



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 triviales.

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()`

El primero lo que hace es que busca los div de que tengan la classe atributo entonces lo que te devuelve son los hijos de p en este caso nos devuelve a "a"

Ruta 2: `//div[@class='attribution']/p/text()`

La ruta numero 2 nos dará todos los nodos de texto directamente que esten dentro del párrafo. Devolverá directamente el contenido del elemento P y este no incluirá otro nodo

ii. Barra simple vs barra doble

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

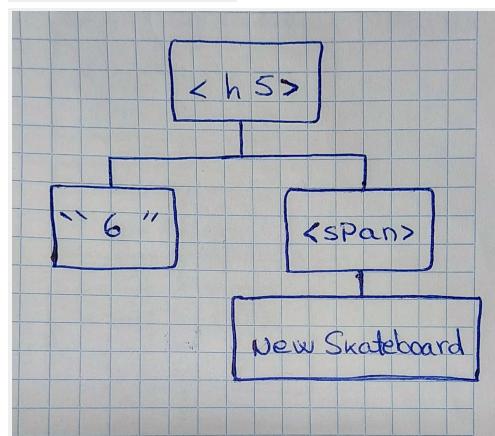
El primero lo es hace toda la ruta dandote los resultados que la cumplan

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

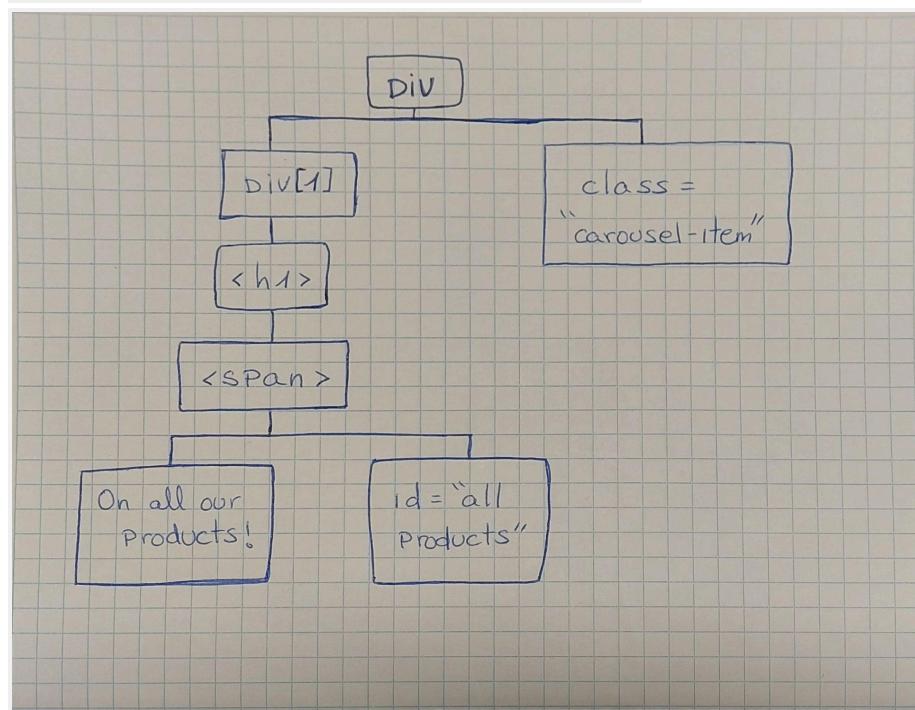
El segundo coge la ruta y llega hasta class=navbar-nav hace un salto hasta llegar a li(que es todo lo que cuelga de él) y te muestra su contenido.

- b. Representa, en forma d'arbre l'estruatura HTML 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

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

altra manera

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

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):



```
//div[@class='container']/nav/a/img/@src
```

images/logo.svg

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

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

images/client-one.png

images/client-two.png

images/client-three.png

- 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()
```

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

Fake Street 123

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

```
//h5[span ='New skateboard']/text()[normalize-space()='12']/..”
```

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

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

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

```
//table[@class='table           table-bordered           table-striped  
table-hover']/tbody/tr/td[text()='Blue']/..
```

- 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

```
/*[@style='color: red;']/node()
```

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

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

queda por acabar

Special

```
//td[text()='$110']/.../.../tr/td[text()='Special']/node()
```

primera linea

I. Troba la ruta a tots els preus dels objectes “Purple” excepte el preu que està pintat en vermell.

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

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