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

XPATH amb HTML
PRACTICA 8.2

Realitzat per :

Sergio y Osama.
Dam1B

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.
<title>ScrapePark.org</title>

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()

Respuesta: La «Ruta 1» muestra todos los nodos y su contenido que contiene div con el atributo `'attribution'.La «Ruta2» esta mostrando con las mismas condiciones el texto que contiene 'div'.`

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>.
```

Práctica realizada por: Osama y Sergio

.

Ruta 2: `//div[@class='attribution']/p/text()`
`© 2022`
`.`

ii. Barra simple vs barra doble

Respuesta:En la «Ruta1» devuelve el contenido de texto dentro del elemento'<a>'que a su vez esta dentro de <li> y finalmente dentro de <ul>. En la «Ruta2» al no ser una ruta absoluta, permite devolver el contenido de texto dentro de todos los elementos <a> que cumplan con este patron.

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

`Products`

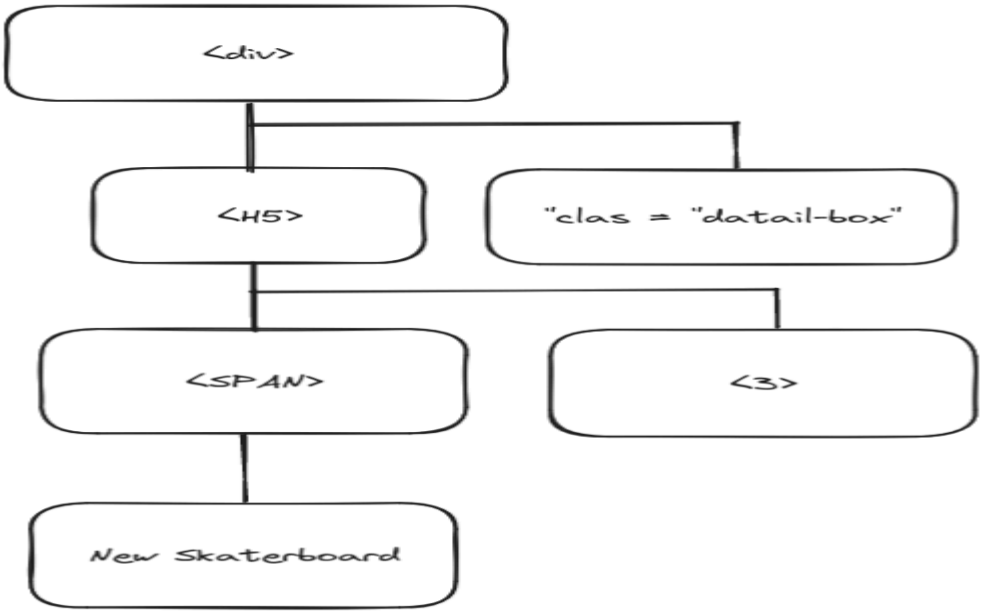
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

```
Home

About
Testimonials
Products

English
Spanish

Contact 1
Contact 2
```

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).
Pagina wb => Mermaid.live
https://excalidraw.com/

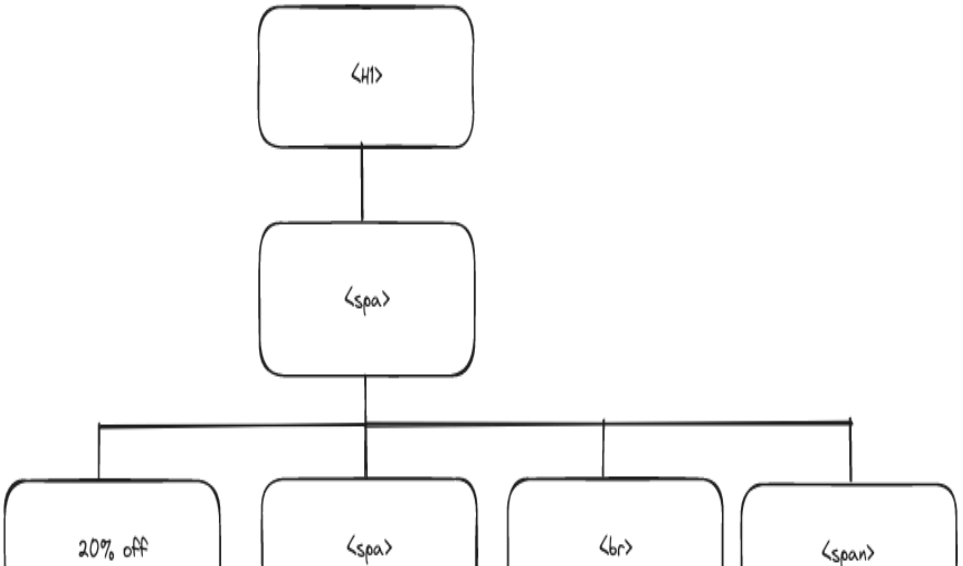
i. `(//div/h5)[6]`

```
<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!</span>
   </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
```

Respuesta: "/html//div[@class='information-f']/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):

Respuesta: "//header//img/@src"

```
images/logo.svg
```

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

Respuesta: "//img[@alt='Customer']/@src"

```
images/client-one.pngme
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()
```

Respuesta: "//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()*].

Respuesta: "//h5[span='New Skateboard' and text()[normalize-space()='12']]"

```
<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 Skateboard 12".**

Respuesta 1: "//h5[span='New Skateboard' and text()[normalize-space()='12']]/..//h6/node()"
Respuesta 2: "//h5[span='New Skateboard' and text()[normalize-space()='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.

i. Troba la ruta XPath a tots els **preus** dels **elements de color 'Blue'**. El resultat ha de ser el següent:

"//tr[td[text()='Blue']] /td[@class='text-center']/../td/node()"
"//tr[td[text()='Blue']]//td/node()"
//tr[td[text()='Blue']]/td[@class='text-center']/text() | //tr[td[text()='Blue']]/td[1]/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/th[@style]/text() | //tr[td[text()]]//td[2]/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 "|"]

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

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

```
Skate
Special
```

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>
```

//tr[4]/td[not(@style)]

"//tr[td[text()='Purple']]/td[1] | //tr[td[text()='$55']]/td[2] | //tr[td[text()='$60']]/td[3] | //tr[td[text()='$167']]/td[5]"

https://www.mclibre.org/consultar/xml/lecciones/xml-xpath.html