



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. 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/pauitc/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()`

© 2022

`All Rights Reserved`

`Created with Free Html Templates`

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

Amb el `node` et mostra tot el que està per sota de l'etiqueta, incluint informació i etiquetes de on estem, en aquest cas de `p`, i `text` et mostra només l'informació que hi ha dins de `p`, s

ii. Barra simple vs barra doble

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

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

La diferència entre barra simple i barra doble és que amb barra doble, seleccionem tot el que hi ha dins del que escrivim després. I, amb barra simple, només selecciona un.

- 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/body/footer/div/div/div[1]/div/div[2]/p[3]/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):

`/html/body/div/header/div/nav/a/img/@src`



images/logo.svg

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

`/html/body/section/div/div[2]/div/div/div/div/div/div/img[@alt='Customer']/@src`

f.

images/client-one.png

images/client-two.png

images/client-three.png

- g. 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()../  
../span/text()
```

Fake Street 123

- h. 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>
```

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

12

```
/html/body/section[3]/div/div[2]/div[12]/div/div[3]/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.

- j. 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/body/table/tbody/tr/td[text() = 'Blue']/../td/text()
```

- k. 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

```
tml/body/table/thead/tr/th[text() = \"Longboard\"]/text() |  
/html/body/table/tbody/tr/td[@style=\"color: red;\"]/text() "
```

- l. 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

L'element que té un preu de \$110 és el següent:

Nom de l'article: Skate

Color: Special

- m. 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>
```

```
/html/body/table/tbody/tr/td[text() = 'Purple']/../td
```