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

```
C:\Users\silvi>cd C:\Users\silvi\OneDrive\Documentos\M04\M4-24

C:\Users\silvi\OneDrive\Documentos\M04\M4-24>git clone https://github.com/pauitic/practica8_2
Cloning into 'practica8_2'...
remote: Enumerating objects: 12, done.
remote: Counting objects: 100% (12/12), done.
remote: Compressing objects: 100% (5/5), done.
remote: Total 12 (delta 3), reused 12 (delta 3), pack-reused 0
Receiving objects: 100% (12/12), done.
Resolving deltas: 100% (3/3), done.
```

```
PS C:\Users\silvi\OneDrive\Documentos\M04\M4-24> & C:/Users/silvi/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/silvi/OneDrive/Documentos/M04/M4-24/practica8_2/web
_scraping.py
<title>ScrapePark.org</title>

PS C:\Users\silvi\OneDrive\Documentos\M04\M4-24>
```

Exercici 2

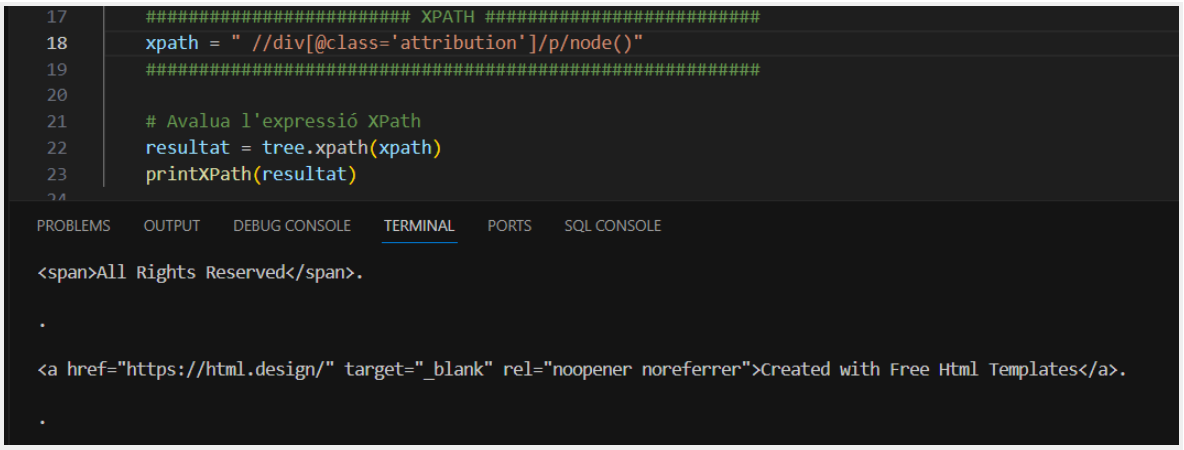
a. Executa les següents rutes XPath i observa el resultat que dona cada una. A continuació, explica les diferències que hi ha entre cada resultat i raona per què produeixen resultats diferents.

i. node() vs text()

Ruta 1: `//div[@class='attribution']/p/node()`
Ruta 2: `//div[@class='attribution']/p/text()`

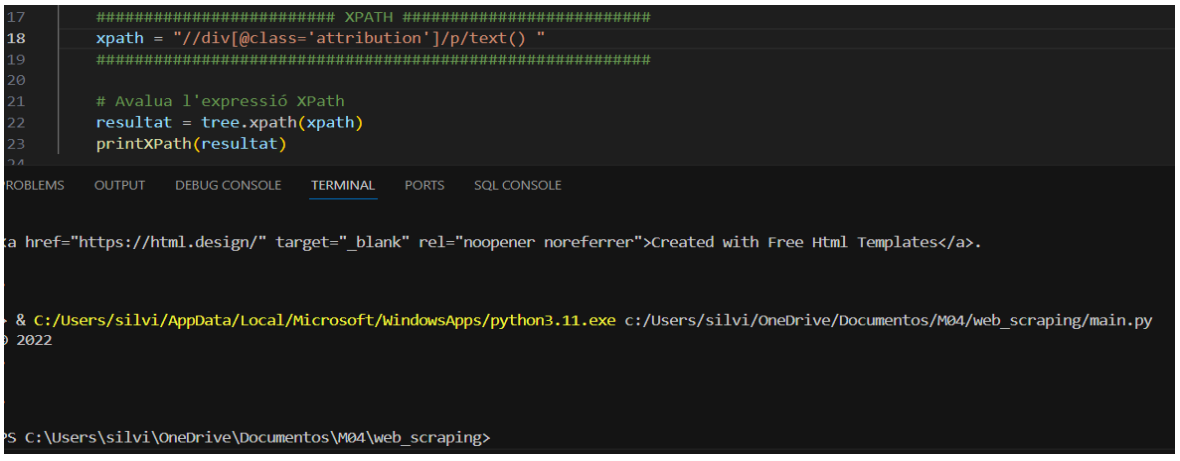
Ruta 1

//div: Esto significa que estamos buscando todos los elementos ** ** en el documento XML.
[@class='attribution']: Aquí estamos filtrando los elementos ** ** que tienen un atributo class con el valor "attribution".
/p: Luego, estamos buscando un elemento **<p>** que sea hijo directo de los elementos ** ** seleccionados anteriormente.
/node(): Finalmente, estamos seleccionando todos los nodos hijos del elemento **<p>**. Esto incluye texto, comentarios, etc.



Ruta 2

Esta ruta selecciona específicamente todos los nodos de texto hijos del párrafo. Esto significa que sólo obtendremos el texto directamente contenido en el <p>, excluyendo cualquier texto que esté dentro de (como dentro de un dentro del <p>) y otros tipos de nodos. Es útil cuando solo te interesa el texto "plano" del párrafo y no cualquier marcado o elemento adicional que pueda contener

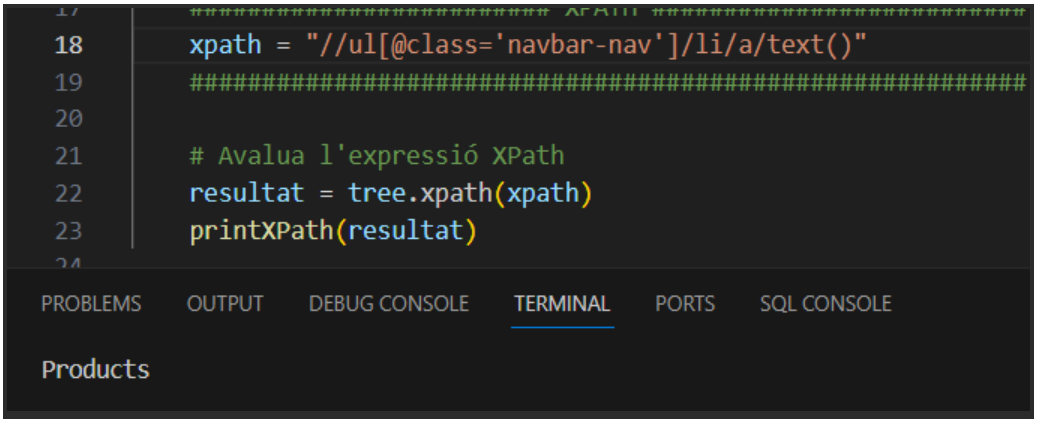


ii. Barra simple vs barra doble

Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

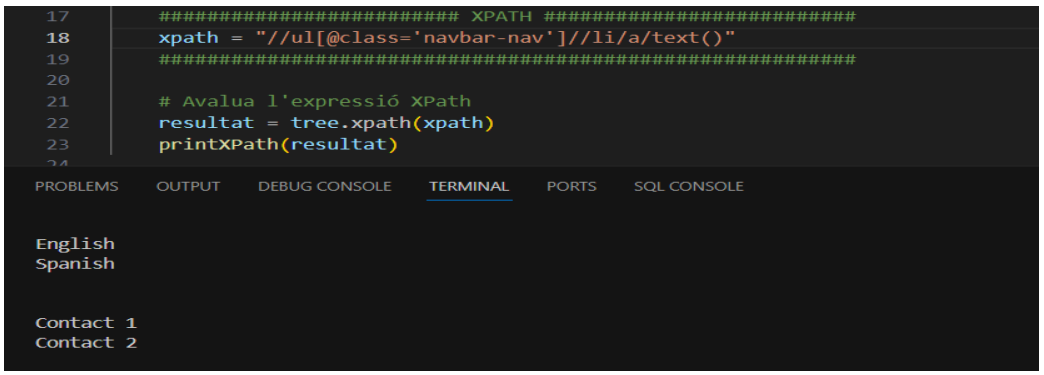
Ruta 1

Esta ruta selecciona el texto de los enlaces que son hijos directos de elementos <li> que, a su vez, son hijos directos del <ul> con clase navbar-nav. Esto significa que solo obtendrá el texto de los enlaces que están exactamente un nivel por debajo en la jerarquía de la lista especificada.



Ruta 2

Esta ruta selecciona el texto de los enlaces dentro de cualquier elemento <li> que sea descendiente (no necesariamente hijo directo) del <ul> con clase navbar-nav. El doble slash // antes de li indica que puede atravesar uno o varios niveles de la estructura del documento para encontrar cualquier <li> que cumpla con ser descendiente del <ul> especificado, no solo los que son hijos directos.



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`

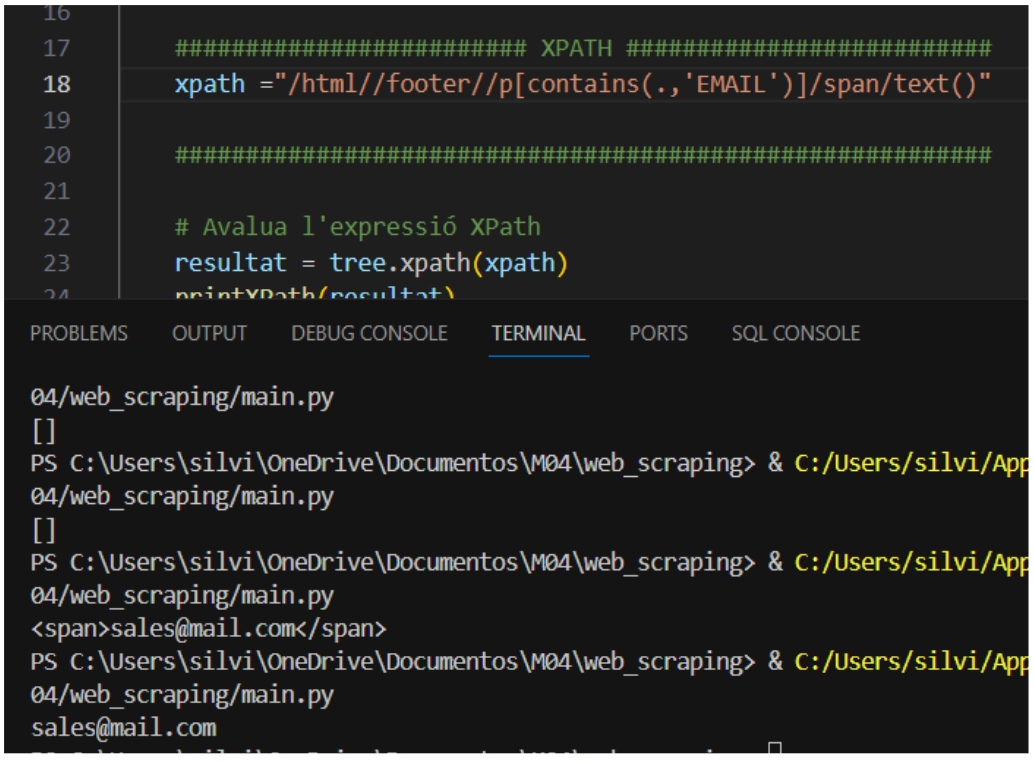
Exercici 3

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.

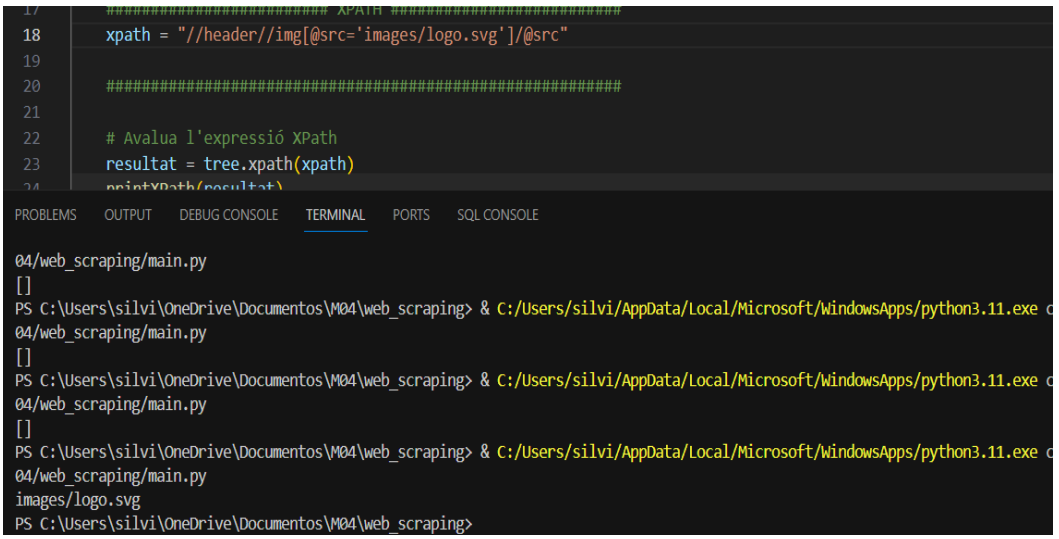
c. Troba la ruta que arriba al **correu** de contacte que es troba al **<footer>** de la pàgina. **Comença la ruta a l'etiqueta <html>**

`/html`

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):

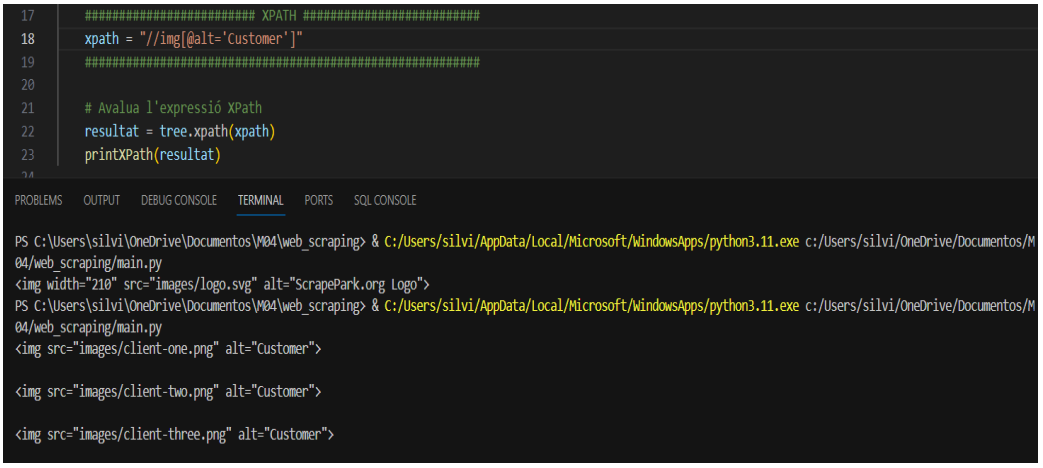


images/logo.svg

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

images/client-one.png
images/client-two.png
images/client-three.png

"//img[@alt='customer]'"



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]/span/text()

```
    ########################## XPATH ##########################
    xpath = "//div[@class='information-f']/p[1]/span/text()"
    ###########################################################

    # Avalua l'expressió XPath
    resultat = tree.xpath(xpath)
    printXPath(resultat)


# Funció que imprimeix correctament el resultat de la cerca amb XPath en fun
# si el resultat és un XML, un HTML, un string, o bé una llista.
def printXPath(result):
    # si és string
    if type(result) is str or type(result) is etree._ElementUnicodeResult:
        print(result)

    # si és xml
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  SQL CONSOLE

    main()
  File "c:\Users\silvi\OneDrive\Documentos\M04\web_scraping\main.py", line 22, in main
    resultat = tree.xpath(xpath)
               ^^^^^^^^^^^^^^^^^
  File "src\\lxml\\etree.pyx", line 1606, in lxml.etree._Element.xpath
  File "src\\lxml\\xpath.pxi", line 290, in lxml.etree.XPathElementEvaluator.__call__
  File "src\\lxml\\xpath.pxi", line 210, in lxml.etree._XPathEvaluatorBase._handle_resul
lxml.etree.XPathEvalError: Invalid expression
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/M
04/web_scraping/main.py
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()*].#----------consiste en llegar a la ruta hasta llegar new Scateboard

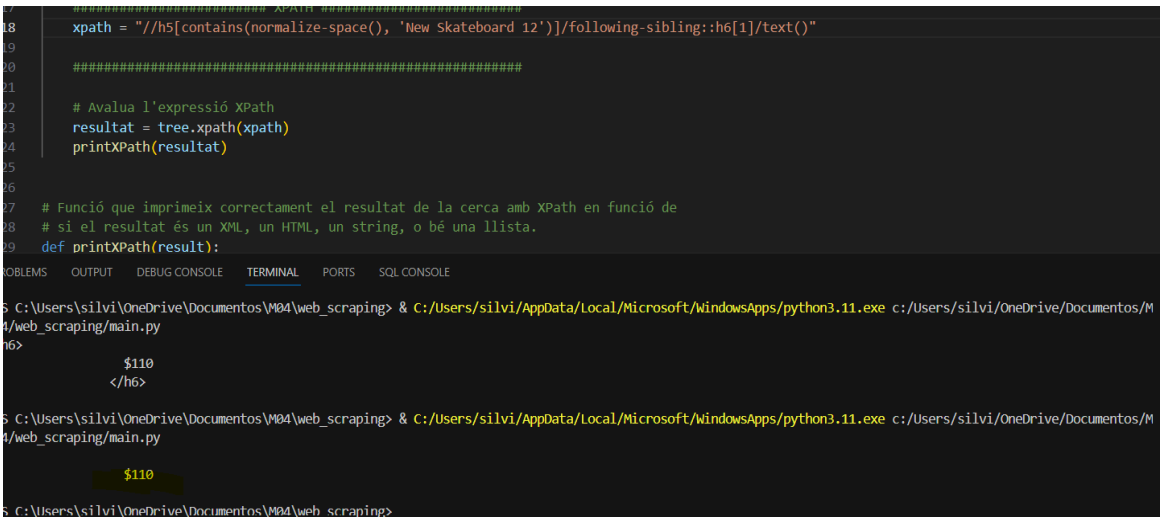
```
<h5>                    <span>New Skateboard</span> 12
</h5>
```

`//h5[normalize-space(.) = 'New Skateboard 12']`

```
    ########################## XPATH ##########################
    xpath = "//h5[contains(normalize-space(),'New Skateboard 12')]"
    ###########################################################

    # Avalua l'expressió XPath
    resultat = tree.xpath(xpath)
    printXPath(resultat)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  SQL CONSOLE

            </h5>

PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/silvi/OneDrive/Documentos/M
04/web_scraping/main.py
[]
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/silvi/OneDrive/Documentos/M
04/web_scraping/main.py
<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 Scateboard 12".**



$ 110

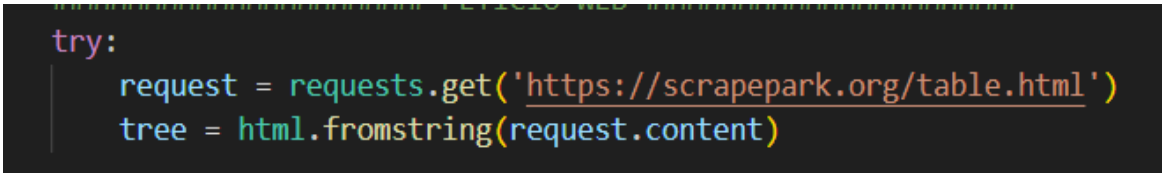
Exercici 4

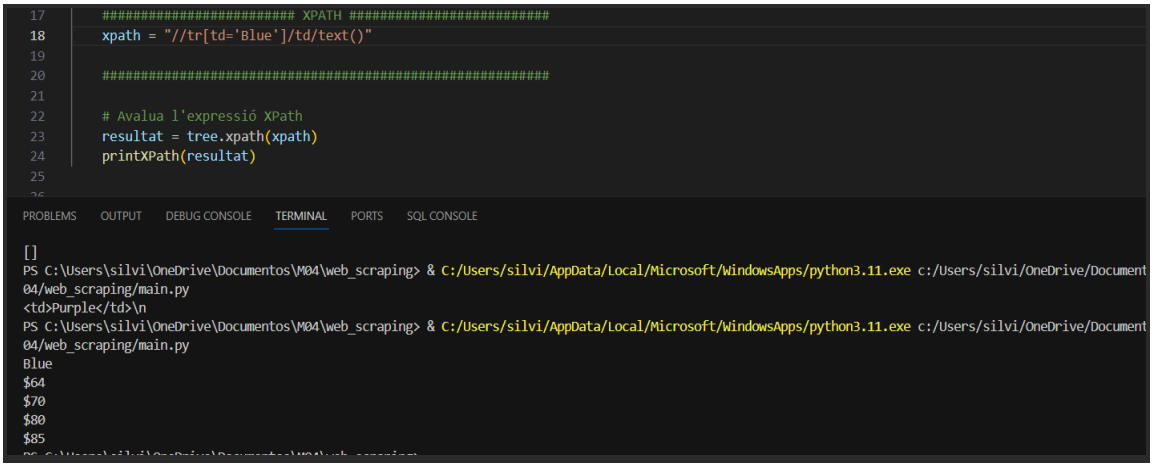
Canvia la ruta a https://scrapepark.org/table.html . Amb l'ajuda del navegador, comprova què hi ha dins d'aquesta pàgina i troba la ruta XPath dels següents elements.

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

```
Blue
$64
$70
$80
$85
```

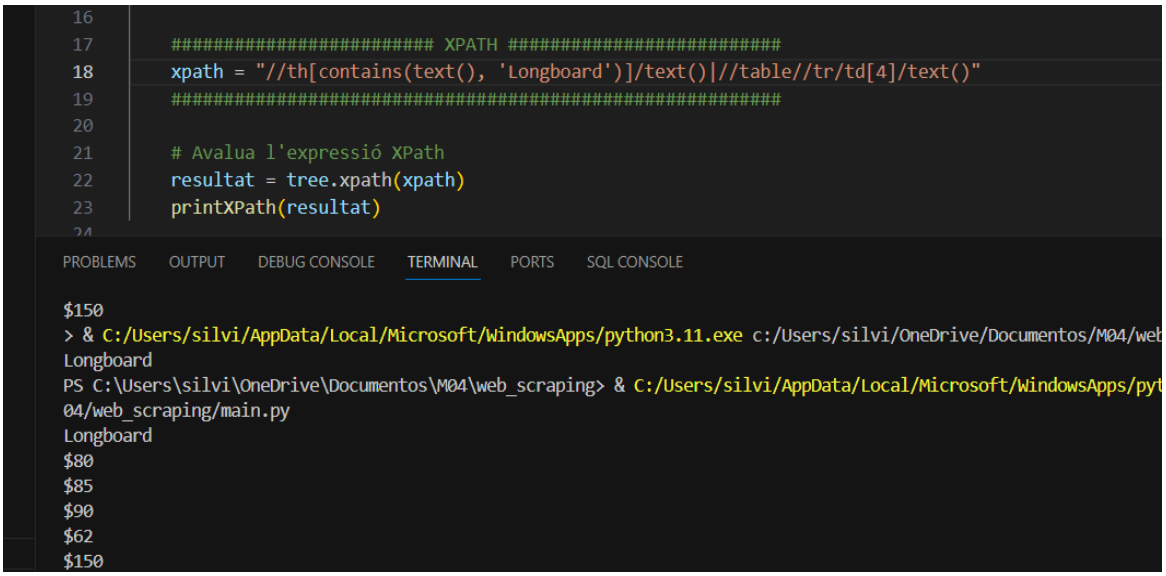
Cambiamos la ruta de 'https://scrapepark.org' a https://scrapepark.org/table.html.





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

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

```
Skate
Special
```

```
"//th[contains(text(), 'Skate')]/text()|//td[contains(text(), 'Special')]/text()"
```

```
16
17      ########################## XPATH ##########################
18      xpath = "//th[contains(text(), 'Skate')]/text()|//td[contains(text(), 'Special')]/text()"
19      ###########################################################
20
21      # Avalua l'expressió XPath
22      resultat = tree.xpath(xpath)
23      printXPath(resultat)
24
25
26  # Funció que imprimeix correctament el resultat de la cerca amb XPath en funció de
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  SQL CONSOLE

lxml.etree.XPathEvalError: Unregistered function
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsA
04/web_scraping/main.py
[]
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsA
04/web_scraping/main.py
Special
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsA
04/web_scraping/main.py
Skate
Special
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping>
```

I. 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>
```

```
18      xpath = "//td[text()='Purple']/following-sibling::td[@class='text-center'][not(contains(@style, 'color: red'))]"
19
20      ###########################################################
21
22      # Avalua l'expressió XPath
23      resultat = tree.xpath(xpath)
24      printXPath(resultat)
25
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  SQL CONSOLE

  File "src\\lxml\\xpath.pxi", line 290, in lxml.etree.XPathElementEvaluator.__call__
  File "src\\lxml\\xpath.pxi", line 210, in lxml.etree._XPathEvaluatorBase._handle_result
lxml.etree.XPathEvalError: Invalid expression
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/silvi/OneDrive/Documentos
04/web_scraping/main.py
[]
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping> & C:/Users/silvi/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/silvi/OneDrive/Documentos
04/web_scraping/main.py
<td class="text-center">$55</td>\n
<td class="text-center">$60</td>\n
<td class="text-center">$72</td>\n
PS C:\Users\silvi\OneDrive\Documentos\M04\web_scraping>
```