Ejercicios sobre Procesamiento de información en Internet (Web Scraping)

Nombres de bebés: de Internet a un DataFrame

En la siguiente url,

```
https://www.enterat.com/servicios/nombres-nino-nina.php
```

hemos encontrado una tabla con los nombres de bebés más populares en 2019, en el sentido de ser los más elegidos por los padres para bautizar a sus hijos:

Los 100 nombres de niño y niña más puestos en España (sean o no en español)

Posición	Nombres de niño	Nombres de niña
1	HUGO	LUCIA
2	LUCAS	SOFIA
3	MARTIN	MARTINA
4	DANIEL	MARIA
5	PABLO	PAULA
6	MATEO	JULIA
7	ALEJANDRO	EMMA

Se pide acceder a esa página desde Python, localizar la tabla y convertirla en un DataFrame para poder procesarla posteriormente.

Apartado 1

En primer lugar, debes hacer lo siguiente mediante un programa en Python: (a) acceder a la `url`, (b) cargar el texto de la página, (c) buscar la tabla, (d) cargar las filas (etiqueta `tr`) y (e) las celdas (etiqueta `td`).

Comprobamos el resultado de estos pasos imprimiendo alguna fila y alguna celda.

In [1]:

```
import requests
bebes_url = "https://www.enterat.com/servicios/nombres-nino-nina.php"
bebes_texto = requests.get(bebes_url).text

from bs4 import BeautifulSoup
bebes_datos = BeautifulSoup(bebes_texto, "html")

bebes_filas = bebes_datos.findAll("tr")
print(bebes_filas[1])

print("..............................")

bebes_celdas = bebes_datos.findAll("td")
print(bebes_celdas[3:6])
```

```
<tr>
<td align="center" width="10%">1</td>
<td align="center" valign="middle" width="117">HUGO</td>
<td align="center" valign="middle" width="117">LUCIA</td>
</tr>
..............................
[<td align="center" width="10%">1</td>, <td align="center" valign="middle" w
idth="117">HUGO</td>, <td align="center" valign="middle" width="117">LUCIA</
td>]
```

Apartado 2

Ahora iniciamos en procesamiento de los datos que tenemos, y vamos comprobando el resultado paso a paso. (a) Selecciona las celdas de una fila cualquiera, y (b) extrae el contenido (el texto) de las celdas de esa fila. (c) El dato entero que indica el número de orden debería ser un entero, y no una cadena de caracteres...

In [2]:

```
fila_1 = bebes_celdas[3:6]
info_1 = [elemento.get_text() for elemento in fila_1]
print(fila_1)
print(info_1)
info_1[0] = int(info_1[0])
print(info_1)
```

```
[<td align="center" width="10%">1</td>, <td align="center" valign="middle" w
idth="117">HUGO</td>, <td align="center" valign="middle" width="117">LUCIA</
td>]
['1', 'HUGO', 'LUCIA']
[1, 'HUGO', 'LUCIA']
```

Apartado 3

Es frecuente necesitar cada columna en una lista (porque los datos de cada columna son homogéneos, y así se facilita su procesamiento en algunos programas. Prepara los datos en tres listas, una por cada columna, y almacena el resultado en sendas listas.

In [3]:

```
num_or = [int(elem.findAll("td")[0].get_text()) for elem in bebes_filas[1:]]
chicos = [    elem.findAll("td")[1].get_text()  for elem in bebes_filas[1:]]
chicas = [    elem.findAll("td")[2].get_text()  for elem in bebes_filas[1:]]

# Los 10 primeros de cada:
print(num_or[:10])
print(chicos[:10])
print(chicas[:10])
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
['HUGO', 'LUCAS', 'MARTIN', 'DANIEL', 'PABLO', 'MATEO', 'ALEJANDRO', 'LEO',
'ALVARO', 'MANUEL']
['LUCIA', 'SOFIA', 'MARTINA', 'MARIA', 'PAULA', 'JULIA', 'EMMA', 'VALERIA',
'DANIELA', 'ALBA']
```

Apartado 4

Pasa el contenido de estas columnas a un DataFrame de la librería `pandas` .

In [4]:

```
from pandas import DataFrame

tabla = DataFrame([num_or, chicos, chicas]).T
tabla.columns = ["Núm. orden", "Chicos", "Chicas"]
print(tabla.head())
```

```
  Núm. orden  Chicos   Chicas
0          1    HUGO    LUCIA
1          2   LUCAS    SOFIA
2          3  MARTIN  MARTINA
3          4  DANIEL    MARIA
4          5   PABLO    PAULA
```