



Generalitat de Catalunya Ajuntament de Barcelona

# 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 <a href="https://scrapepark.org/">https://scrapepark.org/</a>. 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()
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 p\u00e4gina. Comen\u00fca la ruta a l'etiqueta <html>

/html/body/footer//

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

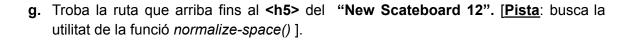
e. Troba la ruta fins a l'atribut src de les imatges amb alt="Client".

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



```
<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 <a href="https://scrapepark.org/table.html">https://scrapepark.org/table.html</a> . 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

**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

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

```
Purple
>
>
$55
$60
$72
```