Melissa Ait
Marcos López Martínez
ASIX-B

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. 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.

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

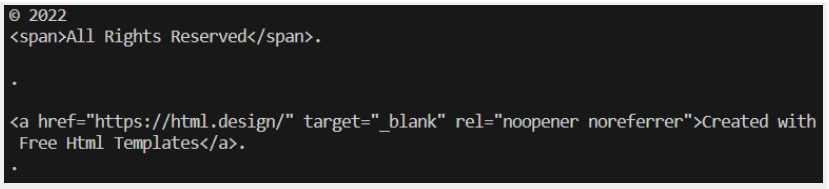
a. Executa les següents rutes XPath i observa el resultat que dona cada una. A continuació, explica les diferències que hi ha entre cada resultat i raona per què produeixen resultats diferents.

i. node() vs text()

Ruta 1: `//div[@class='attribution']/p/node()`

`selecciona todos los nodos hijos del elemento <p> dentro del <div>.`

Esto incluirá no solo los nodos de texto (como lo haría text()), sino también cualquier otro tipo de nodo, como nodos de elemento, comentarios, etc.



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

selecciona solo los nodos de texto directamente descendientes del elemento <p> dentro del .Es decir que solo devolverá el contenido de texto directamente dentro del elemento <p>, excluyendo cualquier otro tipo de nodo.

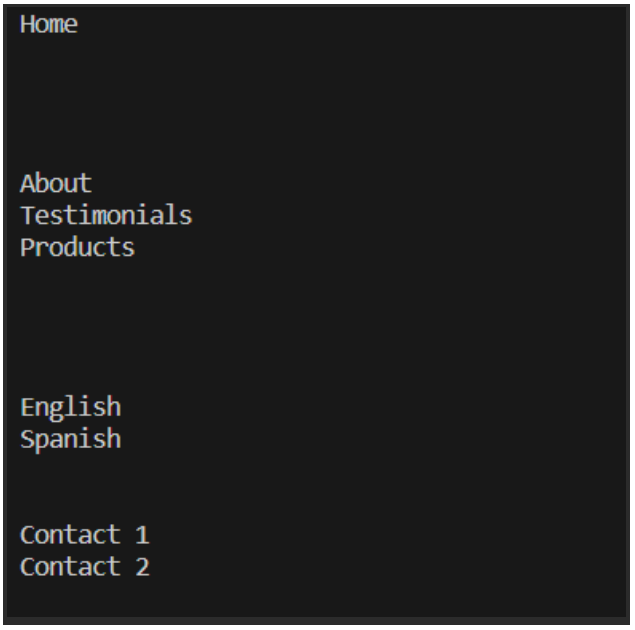
i. Barra simple vs barra doble

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

Selecciona todos los elementos <a> dentro de los elementos <li> que están directamente dentro de un elemento <ul> con la clase 'navbar-nav'.

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

`Selecciona todos los nodos de texto dentro de los elementos <a> que están dentro de cualquier elemento <li>, sin importar su nivel de anidamiento, dentro de un elemento <ul> con la clase '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]`

`html/body/section/div/div/div/div/div/h5/span`

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

`html/body/div/section/div/div/div/div/div/div/div/h1`

Exercici 3

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

`/html`

```
sales@mail.com
```
```
/html/body/footer//div[@class='information-f']/p[contains(text(),'EMAIL')]/span
```

c. 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/body/footer//div[contains(@class,'logo-footer')]/a/img/@src
```

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

```
//img[@alt='Client']/@src
```

```
images/client-one.png
//img[@alt='Client' and @src='images/client-one.png']

images/client-two.png
//img[@alt='Client' and @src='images/client-two.png']

images/client-three.png
//img[@alt='Client' and @src='images/client-three.png']
```

e. 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/body/footer/div[@class='container']/div[@class='row']/div[@class='col-md-4']/div[@class='information-f']/p[1]/span
```

f. 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>
```

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

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

```
110
```

```
//h5[normalize-space()='New Skateboard 12']/../h6/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.

h. 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
```

```
//tr[td='Blue']/td/text()
```

i. 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
```

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

j. **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
```

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

k. 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[position()!=4]/text()
```