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.

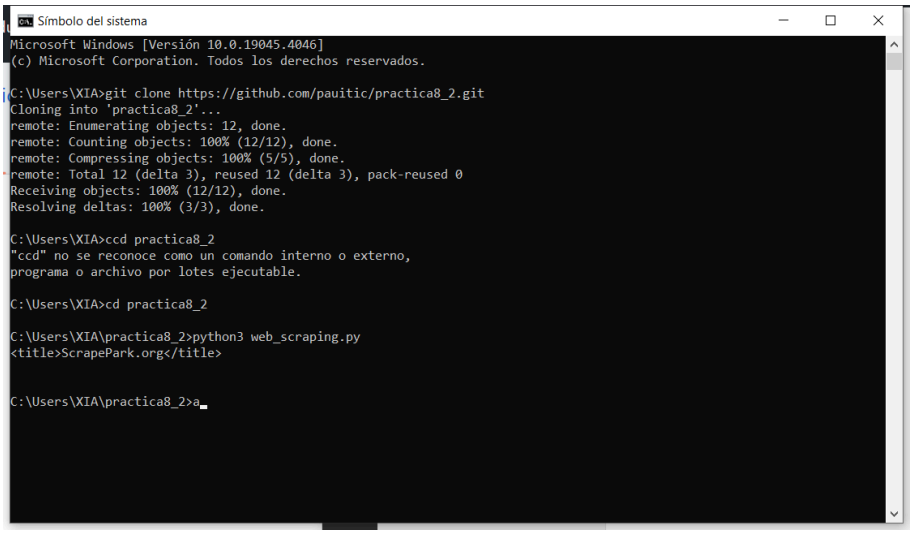
*** S'ha d'entregar l'enllaç del GIT al moodle.**

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 dóna 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:Retorna tots els nodes fills, incloent text, elements i altres nodes continguts dins dels elements <p> que tenen la classe attribution.
-Ruta 2:retorna únicament el text directament contingut dins dels elements <p> que tenen la classe attribution.

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:Recupera el text de tots els elements <a> que es troben directament dins d'un element <li>, el qual és fill directe d'un element <ul> amb la classe navbar-nav.
-Ruta 2:Recupera el contingut de text de tots els elements <a> continguts dins d'un element <li>, el qual pot ser fill directe o net dels elements <ul> amb la classe 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]`

div
└── **h5 (sexta instancia)**

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

div[class='carousel-item']
└── **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/div/div/div/div/p[3]/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):
//header//a/img


```
images/logo.svg
```

e. Troba la ruta fins a l'**atribut *src*** de les imatges amb ***alt="Client"***.
//section[5]//div[@class='carousel-inner']//div[@class='img-box-inner']//img

```
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 Scateboard 12".** [**Pista**: busca la utilitat de la funció *normalize-space()*].
//section[3]//div[@class='row']/div[12]//div[@class='detail-box']/h5/node()

```
<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 Scateboard 12".**

//section[3]//div[@class='row']/div[12]//div[@class='detail-box']/h5/text()

```
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:
//tr[td[contains(text(),'Blue')]]/td/text()

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

j. Troba la ruta que imprimeix **els preus del *longboard*** que es troben a la 4a columna de la taula **pintats en vermell**.
//tr/td[4]/text()

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

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 "|"]
//th[contains(text(), 'Skate')]/text() | //tr[td[contains(text()'$110')]]/td[1]/text()

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

```
Skate
Special
```

I. Troba la ruta a **tots els preus** dels objectes “Purple” **excepte el preu** que està pintat en vermell.
//tr/td[contains(text(),'Purple')]/td[not(contains(@style,'color:
red'))]

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