Grup: Lishi i Karen

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

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()

La diferencia entre utilitzar la funció node() de text() és que node() agafa tant els continguts de valors tipus texts i els elements nodes que conté el node "**p**" que penja de un div que té un atribut de nom class de valor que correspon a "atribution" i en canvi, text(), només agafa el text que conté directament el node "**p**", perquè la resta de text està contingut a altres subnodes "**span**" i "**a**" i per això no agafa a aquests textos, en canvi la funció node() també selecciona altres subnodes "span" i "a" inclosos.

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

```
© 2022
<span>All Rights Reserved</span>.

.

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

.
```

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

```
© 2022
.

.
```

ii. Barra simple vs barra doble

La diferència que hi ha entre la barra simple i la barra doble es que la 1r ruta al ser una barra simple obligatoriament té penjar directament del primer node() **"ul"** amb atribut de valor 'navbar-nav', només els subnodes() fills **"li"** directes del node pare **ul** de l'atribut indicat, ens retorna el valor de tipus text que conté el node() "**a".** En canvi la segona ruta de barra doble ens mostra el resultat a partir del primer node **"ul"** qualsevol subnode "**li**" que pengi directament o no, dins del node **"ul"**, que pot penjar inclus d'un altre subnode() fill del node pare **"ul",** com ara nodepare(ul)>li>ul>li>a. Que retorna el valor text del node() **"a"** que penja de un node pare **"li"** que penja desde qualsevol lloc desde **dins** del node pare **"ul"** amb l'atribut indicat.

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

```
Home


Products
```

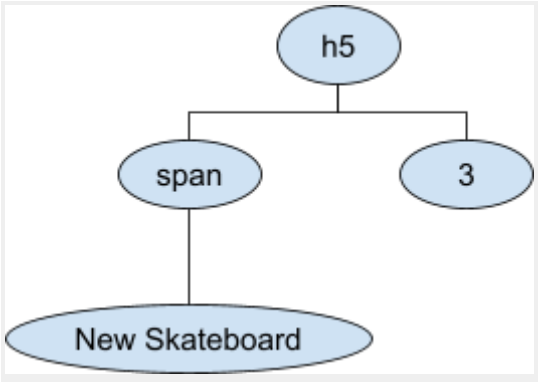
Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

```
Home


About
Testimonials
Products


English
Spanish


Contact 1
Contact 2
```

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).

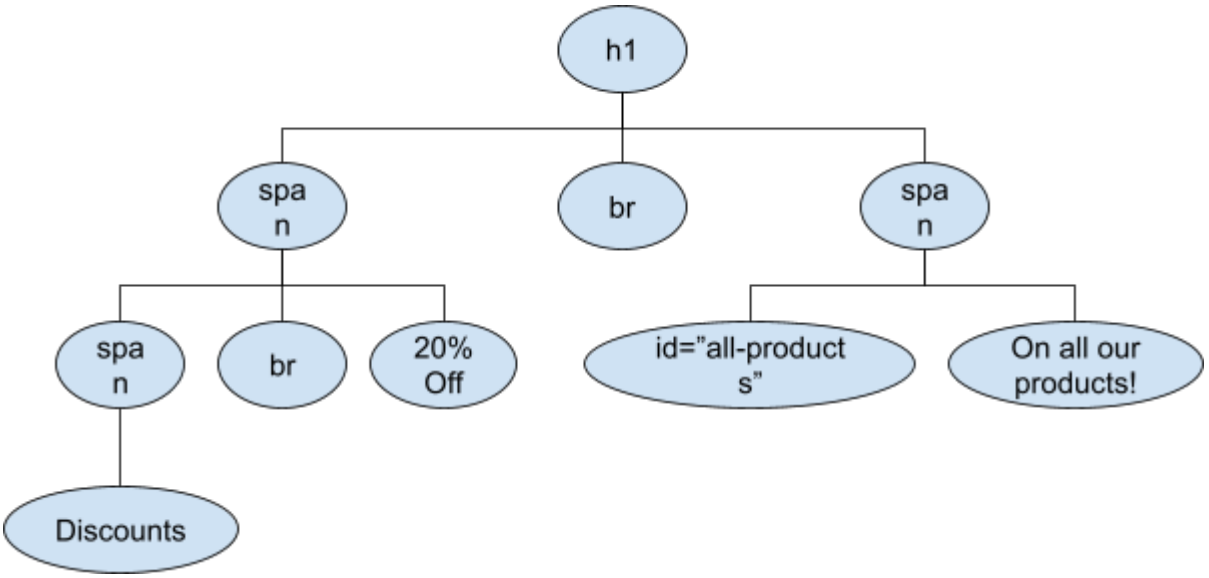
i. `(//div/h5)[6]`

```
<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>
```



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
```

/html//div[@class='information-f']/p[3]/span/text()

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//@src

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']/@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>
```

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

h. Partint de la ruta de l'apartat anterior, Troba la ruta que arriba fins al "**preu" -> h6** (text) del **"New Scateboard 12".**

```
$110
```

//h5[span='New Skateboard']/text()[normalize-space()='12']/../../h6/text()

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
```

//td[text()='Blue']/../td/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
```

Es poden fer de les següents dues maneres:

//tr/*[@style='color: red;']/text()
//tr/node()[@style='color: red;']/text()

A més si volguessim filtrar per columna, també podem fer:

//tr[//@style='color: red;']/*[4]/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']/../*[1]/text()|//td[text()='$110']/../../..//th[2]/text()

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>
```

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