



# 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](https://github.com/pauitc/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()`

© 2022

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

.

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

.

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

© 2022

.

.

`node()` retorna tot el contingut que hi ha dins de `p` fills inclosos, `text()` retorn tan sols nodes de `text`

## 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

Aboutsexooo

Testimonials

Products

English

Spanish

Contact 1

Contact 2

La diferència entre la primera ruta i la segona és que en la primera, la barra simple només retorna tots els nodes `'li'` que són fills directes de `'ul'` amb la classe `'navbar-nav'` i el segon es que utilitza doble barra llavors tots els elements `'li'` que són descendents de `'ul'` amb la classe `'navbar-nav'`q independentment del nivell d'anidament.

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).

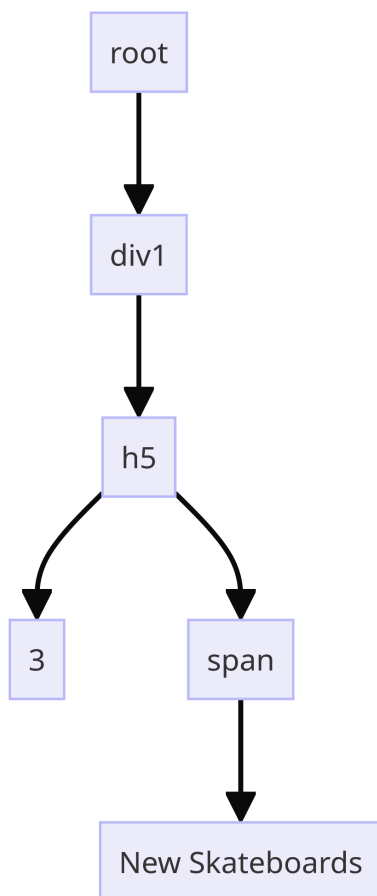
i. `//*[@div/h5][6]`

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

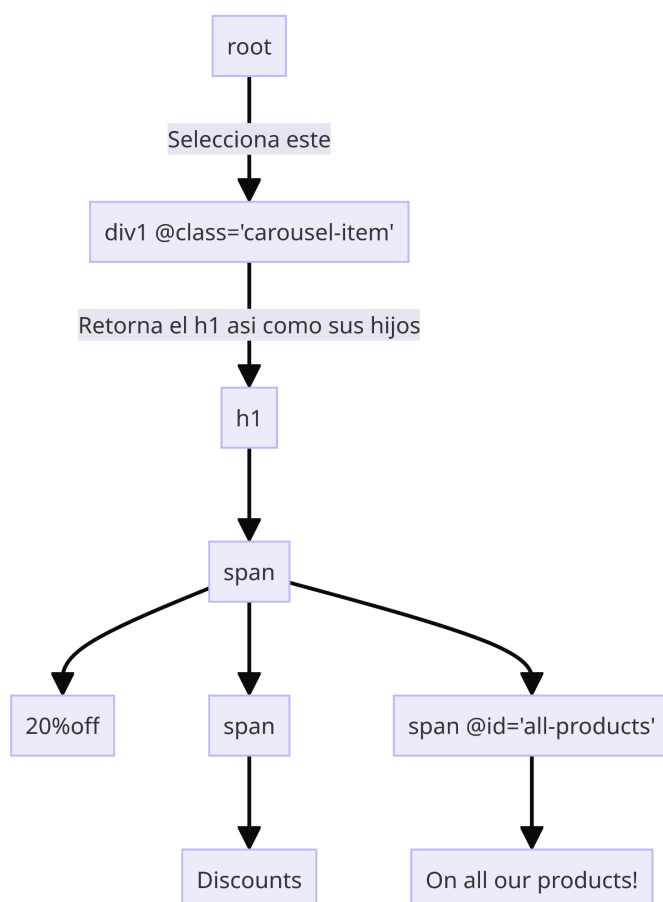
ii. `//*[@div[@class='carousel-item']][1]//h1`

```
<h1>
  <span>
    <span>Discounts</span><br>20% Off
  </span>
  <br>
  <span id="all-products">On all our products!</span>
</h1>
```

i.



ii.



## 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/div[1]/div[2]/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):

```
/html/body/footer/div/div/div/div[1]/div[1]/a/img/@src
```



images/logo.svg

- e. Troba la ruta fins a l'**atribut src** de les imatges amb **alt="Cliente" or "Costumer"**.

```
(//img[@alt='Customer'])/@src
```

```
//div[@class='img-box-inner']/img/@src
```

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]/span/text()
```

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

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()* ].

```
//div[@class='detail-box']/h5[text()='12']  
//h5[span='New Skateboard' and text()='12']  
//h5[span='New Skateboard']/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”.

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

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:

```
//tbody/tr/td[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**.

```
//table//th[@style='color: red;']/text() | //td[@style='color: red;']/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']/../../thead/tr/th[text()='Skate']/text() |  
//td[text()=' $110']/../td[text()='Special']/text()
```

Skate

Special

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

```
//td[text()='Purple']/../node()[not(@style='color: red;')]
```

<td>Purple</td>

<td class="text-center">\$55</td>

<td class="text-center">\$60</td>

<td class="text-center">\$72</td>