



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.

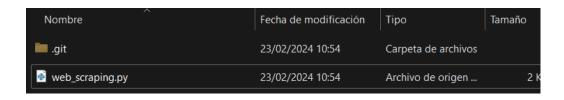
Guió

Amb l'ajuda de l'inspector d'elements del navegador, investiga com està formatada la pàgina 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()

El node() ens imprimeix tot el que hi ha dintre de la ruta.

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

El text() ens mostra només la informació (el text) de dintre de les etiquetes independentment de si tenim etiquetes fills.

ii. Barra simple vs barra doble

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

Aquesta ruta ens retorna tots els nodes fills de text els quals el seu fill té una llista "li" amb l'atribut @class='navbar-nav' i al tenir "/" ens selecciona els fills directes d'aquest.

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

Aquesta ruta ens selecciona només els fills directes, i utilitza "//" per seleccionar qualsevol node que coincideixi amb l'atribut: @class='navbar-nav

- **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]

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>

sales@mail.com

/html/body/footer//div/div[1]/div/div[2]/p[3]/span/node()

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

//div/header/div/nav/a/img/@src

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

```
//*[@id='carouselExample3Controls']/div[1]//div[1]//div[1]//div/img/@
src
```

images/client-one.png

//*[@id='carouselExample3Controls']/div[1]/div[2]//div[1]//div/img/@
src

images/client-two.png

//*[@id="carouselExample3Controls"]/div[1]/div[3]//div[1]//div/img/@
src

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:

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

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 /html/body/table/tbody/tr[1]/td[1]/text()
$64: /html/body/table/tbody/tr[1]/td[2]/text()
$70: /html/body/table/tbody/tr[1]/td[3]/text()
$80: /html/body/table/tbody/tr[1]/td[4]/text()
$85: /html/body/table/tbody/tr[1]/td[5]/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 /html/body/table/thead/tr/th[4]/text()
$80 /html/body/table/tbody/tr[1]/td[4]/text()
$85 /html/body/table/tbody/tr[1]/td[5]/text()
$90 /html/body/table/tbody/tr[3]/td[4]/text()
$62 /html/body/table/tbody/tr[4]/td[4]/text()
$150 /html/body/table/tbody/tr[5]/td[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 "|"]

Nom: Skate
Color: Special

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

```
Purple
Class="text-center">$55
Class="text-center">$60
Class="text-center">$72

Class="text-center">$72

//td[contains(text(), 'Purple')]/td[position() > 1]
```