



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()`

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

La diferència entre la ruta 1 i la ruta 2 és que la primera llista tot el contingut dintre de la `<p>` juntament amb les seves etiquetes corresponents, en canvi la segona llista tot el contingut dintre de la `<p>` però sense les seves etiquetes, és a dir l'informació de les etiquetes.

ii. Barra simple vs barra doble

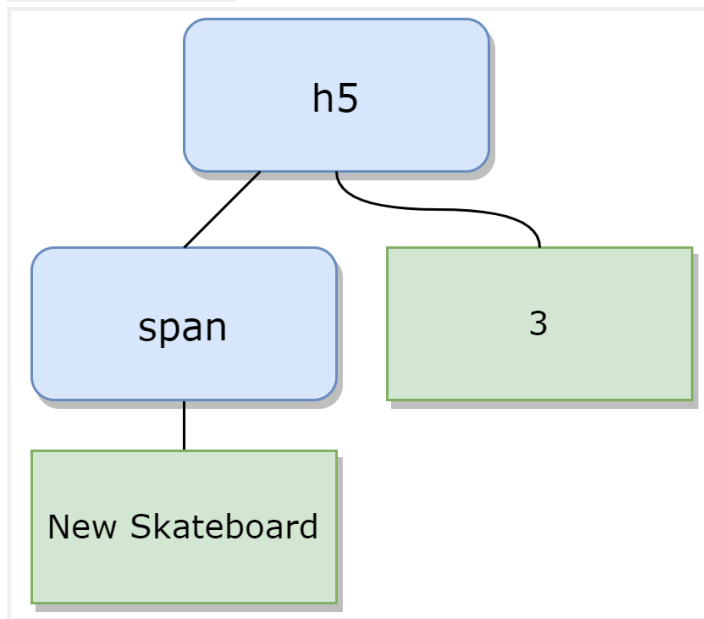
Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`

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

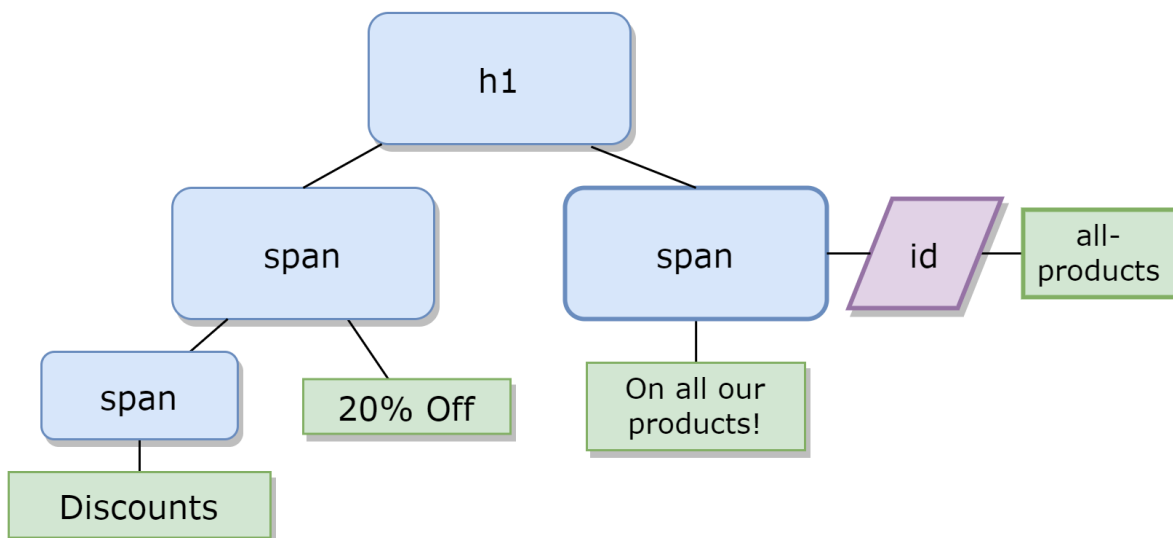
La barra simple és més precisa i imprimeix el text de de la ruta en específic, per tant retorna menys informació. Per altra banda la doble barra agafa tots els elements des d'on estiguin i si tenen els mateixos fills, els imprimeix, per tant retorna més informació perquè està imprimint informació de més elements.

- 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`

`/html/body/footer//div/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`

`//header//img/@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`

`//div[@class='img-box-inner']/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/node()`

- g. Troba la ruta que arriba fins al **<h5>** del **"New Skateboard 12"**. **[Pista:** busca la utilitat de la funció `normalize-space()`].

`<h5>`

`New Skateboard 12`

`</h5>`

`"//h5[node()][normalize-space()='12']"`

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

12

```
//div[@class='detail-box']/h5/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

```
//tr[1]/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

```
//th[4]/node()|//td[@style]/node()
```

- 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.

```
<td>Purple</td>
```

```
<td class="text-center">$55</td>
```

```
<td class="text-center">$60</td>
```

```
<td class="text-center">$72</td>
```