



# Pràctica 8.2: Web Scraping (XPath)

Grup: Eric Ayxendri, Alex Borda Segarra, Jose Ignacio Loayza, Santiago Muñoz

## 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)

```
<title>ScrapePark.org</title>
```

## 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 noreferrer">
```

Devuelve los nodos dentro de la etiqueta "p" que pertenece a div con clase attribution.

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

```
© 2022
.
.
```

Devuelve solo el texto plano dentro de "p"

- ii. Barra simple vs barra doble

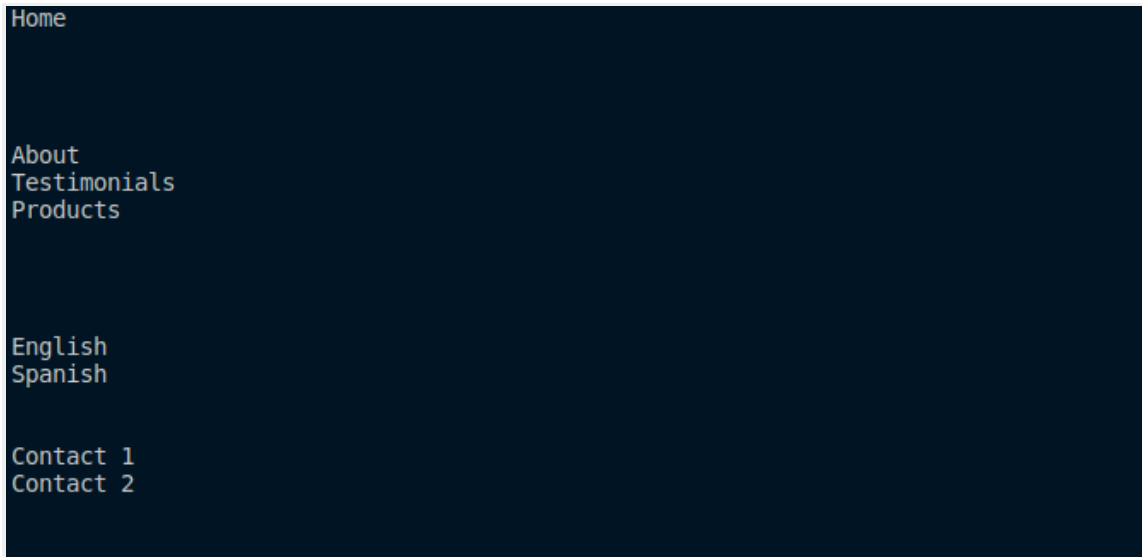
Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`

```
Home

Products
```

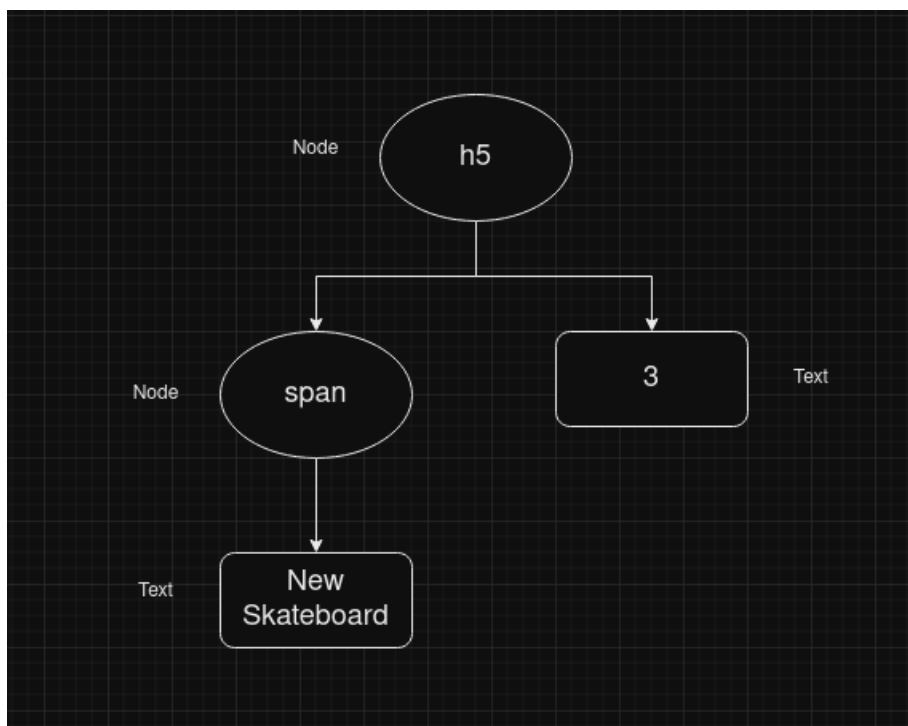
Ens retorna tot el text pla que estigui dins del node a, però aquest ha de penjar del node li, i aquest del node ul amb l'atribut class igual a 'navbar-nav'.

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

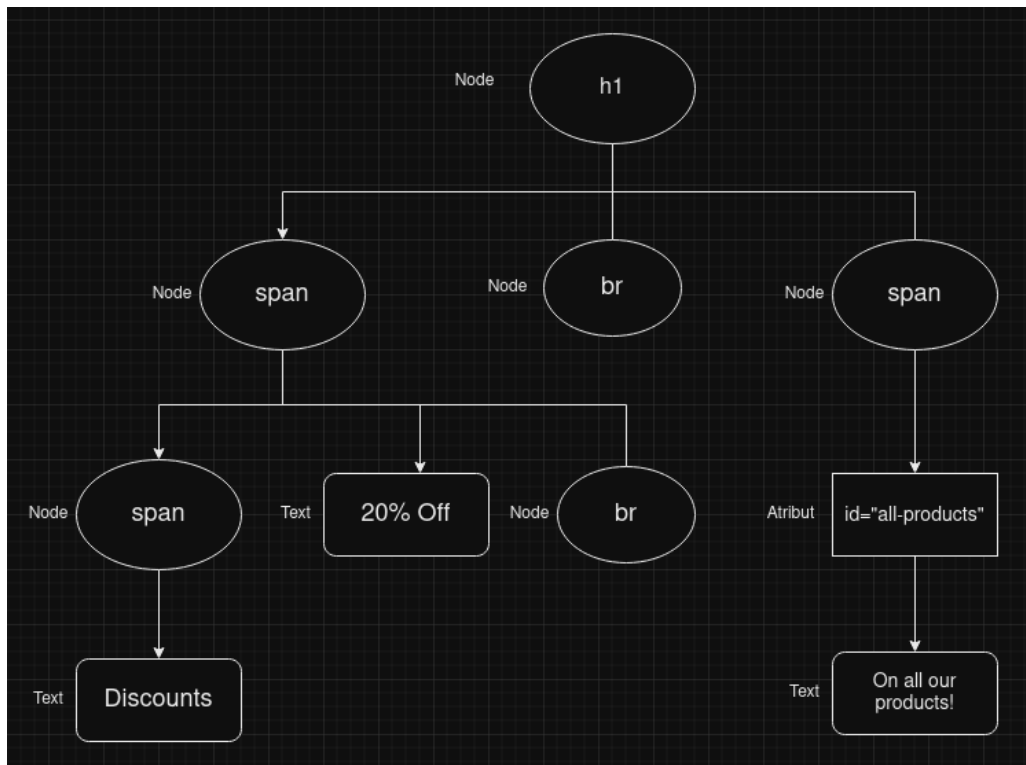


Devuelve el texto dentro de todos los links dentro de todos los li sin importar el camino para llegar a la etiqueta li, solo que estén dentro de la ul con la clase "navbar-nav"

- 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]`



ii. `//div[@class='carousel-item'][1]//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//strong[text()='EMAIL']/../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

`//header//nav//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

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

Fake Street 123

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

- 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>
```

```
/html//div/h5[text()='12 and node()='New Skateboard']
```

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

```
normalize-space(//div/h5[text()='12 and node()='New Skateboard'])
```

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:

Blue  
\$64  
\$70  
\$80  
\$85

```
/html//tbody/tr[node()='Blue']//td[position()>1]/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

```
/html//table//tr/node()[@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 ']
```

Skate

Special

```
//td[text()=' $110 ']/../../thead//th[2]/text() |  
//td[text()=' $110 ']/../td[1]/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>

```
//td[text()='Purple']/../td[not(@style)]
```