Institut TIC de Barcelona
Consorci d'Educació de Barcelona
Generalitat de Catalunya
Ajuntament de Barcelona

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()`
`Se muestra la etiqueta <span> y  enlace que está pegada en footer.`
Ruta 2: `//div[@class='attribution']/p/text()`
`Se muestra la información de copyright.`

ii. Barra simple vs barra doble

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

`Se muestra la "home", "product" a dónde "home" y "product" no tiene información dentro de ellos.(no tiene hijos)`

Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`
`Devuelve la etiqueta home, product y la información dentro de otras etiquetas (content,language,contact) se muestra información de hijos de estas tres content,language,contact etiquetas.`

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]`
`Se devuelve la información de h5 y también muestra que esta dentro de h5.`

ii. `//div[@class='carousel-item'][1]//h1`
`se muestra información de "h1" y la etiqueta "span" que está dentro de h1.`

Exercici 3

c. 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/body/footer/div/div/div[1]/div/div[2]/p[3]/span/span/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):

`images/logo.svg`
`html/body/header/div/a/img`

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`
`//body/section[5]//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:

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

```
Fake Street 123
```

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/body/section[3]/div/div[2]/div[12]/div/div[3]/h5
```

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

```
12
/html/body/section[3]/div/div[2]/div[12]/div/div[3]/h5/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
//tbody/tr/td[text()='Blue']/../node()/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
```

```
//thead/tr/th[@style]/text() | //tbody//tr/td[@style]/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 "|"]

```
Skate
Special
//td[text()='$110']/../../../thead/tr/th[text()='Skate']/text(
) | //td[text()='$110']/../td[text()='Special']/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>
```

```
//tbody/tr/td[text()='Purple']/../node()[normalize-space()=not
(@style)]
```