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. 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.

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

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()`
Selecciona tots els nodes fills de l'element p dins del div amb la classe d'atribució
Ruta 2: `//div[@class='attribution']/p/text()`
Selecciona només el contingut de text de l'element p dins del div amb la classe d'atribució

A diferència entre els resultats és que la primera ruta inclourà tots els nodes fills, mentre que la segona ruta només inclourà el text contingut dins de l'element p.

ii. Barra simple vs barra doble

Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`
Seleccionarà tots els nodes de text continguts dins dels elements a que són fills directes dels elements li que, al seu torn, són fills directes de l'ul amb la classe navbar-nav.

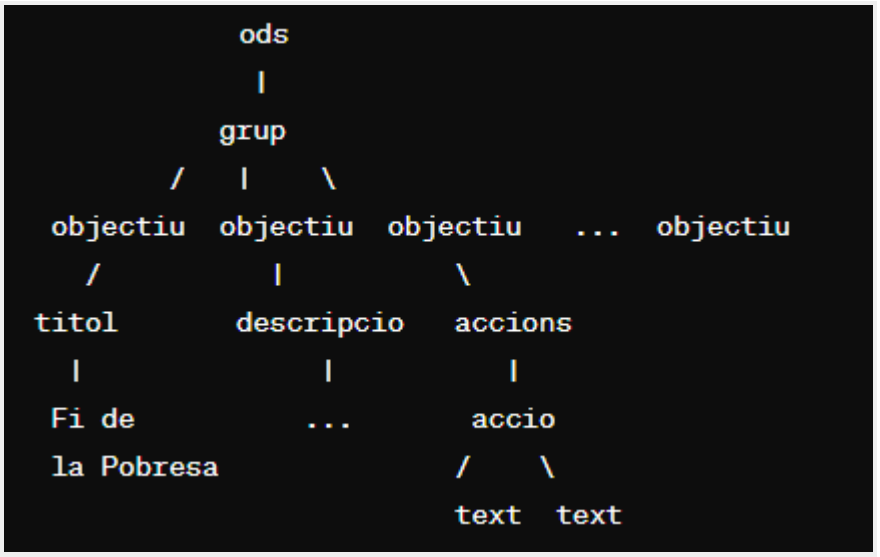
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

seleccionarà tots els nodes de text continguts dins dels elements a que es trobin en qualsevol profunditat dins dels elements li que, al seu torn, són fills de l'ul amb la classe navbar-nav.

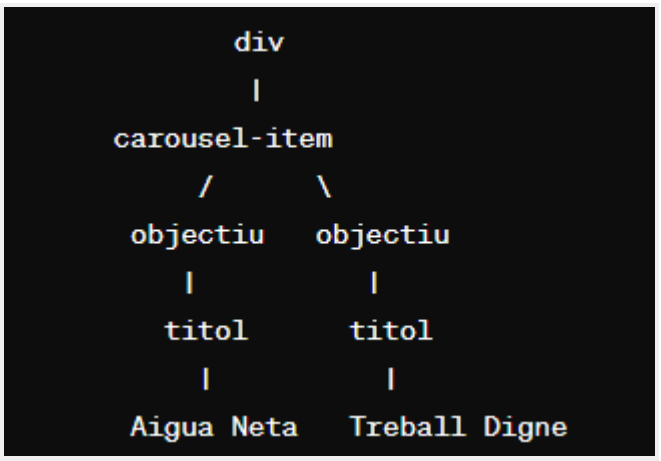
La diferència principal aquí és que la primera ruta només seleccionarà els nodes de text dels elements a que són fills directes dels elements li, mentre que la segona ruta seleccionarà els nodes de text dels elements a que es trobin en qualsevol nivell de profunditat dins dels elements li.

b. Representa, en forma d'arbre l'estructura XML 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

c. Descobreix 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`

```
<footer>
  <div class="container">
    <div class="row">flex
      <div class="col-md-4">
        <div class="full">
          <div class="logo-footer">…</div>
          <div class="information-f">
            <p>…</p>
            <p>…</p>
            <p>
              <strong>EMAIL</strong>
              ": "
              <span>sales@mail.com</span> == $0
```

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`

```
<!DOCTYPE html>
<html>
  <head>…</head>
  <body>
    <div class="hero-area">flex
      <!-- header section strats -->
      <header class="header-section">
        <div class="container">
          <nav class="navbar navbar-expand-lg custom-nav-container">flex
            <a class="navbar-brand" href="/">
              <img width="250" src="images/logo.svg" alt="ScrapePark.org Log
              o"> == $0
```

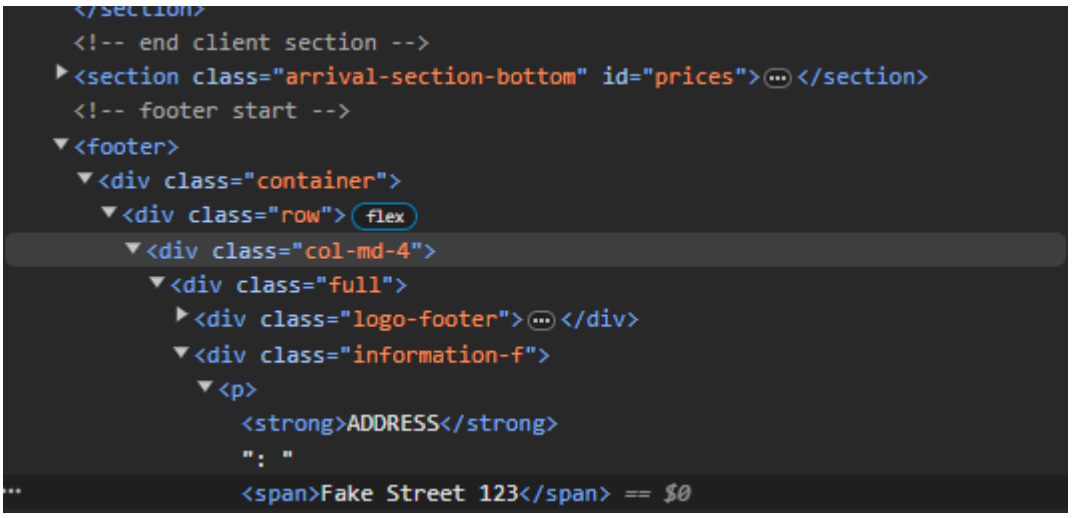
e. Troba la ruta fins a l'**atribut *src*** de les imatges amb ***alt="Client"***.

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

f. Troba la ruta fins a l'**adreça** de la pàgina web ***"Fake Street 123".*** Fes que l'adreça XPath parteixi la següent ubicació:

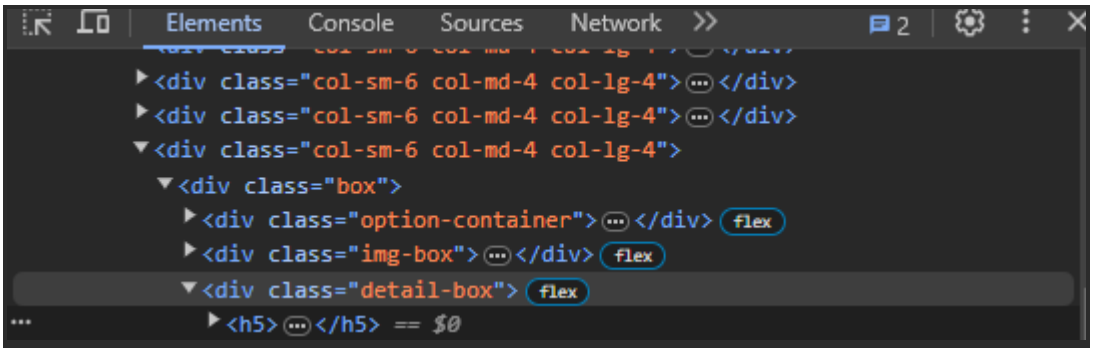
```
//div[@class='information-f']/p[1]/strong/text()
```

```
Fake Street 123
```



g. 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>
```



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

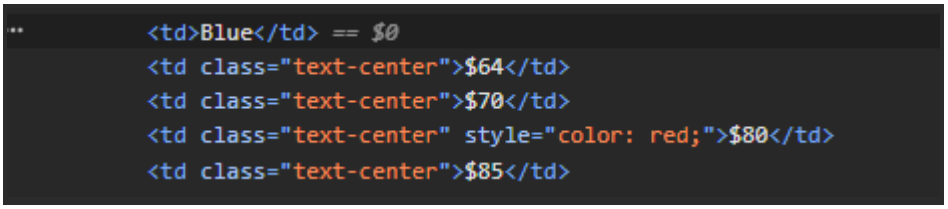
12

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:

```
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**.

```
Longboard
$80
$85
$90
$62
$150
```

```
<th class="text-center">Skate</th>
<th class="text-center">Cruiser</th>
<th class="text-center" style="color: red;">Longboard</th> == $0
<th class="text-center">Freeboard</th>
```

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>Blue</td>
td class="text-center">$64</td>
td class="text-center">$70</td>
td class="text-center" style="color: red;">$80</td
td class="text-center">$85</td>
r>
></tr>
></tr>
></tr>
>
td>Special</td>
td class="text-center">$110</td> == $0
```

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>Purple</td> == $0
<td class="text-center">$55</td>
<td class="text-center">$60</td>
<td class="text-center" style="color: red;">$62</td>
<td class="text-center">$72</td>
```