Pràctica 8.1: XPath

Objectius

Els objectius d'aquesta pràctica d'XPath se centren a desenvolupar una comprensió fonamental de la sintaxi i les funcionalitats bàsiques d'XPath. En primer lloc, caldrà familiaritzar-se amb la sintaxi bàsica de l'XPath i adquirir habilitats per seleccionar elements específics d'un document XML mitjançant camins absoluts i relatius. Un segon objectiu important és la pràctica en la manipulació de text i atributs, que inclourà l'extracció de contingut de text d'elements i l'accés als valors dels atributs.

Lliuraments

Aquesta pràctica està formada per la part 8.1 i la part 8.2. El lliurament es farà en una sola entrega una vegada ambdues parts estiguin realitzades. Per aquesta raó, de moment no cal lliurar aquesta pràctica.

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

Exercici 1

Completa la taula explicant els resultats esperats per a cada una de les expressions XPath donades. En casos on la resposta sigui **un objecte sigui substancialment gran**, només cal que indiquis de quin element/s es tracta.

- Si dona **error** o no dona **cap resultat**, **explica'n la raó**

Exercici 2

Una vegada tinguis la taula de l'exercici 1 plena, avalua les expressions XPath fent servir el codi Python 3 que es troba a l'arxiu *xpath_evaluator.py*.

- Assegura't tenir instal·lada la llibreria 'lxml'. Per instal·lar-la, pots executar al terminal: `pip install lxml`

	Ruta XPath	Explica el resultat
1	**/llenguatges/llenguatge/nom**	Llista de tots els noms dels llenguatges de programació
2	**/llenguatges/llenguatge/nom/node()**	Llista de nodes de text que contenen els noms dels llenguatges de programació
3	**/llenguatges/llenguatge/paradigmes[@tipat]**	Llista de tots els elements "paradigmes" amb atribut "tipat" definit
4	**/llenguatges/llenguatge/paradigmes[@tipat=”false”]**	Llista de tots els elements "paradigmes" amb atribut "tipat" amb valor "false"
5	**nom**	No hi ha coincidències, ja que "nom" no és un element dins de l'arrel "llenguatges"
6	**/nom**	No hi ha coincidències, ja que "nom" no és un element fill directe de l'arrel
7	**/nom/**	No hi ha coincidències, ja que "nom/" no és una expressió XPath vàlida
8	**//nom**	Llista de tots els nodes de text que contenen la paraula"nom" en qualsevol lloc del document
9	**//nom/node()**	Llista de tots els nodes fills dels elements "nom"
10	**//llenguatge/nom**	Llista de tots els noms dels llenguatges de programació
11	**//llenguatge/nom/node()**	Llista de tots els nodes de text que contenen els noms dels llenguatges de programació
12	**//nom/node() | //popularitat/node()**	Llista de tots els nodes fills dels elements "nom" i de tots els nodes de text que contenen la popularitat dels llenguatges de programació
13	**//mode_execucio/***	Llista de tots els nodes fills dels elements

		"mode_execucio"
14	**//llenguatge[2]**	Segon element "llenguatge"
15	**//llenguatge[last()-1]**	Penúltim element "llenguatge"
16	**//llenguatge[@fundacio]**	Llista de tots els elements "llenguatge" amb atribut "fundacio" definit
17	**//@fundacio**	Llista de tots els atributs "fundacio"
18	**//@fundacio[.>2000]**	Llista de tots els atributs "fundacio" amb valor superior a 2000
19	**//@fundacio[.>2000]/nom**	Llista dels noms dels llenguatges amb atribut "fundacio" superior a 2000
20	**//@fundacio[.>2000]/../nom**	Llista dels noms dels llenguatges amb atribut "fundacio" superior a 2000
21	**//llenguatge[mode_execucio='Java Virtual Machine'] /nom/text()**	Llista dels noms dels llenguatges que s'executen a la Java Virtual Machine
22	**//llenguatge[nom='Kotlin']/popularitat/node()**	Node de text que conté la popularitat del llenguatge Kotlin
23	**//mode_execucio[.='Interpretat']/..**	Llista dels elements pares dels elements "mode_execucio" amb valor "Interpretat"
24	**//paradigmes[node()='Imperatiu']/../nom**	Llista dels noms dels llenguatges que tenen el paradigma "Imperatiu"
25	**//***	Llista de tots els nodes del document

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<llenguatges>
   <llenguatge fundacio="2011">
       <nom>Kotlin</nom>
       <mode_execucio>Java Virtual Machine</mode_execucio>
       <popularitat>Creixent</popularitat>
       <paradigmes tipat="true">
           <paradigma>Orientat a objectes</paradigma>
           <paradigma>Imperatiu</paradigma>
           <paradigma>Funcional</paradigma>
       </paradigmes>
   </llenguatge>
   <llenguatge fundacio="1983">
       <nom>C++</nom>
       <mode_execucio>Compilació</mode_execucio>
       <popularitat>Alta</popularitat>
       <paradigmes tipat="true">
           <paradigma>Procedimental</paradigma>
           <paradigma>Imperatiu</paradigma>
           <paradigma>Orientat a objectes</paradigma>
           <paradigma>Programació genèrica</paradigma>
       </paradigmes>
   </llenguatge>
   <llenguatge fundacio="1995">
       <nom>Java</nom>
       <mode_execucio>Java Virtual Machine</mode_execucio>
       <popularitat>Alta</popularitat>
       <paradigmes tipat="true">
           <paradigma>Orientat a objectes</paradigma>
           <paradigma>Basat en classes</paradigma>
       </paradigmes>
   </llenguatge>
   <llenguatge fundacio="1991">
       <nom>Python</nom>
       <mode_execucio>Interpretat</mode_execucio>
       <popularitat>Molt alta</popularitat>
       <paradigmes tipat="false">
           <paradigma>Imperatiu</paradigma>
           <paradigma>Funcional</paradigma>
       </paradigmes>
   </llenguatge>
   <llenguatge fundacio="1995">
       <nom>JavaScript</nom>
       <mode_execucio>Interpretat</mode_execucio>
       <popularitat>Alta</popularitat>
       <paradigmes tipat="false">
           <paradigma>Orientat a objectes</paradigma>
           <paradigma>Funcional</paradigma>
       </paradigmes>
   </llenguatge>
   <nom>Això no és un llenguatge de programació :(</nom>
```

```
</llenguatges>
```

Exercici 3

El fitxer ***ods.xml*** conté els **Objectius de Desenvolupament Sostenible (ODS)** en català. Els ODS són un conjunt de 17 objectius interconnectats adoptats per les Nacions Unides per abordar els reptes mundials, com la pobresa, la desigualtat, el canvi climàtic i la justícia social, amb l'objectiu de millorar la vida de les persones i protegir el planeta.

https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Dissenyeu l'expressió XPath per accedir a la informació sol·licitada al fitxer *ods.xml*. Recordeu que heu de satisfer les condicions de l'enunciat i, tot i conèixer l'XML, heu d'extreure la informació sense afegir cap informació addicional. Per exemple, si se us demana seleccionar l'ODS titulat "Igualtat de Gènere", **heu de seleccionar-lo exclusivament fent servir aquesta informació**. L'exercici seria incorrecte si seleccioneu la informació fent servir el fet que és l'ods=5.

1. El text de dins de l'etiqueta **<titol>** de **tots els ODS** (Objectius de Desenvolupament Sostenible).

```
Fi de la Pobresa
Fam Zero
Salut i Benestar
Educació de Qualitat
Igualtat de Gènere
Aigua Neta i Sanejament
Reducció de les Desigualtats
Ciutats i Comunitats Sostenibles
Pau, Justícia i Institucions Sòlides
Energia Assequible i No Contaminant
Indústria, Innovació i Infraestructura
Producció i Consum Responsables
Acció pel Clima
Vida Submarina
Vida d'Ecosistemes Terrestres
Treball Digne i Creixement Econòmic
Aliances per a Assolir els Objectius
```

```
//titol/node()
```

2. La descripció (text) de l'ODS titulat **"Igualtat de Gènere"**.

```
Promou la igualtat de gènere i empoderar totes les dones i nenes. Inclou objectius com eliminar
la violència de gènere i garantir la participació igualitària en la presa de decisions.
```

```
//objectiu[@ods=5]/descripcio/text()
```

3. Les **accions** (text) de l'ODS que té el títol de **"Fi de la Pobresa".**

```
Implementar polítiques socials inclusives
Garantir protecció social per a tots
```

```
//objectiu[@ods=1]/accions/accio/text()
```

4. El **títol** (text) de les ODS **13, 14, 15 i 16**.

```
Pau, Justícia i Institucions Sòlides
Acció pel Clima
Vida Submarina
Vida d'Ecosistemes Terrestres
```

```
//objectiu[@ods>=13 and @ods<17]/titol/text()
```

5. El **títol** i la **descripció** (text) de l'ODS amb **ods="10".**

```
Reducció de les Desigualtats
Busca reduir les bretxes entre països i dins d'ells. Inclou objectius com empoderar les persones
marginades i promoure polítiques inclusives.
```

```
//objectiu[@ods=10]/titol/text() |
//objectiu[@ods=10]/descripcio/text()
```

6. El llistat **d'accions**, en XML, dels ODS que pertanyen al grup de **tipus econòmic**.

```
<accio>Fomentar l'emprenedoria i la innovació</accio>
<accio>Reduir la bretxa salarial de gènere</accio>
<accio>Promoure la cooperació internacional en el desenvolupament</accio>
<accio>Facilitar l'accés a la tecnologia i la innovació</accio>
```

```
//grup[@tipus="economic"]/objectiu/accions/accio
```

7. El text de dins de l'etiqueta **<titol>** del **5é objectiu** dins del grup d'ODS de tipus **"ambiental"**.

```
Vida Submarina
```

```
//grup[@tipus="ambiental"]/objectiu[@ods=14]/titol/node()
```

8. Dins el grup de tipus **social**, selecciona el **7è objectiu** i retorna la **2a acció**.

```
<accio>Garantir igualtat d'oportunitats per a tothom</accio>
```

```
//grup[@tipus="social"]/objectiu[7]/accions/accio[2
```

9. El **tipus del grup** al qual pertany l'ODS que té l'acció "**Promoure l ús d energies renovables**".

```
ambiental
```

```
//grup[2]/@tipus
```

10. **Busca les etiquetes amagades <start> i <end>** (fes ctrl+F) i **troba el camí** per començar a l'etiqueta <start> i arribar a imprimir el text que hi ha dins l'etiqueta <end>. La teva ruta ha de començar per ***//*start** i has d'intentar fer l'expressió XPath més breu possible.

```
<end>Busca promoure l'ocupació digna, el creixement econòmic inclusiu i la protecció social.
Inclou objectius com reduir la informalitat laboral i garantir salaris justos.</end>
```

```
//start/ancestor::*//end
```

11. Les **5 primeres accion**s per les ODS de tipus **ambiental.** [opcional]

```
<accio>Desenvolupar infraestructures per a energies renovables</accio>
<accio>Promoure l'eficiència energètica a la indústria</accio>
<accio>Invertir en investigació i desenvolupament</accio>
<accio>Desenvolupar infraestructures de transport sostenible</accio>
<accio>Fomentar el reciclatge i la reutilització de productes</accio>
```

```
//grup[@tipus="ambiental"]/objectiu[@ods=7]/accions/accio |
//grup[@tipus="ambiental"]/objectiu[@ods=9]/accions/accio |
//grup[@tipus="ambiental"]/objectiu[@ods=12]/accions/accio[1]
```

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.

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()`

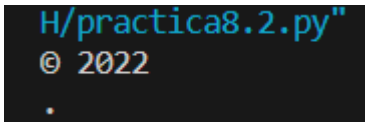
```
actica8.2.py"
© 2022
<span>All Rights Reserved</span>.

.

<a href="https://html.design/" target="_blank" rel="noopener noreferrer">Created with Free Htm
l Templates</a>.
```

- `Footer: XPATH node(), imprime todo lo que esta dentro de la etiqueta <p>.`

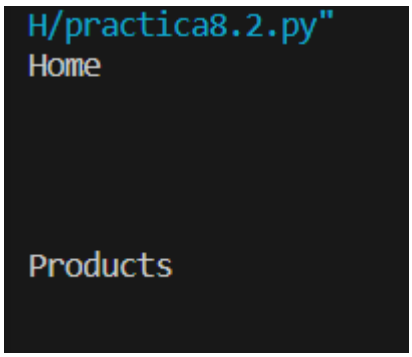
Ruta 2: `//div[@class='attribution']/p/text()`



- `Footer: XPATH text(), imprime solo el texto que contiene la etiqueta, no admite referencias o atributos.`

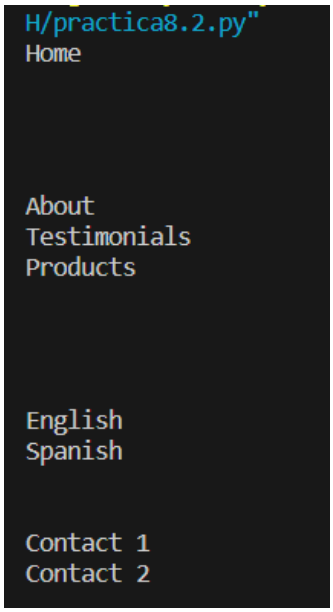
ii. Barra simple vs barra doble

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



- `Con barra simple "\" solo imprime las <li> padres.`

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

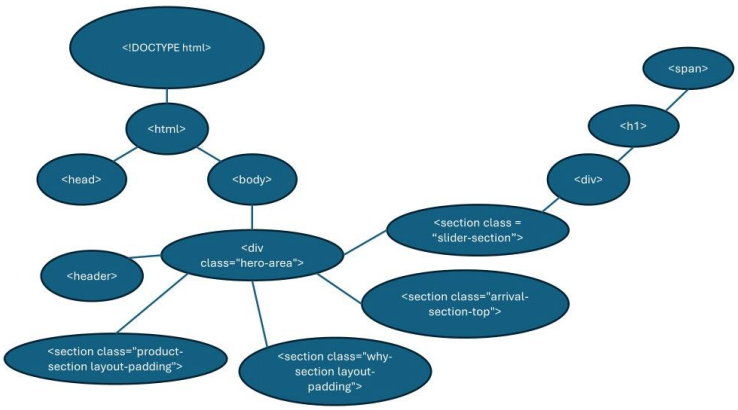


- `Con doble barra "\\" nos indica que todas la <li> imprimirá tanto las etiquetas padre como las que estan dentro de este.`

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

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[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//div[@class='logo-footer']/a/img/@src
```

```
images/logo.svg
```

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

```
//section[@class='client-section layout-padding']//div[@class = 'img-box-inner']/img/@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]/span/text()
```

```
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>
```

```
(((//section[@class = 'product-section layout-padding']//div[@class = 'detail-box']/h5)[12])
```

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

```
12
```

```
normalize-space(((//section[@class = 'product-section layout-padding']//div[@class = 'detail-box']/h5/text())[24])
```

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
```

```
//tbody/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
```

```
//thead/tr/th[4]/text() | //tbody/tr/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 "|"]

```
//td[text()='$110']
```

```
Skate
Special
```

```
//td[text()='$110']/ancestor::tr/td[1]/text() |
//td[text()='$110']/ancestor::table/thead/tr/th[2]/text()
```

l. 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>
```

```
//tbody/tr[4]/td[1] | //tbody/tr[4]/td[2] | //tbody/tr[4]/td[3] |
                    //tbody/tr[4]/td[5]
```