

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.

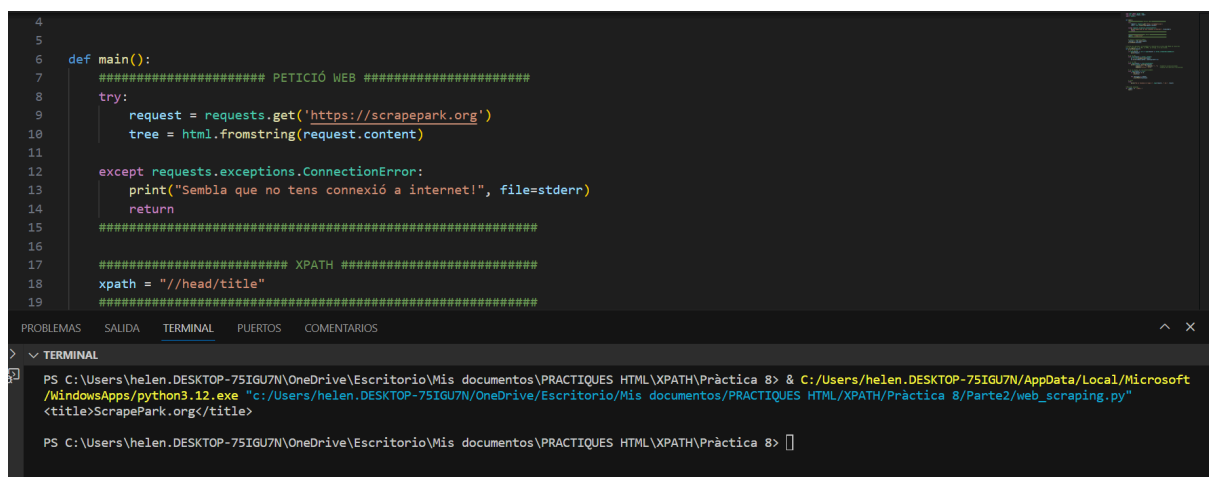
Guió

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/pauitc/practica8_2



```
4
5
6 def main():
7     ##### PETICIÓ WEB #####
8     try:
9         request = requests.get('https://scrapepark.org')
10        tree = html.fromstring(request.content)
11
12    except requests.exceptions.ConnectionError:
13        print("Sembla que no tens connexió a internet!", file=stderr)
14        return
15    #####
16
17    ##### XPATH #####
18    xpath = "//head/title"
19    #####
```

PROBLEMAS SALIDA TERMINAL PUERTOS COMENTARIOS

PS C:\Users\helen.DESKTOP-75IGU7N\OneDrive\Escritorio\Mis documentos\PRACTIQUES HTML\XPATH\Pràctica 8> & C:/Users/helen.DESKTOP-75IGU7N/AppData/Local/Microsoft/WindowsApps/python3.12.exe "c:/Users/helen.DESKTOP-75IGU7N/OneDrive/Escritorio/Mis documentos/PRACTIQUES HTML/XPATH/Pràctica 8/Parte2/web_scraping.py"

<title>ScrapePark.org</title>

PS C:\Users\helen.DESKTOP-75IGU7N\OneDrive\Escritorio\Mis documentos\PRACTIQUES HTML\XPATH\Pràctica 8>

Exercici 2

- 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()`

```
© 2022
<span>All Rights Reserved</span>.

.

<a href="https://html.design/" target="_blank" rel="noopener noreferrer"
>Created with Free Html Templates</a>.

.
```

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

```
© 2022
.
.
```

La ruta 1 mostra tots els resultats i objectes de l'element p, mentre que la ruta 2 només mostra el contingut que siguin fills directes de l'element p.

ii. Barra simple vs barra doble

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

```
Home

Products
```

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

Home

About
Testimonials
Products

English
Spanish

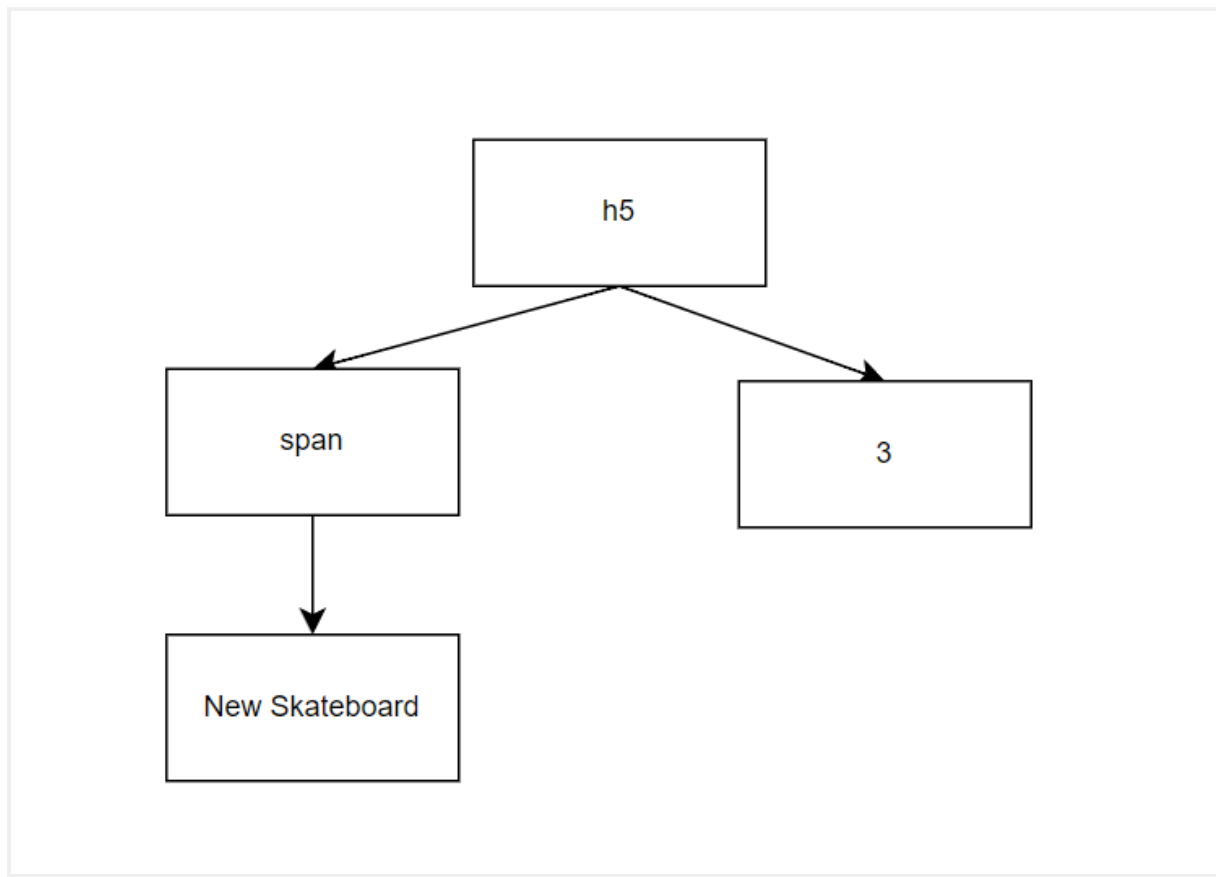
Contact 1
Contact 2

Amb la ruta 1 mostra només el contingut dintre del a de dintre de l'element `` dintre de l'element `ul` amb atribut `class navbar-nav`. Mentre que la ruta 2 mostra el contingut de tots els elements a dintre d'un element `li`.

- 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).

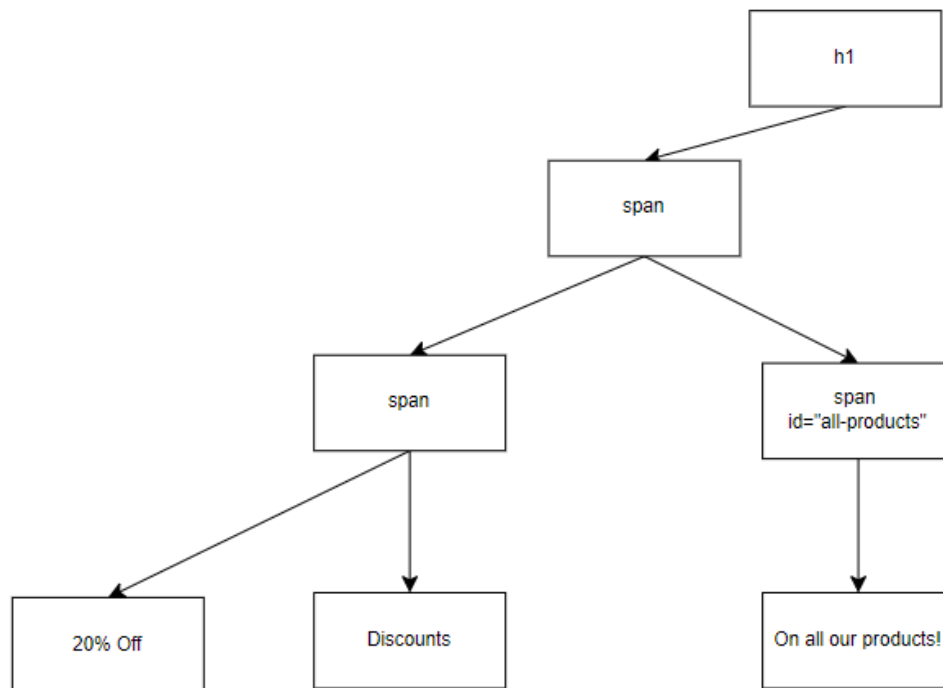
- i. `(//div/h5) [6]`

```
<h5>  
    <span>New Skateboard</span> 3  
</h5>
```



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

```
<h1>
    <span>
        <span>Discounts</span><br>20% Off
    </span>
    <br>
    <span id="all-products">On all our products!</sp
an>
</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[@class="information-f"]/p[3]/span/node()
```

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/footer//div[@class="logo-footer"]/a/img/@src
```

- 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

```
/html/body/section[@class='client-section  
layout-padding']//div/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:

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

Resposta:

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

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</span> 12  
</h5>
```

```
//div[@class='detail-box']/h5[normalize-space() = 'New Skateboard 12']
```

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

```
//div[@class='detail-box']/h5[normalize-space() = 'New Skateboard  
12']/../h6/node()
```

110

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

```
/html/body/table/tbody/tr[node() = 'Blue']/td/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

```
//th[@style= 'color: red;]/text() | /html/body/table/tbody/tr/td[@style  
= 'color: red;]/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 “[”]

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

Skate
Special

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

```
/html/body/table/tbody/tr[node() = 'Purple']/td[not (@style)]
```

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