

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	<nom>Kotlin</nom> <nom>C++</nom> <nom>Java</nom> <nom>Python</nom> <nom>JavaScript</nom>
2	/llenguatges/llenguatge/nom/node()	Kotlin C++ Java Python JavaScript
3	/llenguatges/llenguatge/paradigmes[@tipat]	Retornarà totes les etiquetes paradigmes amb atribut tipat = true Retorna tots els tipats es igual si true o false
4	/llenguatges/llenguatge/paradigmes[@tipat="false"]	Retornarà totes les etiquetes paradigmes amb atribut tipat = false
5	nom	Retornarà la primera etiqueta nom que troba
6	/nom	Retornarà la primera etiqueta nom que estiguin des de l'arrel []
7	/nom/	Retorna l'estructura d'etiquetes de l'etiqueta nom Retorna un error
8	//nom	retorna totes les etiquetes nom del document
9	//nom/node()	Retornara tots els valors que tinguin l'etiqueta nom
10	//llenguatge/nom	Retorna l'etiqueta nom de tots els llenguatges
11	//llenguatge/nom/node()	Retorna el valor del nom de tots els llenguatges
12	//nom/node() //popularitat/node()	Retorna el valor de tots els noms i el valor de la seva popularitat
13	//mode_execucio/*	Retorna tots els elements per sota de tots els mode_execucio Donara error, no retorna res. No hi ha res a sota de l'etiqueta

14	//llenguatge[2]	Retorna el segon llenguatge del document
15	//llenguatge[last()-1]	Retorna el penultim llenguatge
16	//llenguatge[@fundacio]	Retorna el llenguatge que tingui atribut fundacio
17	//@fundacio	Retorna els elements que tinguin atribut fundacio Retorna el valor dels atributs fundacio
18	//@fundacio[.>2000]	Retorna els elements que tinguin atribut fundacio i el valor d'aquest sigui major a 2000
19	//@fundacio[.>2000]/nom	Retorna l'etiqueta nom dels elements que tinguin atribut fundacio i el valor d'aquest sigui major a 2000 Donaria error, no retorna res.
20	//@fundacio[.>2000]/../nom	Retorna l'etiqueta nom del element anterior als elements que tinguin atribut fundacio i el valor d'aquest sigui major a 2000 Donaria error, no retorna res.
21	//llenguatge[mode_execucio='Java Virtual Machine']/nom/text()	Retorna el text de la etiqueta nom del element llenguatge que tingui atribut mode_execucio = Java Virtual Machine Resultat: Kotlin Java
22	//llenguatge[nom='Kotlin']/popularitat/node()	Retorna el valor de popularitat dels llenguatge que tinguin com a nom "Kotlin" Resultat: Creixent
23	//mode_execucio[.='Interpretat']/..	Retornara l'element pare de l'etiqueta que sigui Interpretat
24	//paradigmes[node()='Imperatiu']/../nom	Dona error, paradigmes conte etiquetes, no un valor.
25	//*	Retorna tot el document

```
<?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

//objectiu/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[titol="Igualtat de Gènere"]/descripcio/node()

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[titol="Fi de la Pobresa"]/accions//accio/node()

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 or @ods=14 or @ods=15 or @ods=16]/titol/node()

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/node() | //objectiu[@ods=10]/descripcio/node()

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

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

Vida Submarina

//grup[@tipus="ambiental"]/objectiu[5]/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

//accio[text()="Promoure l'ús d'energies renovables"]/ancestor::grup/@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::ods//end

11. Les **5 primeres accions** 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/accions/accio)[position() <= 5]

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()`

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

node ens retorna el contingut del paràgraf, en canvi text ens retorna el text, en aquest cas no hi ha text directe

- ii. Barra simple vs barra doble

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

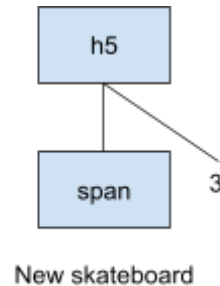
Retorna dels ul el primer li

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

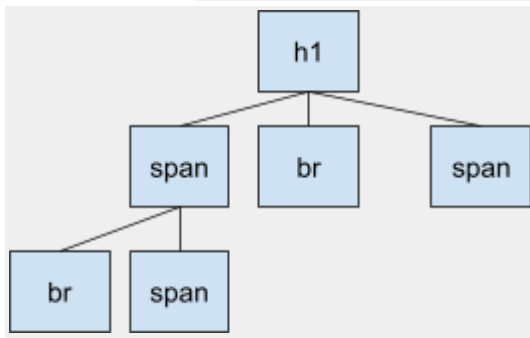
Retorna de tots els ul tots els li

- 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//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):



`//header//a[@class='navbar-brand']//@src`

images/logo.svg

- 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

```
//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()
```

Solució:

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

```
<h5>                                <span>New Skateboard</span> 12
</h5>
```

```
//h5/text()[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"**.

```
//h5/text()[normalize-space()='12']/../../h6/text()
```

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

```
//tr//td[text()='Blue']/../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[@style='color: red;']/text() | //td[@style='color: red;']/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 ']/preceding::td[text() = 'Special']/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>  
//tr[td[text() = 'Purple']]/td
```

