



Pràctica 8.2: Web Scraping (XPath)

Daniel Vallespin Mellado

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

```
git clone https://github.com/pauitic/practica8_2.git
```

Exercici 2

- 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.

- 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()`

© 2022

.

.

Les diferències són que al node retorna tot el que hi ha dins ja que el fill del div és una etiqueta `<p>` i amb `node()` obtenim tot el seu contingut de dins.

Quan utilitzem `text()` tan sol obtenim el text que hi ha directament en la primera etiqueta que hi ha text en aquest cas `<p>` té directament "2022 . .",

```
<div class="attribution">
  <p> == $0
  "© 2022 "
  <span>Todos los derechos reservados</span>
  ". "
  <a href="https://html.design/" target="_blank" rel="noopener noreferrer">Creado con
  Free Html Templates</a>
  ". "
</p>
```

ii. Barra simple vs barra doble

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

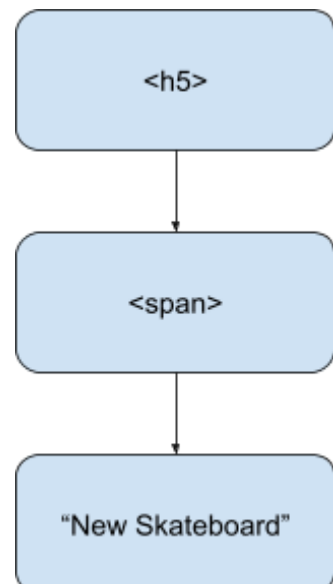
En aquest cas quan utilitza `"/` això retornarà tot els `text()` de les etiquetes `` que el seu pare directe segueix `<ul class='navbar-nav'>`.

Amb `"/` agafa totes les etiquetes `` dins de `<ul class='navbar-nav'>` pero a qualsevol nivell.

- 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).

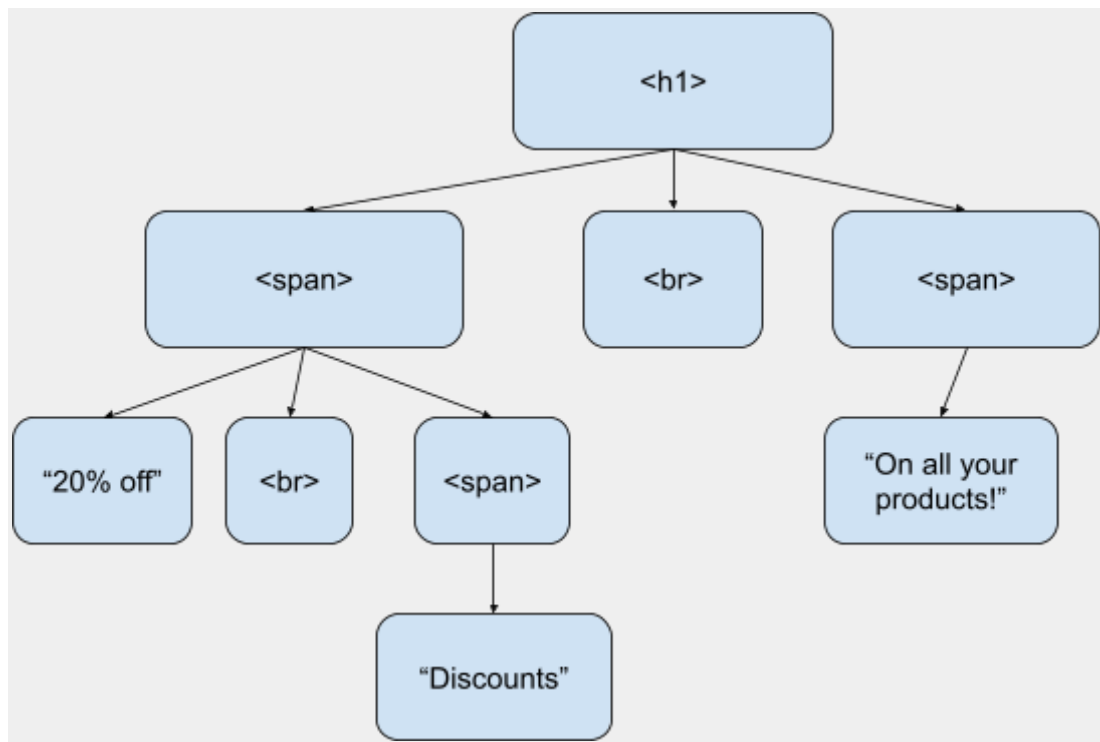
- i. `(//div/h5) [6]`

```
<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/body/footer//div[@class="information-f"]/p[last()]/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):



Header:

```
//nav/a/img/@src
```

images/logo.svg

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

```
//img[@alt="Customer"]/@src
```

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()
```

```
//div[@class="information-f"]/p[1]/strong/../span/text()
```

Fake Street 123

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

```
//section[3]//div[@class="row"]/div[last()]/h5
```

```
<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 Skateboard 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
\$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.

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

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

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

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