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.

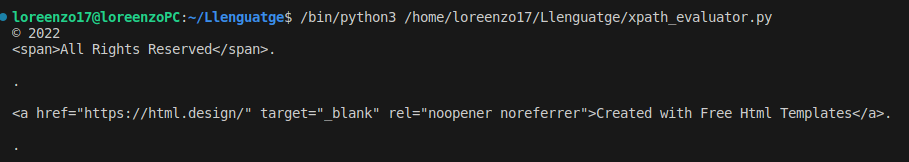
<https://github.com/pauitic/practica8_2>

## Exercici 2

1. 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.
   1. node() vs text()

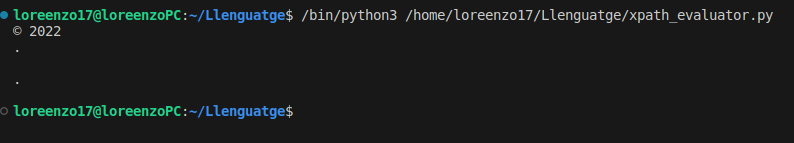
**Ruta 1**:  **//div[@class='attribution']/p/node()**

**Aquesta expressió retorna tots els nodes fills de l'element p que és fill de l'element div amb l'atribut class igual a 'attribution'. Això inclou tant els nodes de text com els elements HTML continguts dins de l'element p.**

****

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

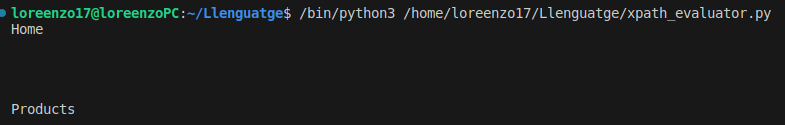
**En canvi, aquesta expressió només retorna els nodes de text directament fills de l'element p. Per tant, només obtens els continguts de text directament visibles sense cap element HTML adjacent.**

****

* 1. Barra simple vs barra doble

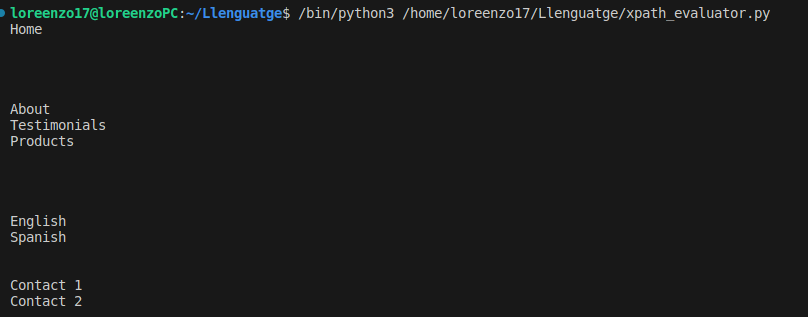
**Ruta 1**:  **//ul[@class='navbar-nav']/li/a/text()**

**Aquesta expressió retorna el text contingut dins dels elements a que són fills directes dels elements li, que a la vegada són fills de l'element ul amb l'atribut class='navbar-nav' pero nomes retorna Home i Products ja que son els continguts de l’element a.**

****

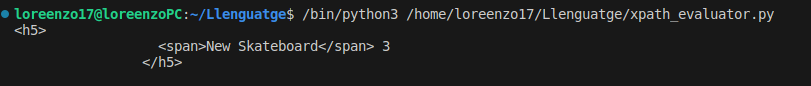
**Ruta 2**:  **//ul[@class='navbar-nav']//li/a/text()**

**Aquesta expressió retorna els elements a que són fills directes dels elements li, que a la vegada són fills directes de l'element ul amb l'atribut class='navbar-nav' i en aquesta ocasió si que retorna tot el contingut de l’element a incloent el text .**



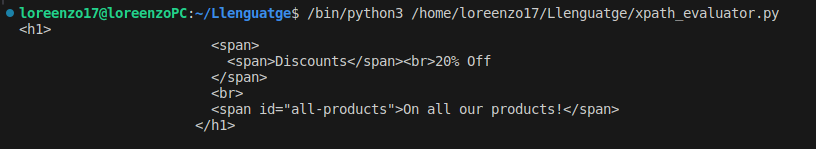
1. Representa, en forma d’arbre l’estructura HTML que resulta d’avaluar la següent ruta XPath (pots ignorar els salts de línia i espais).
   1. **(//div/h5)[6]**

**Aquesta expressió selecciona el sisè element h5 que apareix a la pàgina web com aqui podem veure el contingut complet d'aquest sisè element h5 com inclou el contingut del element span, que seria New Skateboard**

****

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

**En aquest cas selecciona l'element h1 que està dins del primer div amb l'atribut class igual a 'carousel-item'. Retorna el contingut complet d'aquest element h1, incloent els elements span i br que estan dins**



## 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.

1. 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//div[@class]/p[3]/span/node()**

****

sales@mail.com

1. 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):



**Hem escollit la imatge que està al footer:**

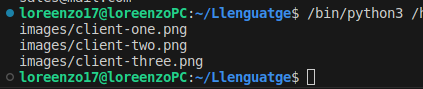
**//footer//img/@src**



images/logo.svg

1. Troba la ruta fins a l’**atribut *src*** de les imatges amb ***alt=”Client”***.

**//img[@alt='Client']/@src**



images/client-one.png

images/client-two.png

images/client-three.png

1. Troba la ruta fins a l’**adreça** de la pàgina web ***“Fake Street 123”*.** Fes que l’adreça XPath parteixi la següent ubicació:

**//footer//div[@class='information-f']/p[1]/span/node()**

****

Fake Street 123

1. Troba la ruta que arriba fins al **<h5>** del **“New Scateboard 12”.** [**Pista**: busca la utilitat de la funció *normalize-space()* ].

<h5> <span>New Skateboard</span> 12 </h5>

1. Partint de la ruta de l’apartat anterior, Troba la ruta que arriba fins al **preu** (text) del **“New Scateboard 12”.**

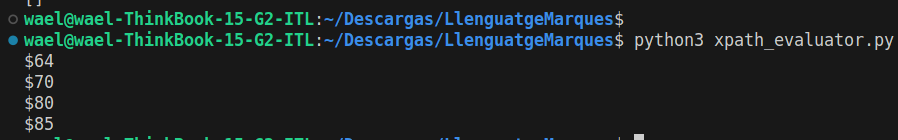
$110

## 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.

1. 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()



Blue

$64

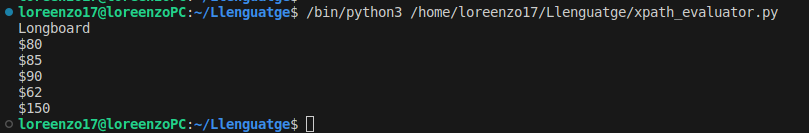
$70

$80

$85

1. Troba la ruta que imprimeix **els preus del *longboard***que es troben a la 4a columna de la taula **pintats en vermell**.

//body/table//tr/th[4]/text() | //body/table//tr/td[4]/text()



Longboard

$80

$85

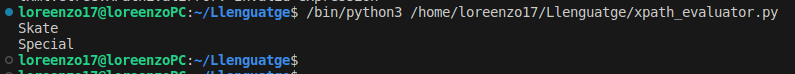
$90

$62

$150

1. **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[2]/text()|//td[text()=’$110’]/../td[1]/text()

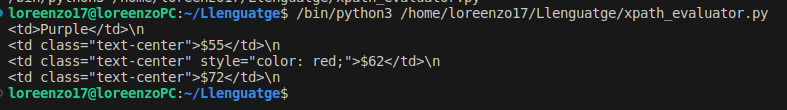


Skate

Special

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

//td[../td[text()='Purple'] and position() != 3]



<td>Purple</td>

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

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

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