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()`

Ruta 1 selecciona tots els nodes fills, incloent text, elements, atributs, etc. dels elements p. En canvi, Ruta 2, només selecciona els nodes de text.

ii. Barra simple vs barra doble

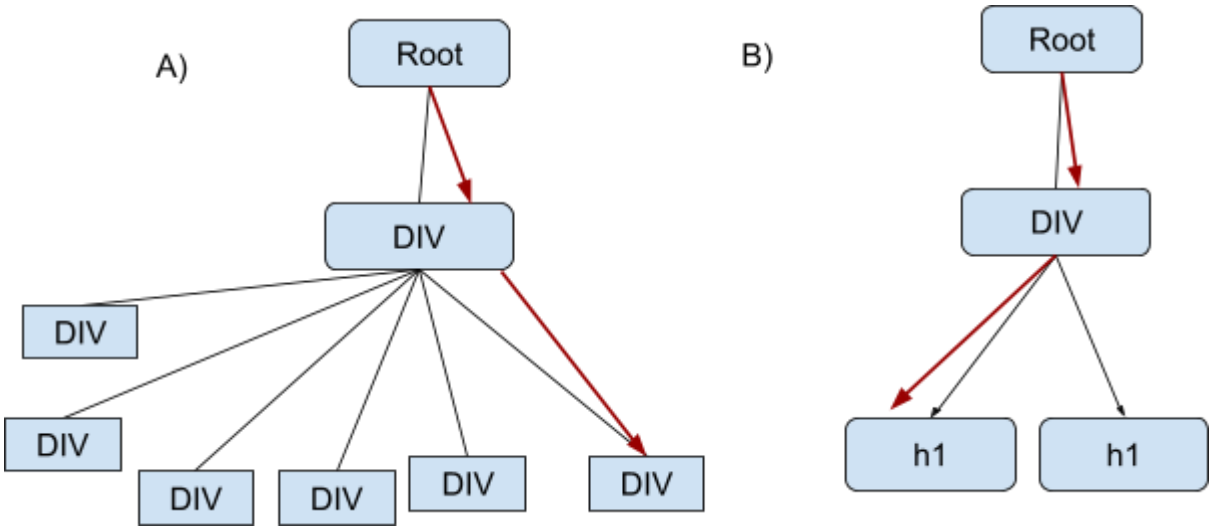
Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

Ruta 1 només selecciona els li que són fills directes del ul mentre que Ruta 2 Selecciona tots els descendents de l'element ul.

b. Representa, en forma d'arbre l'estructura 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/body/footer/div/div/div[1]/div/div[2]/p[3]/span
`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):

/html/body/footer/div/div/div[1]/div/div[1]/a/img
images/logo.svg

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

images/client-one.png
/html/body/section[5]/div/div[2]/div[1]/div[1]/div/div[1]/div/div/img
images/client-two.png
/html/body/section[5]/div/div[2]/div[1]/div[2]/div/div[1]/div/div/img
images/client-three.png
/html/body/section[5]/div/div[2]/div[1]/div[3]/div/div[1]/div/div/img

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/body/footer/div/div/div[1]/div/div[2]/p[1]/span

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

<h5>
New Skateboard 12
</h5>
/html/body/section[3]/div/div[2]/div[12]/div/div[3]/h5/span

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

$110
/html/body/section[3]/div/div[2]/div[12]/div/div[3]/h6

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: **//tbody/tr[1]/td[1]**
$64: **//tbody/tr[1]/td[2]**
$70: **//tbody/tr[1]/td[3]**
$80: **//tbody/tr[1]/td[4]**
$85: **//tbody/tr[1]/td[5]**

or do the following to get all the value in the row at once:
//tbody/tr[1]/td

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

Longboard: **//thead/tr/th[4]**
$80: **//thead/tbody[1]/th[4]**
$85: **//thead/tbody[2]/th[4]**
$90: **//thead/tbody[3]/th[4]**
$62: **//thead/tbody[4]/th[4]**
$150: **//thead/tbody[5]/th[4]**

or you can do the following to get all the price value at once.
//tbody/tr/td[4]

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

//thead/tr/th[2] | //tbody/tr[5]/td[1]

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>
```

//table/tbody/tr[4]/td[not(@style='color: red;')]