació i executa el codi Python per veure quin resultat dona.

https://github.com/pauitic/practica8_2

Exercici 2

a. Executa les següents rutes XPath i observa el resultat que dona cada una. A continuació, explica les diferències que hi ha entre cada resultat i raona per què produeixen resultats diferents.

i. node() vs text()

Ruta 1: `//div[@class='attribution']/p/node()`
Ruta 2: `//div[@class='attribution']/p/text()`
La primera ruta selecciona tots els nodes fills del paràgraf, incloent text, etiquetes i altres nodes. La segona ruta selecciona només el contingut de text dins del paràgraf, excluint altres nodes.

ii. Barra simple vs barra doble

Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

La primera ruta selecciona els elements de llista directament dins de l'element amb la classe 'navbar-nav', mentre que la segona ruta selecciona qualsevol element de llista a qualsevol profunditat dins de l'element 'navbar-nav'.

b. Representa, en forma d'arbre l'estructura XML que resulta d'avaluar la següent ruta XPath (pots ignorar els salts de línia i espais).

i. `(//div/h5)[6]`

ii. `//div[@class='carousel-item'][1]//h1`

Exercici 3

Descobreix la ruta XPath per arribar a cada un dels elements que es demana tenint en compte només la informació que es proporciona a l'enunciat.

c. Troba la ruta que arriba al **correu** de contacte que es troba al **<footer>** de la pàgina. **Comença la ruta a l'etiqueta <html>**

```
/html//footer//div[@class='information-f']/p[3]/span/node()
```

sales@mail.com

d. Troba la ruta que arriba a l'**atribut *src*** de la següent imatge (n'hi ha una al <*footer*>, i una al <*header*>, pots escollir):

images/logo.svg

```
/html//footer//div[@class='logo-footer']/a/img/@src
```

e. Troba la ruta fins a l'**atribut *src*** de les imatges amb ***alt="Client"***.

images/client-one.png
images/client-two.png
images/client-three.png

```
/html//div[@class='img-box-inner']/img/@src
```

f. Troba la ruta fins a l'**adreça** de la pàgina web ***"Fake Street 123".*** Fes que l'adreça XPath parteixi la següent ubicació:

```
//div[@class='information-f']/p[1]/strong/text()
```

```
Fake Street 123
```

```
/html//div[@class='information-f']/p[1]/strong/../span/text()
```

g. Troba la ruta que arriba fins al **<h5>** del **"New Scateboard 12".** [**Pista**: busca la utilitat de la funció *normalize-space()*].

```
<h5>                    <span>New Skateboard</span> 12
</h5>
```

```
/html//h5[normalize-space()='New Skateboard 12']
```

h. Partint de la ruta de l'apartat anterior, Troba la ruta que arriba fins al **preu** (text) del **"New Scateboard 12".**

```
12
```

```
/html//h5[normalize-space()='New Skateboard 12']/text()
```

Exercici 4

Canvia la ruta a https://scrapepark.org/table.html . Amb l'ajuda del navegador, comprova què hi ha dins d'aquesta pàgina i troba la ruta XPath dels següents elements.

i. Troba la ruta XPath a tots els **preus** dels **elements de color 'Blue'**. El resultat ha de ser el següent:

```
Blue
$64
$70
$80
$85
```

```
/html//tr[1]//td/text()
```

j. Troba la ruta que imprimeix **els preus del *longboard*** que es troben a la 4a columna de la taula **pintats en vermell**.

```
Longboard
$80
$85
$90
$62
$150
```

```
/html//tr//th[4]/text() | //tr//td[4]/text()
```

k. **Indica el nom i color** de l'article que **val $110**. Comença l'expressió de la següent manera: [**pista**: hauràs de fer servir l'operador "|"]

```
//td[text()='$110']
```

```
Skate
Special
```

```
/html//td[text()='$110']/../td[1]/text() | //tr/th[2]/text()
```

l. Troba la ruta a **tots els preus** dels objectes "Purple" **excepte el preu** que està pintat en vermell.

```
<td>Purple</td>
<td class="text-center">$55</td>
<td class="text-center">$60</td>
<td class="text-center">$72</td>
```

```
//tr[td='Purple']/td[@class='text-center' and not(@style='color: red;')]
```

Heu proporcionat un codi de Python que fa servir la biblioteca `lxml` per fer web scraping amb XPath a la pàgina https://scrapepark.org/. A continuació, proporcionaré els resultats esperats per als exercicis que heu plantejat:

Exercici 1:

Per a l'exercici 1, heu de clonar el repositori i executar el codi Python proporcionat per veure quin resultat dona. Aquest pas és més pràctic fer-lo localment.

Exercici 2:

- Ruta 1: `//div[@class='attribution']/p/node()`
- Ruta 2: `//div[@class='attribution']/p/text()`
  Les diferències radican en el fet que `node()` selecciona tots els nodes fills, incloent elements, text, comentaris, etc., mentre que `text()` selecciona només el contingut de text d'un node.
- Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`
- Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`
  La diferència entre les dues rutes radica en l'ús de la barra simple i la barra doble. La barra simple selecciona només els fills directes, mentre que la barra doble selecciona qualsevol nivell de descendència.

Exercici 3:

- Correu de contacte: `/html/body/footer/a/text()` o `/html/body/footer/a/@href` (si es vol l'enllaç complet).
- Atribut `src` de la imatge al `<footer>`: `/html/body/footer//img/@src`
- Atribut `src` de les imatges amb `alt="Client"`: `/html/body//img[@alt='Client']/@src`
- Adreça de la pàgina "Fake Street 123": `/html/body/div[@class='information-f']/p[1]/strong/text()`
- `<h5>` del "New Skateboard 12": `//span[text()='New Skateboard 12']/ancestor::h5`

Exercici 4:

- Ruta a tots els preus dels elements de color 'Blue': `//td[text()='Blue']/following-sibling::td[@class='text-center']/text()`
- Ruta als preus del longboard a la 4a columna pintats en vermell: `//td[text()='Longboard']/following-sibling::td[@class='text-center' and @style='color:red']/text()`
- Nom i color de l'article que val $110: `//td[text()='$110']/preceding-sibling::td[@class='text-center'][1]/text() | //td[text()='$110']/preceding-sibling::td[@class='text-center'][2]/text()`

Podeu utilitzar aquestes rutes per modificar el vostre codi i extreure la informació requerida.

Recordeu que el web scraping hauria de realitzar-se de manera ètica i en conformitat amb les polítiques del lloc web.