Aprèn a realitzar web scraping.

Nivell 1

- Exercici 1:

Realitza web scraping d'una pàgina de la borsa de Madrid (https://www.bolsamadrid.es) utilitzant BeautifulSoup i Selenium.

In [213…
```
import requests
from bs4 import BeautifulSoup
import re
import pandas as pd
pd.set_option('display.max_columns', None)
```

In [214…
```
# asignar url
url = ('https://www.bolsamadrid.es')
page = requests.get(url)
```

In [215…
```
soup = BeautifulSoup(page.content, 'html.parser')
```

In [216…
```
#obtención de la ruta al listado completo de acciones
results = soup.find('a', string='Acciones')
results = results.get('href')
print(results)
```

```
/esp/aspx/Mercados/Precios.aspx?indice=ESI100000000
```

In [217…
```
#link a la lista de las acciones
link = (url+results)
print(link)
```

```
https://www.bolsamadrid.es/esp/aspx/Mercados/Precios.aspx?indice=ESI100000000
```

In [218…
```
html = requests.get(link)
soup = BeautifulSoup(html_acciones.content, 'html.parser')
```

In [219…
```
#obtenemos los nombres de las columnas
column_names = soup.find(id='ctl00_Contenido_tblAcciones')
column_names.find_all('th')
```

Out[219…
```
[<th scope="col">Nombre</th>,
 <th scope="col">Últ.</th>,
 <th scope="col">% Dif.</th>,
 <th scope="col">Máx.</th>,
 <th scope="col">Mín.</th>,
 <th scope="col">Volumen</th>,
 <th scope="col">Efectivo (miles €)</th>,
 <th scope="col">Fecha</th>,
 <th class="Ult" scope="col">Hora</th>]
```

In [220…
```
#se crea un array para almacenar los nombres de las columnas
array = []
for element in column_names.find_all("th"):
    array.append(element.get_text())
print(array)
```

```
['Nombre', 'Últ.', '% Dif.', 'Máx.', 'Mín.', 'Volumen', 'Efectivo (miles €)', 'Fecha', 'Hora']
```

In [221…
```
#obtención de los valores de los elementos que llenarán el array
stocks = []
for element in column_names.find_all("td"):
    stocks.append(element.text)
print (stocks)
```

```
['ACCIONA', '180,1000', '0,61', '181,1000', '178,6000', '12.102', '2.178,40', '04/07/2022', '11:06:30', 'ACCION
A ENER', '38,0600', '1,28', '38,3000', '37,8400', '29.402', '1.119,37', '04/07/2022', '11:06:17', 'ACERINOX',
'8,6020', '-1,38', '8,8600', '8,5840', '346.897', '3.020,63', '04/07/2022', '11:06:26', 'ACS', '21,7300', '0,6
5', '21,8400', '21,5800', '163.729', '3.552,45', '04/07/2022', '11:06:33', 'AENA', '127,1500', '2,66', '127,150
0', '124,5500', '21.286', '2.684,12', '04/07/2022', '11:06:30', 'AMADEUS', '53,0600', '0,19', '53,7400', '53,02
00', '49.611', '2.644,70', '04/07/2022', '11:06:46', 'ARCELORMIT', '21,7000', '0,51', '22,0000', '21,5750', '5
7.286', '1.247,10', '04/07/2022', '11:06:25', 'B.SANTANDER', '2,6710', '-0,48', '2,6980', '2,6615', '2.867.08
7', '7.681,44', '04/07/2022', '11:06:46', 'BA.SABADELL', '0,7378', '-2,97', '0,7640', '0,7312', '12.846.357',
'9.577,56', '04/07/2022', '11:06:29', 'BANKINTER', '5,7640', '-2,64', '5,9480', '5,7020', '637.686', '3.704,9
7', '04/07/2022', '11:06:08', 'BBVA', '4,3385', '0,58', '4,3680', '4,3170', '2.606.043', '11.312,47', '04/07/20
22', '11:06:15', 'CAIXABANK', '3,2530', '-1,81', '3,3220', '3,2200', '3.466.757', '11.304,29', '04/07/2022', '1
1:05:51', 'CELLNEX', '38,1600', '0,00', '38,4800', '37,8900', '110.997', '4.237,92', '04/07/2022', '11:04:40',
'ENAGAS', '21,5800', '1,46', '21,6000', '21,3600', '181.390', '3.897,30', '04/07/2022', '11:07:00', 'ENDESA',
'18,6100', '0,32', '18,7350', '18,4850', '1.136.910', '21.162,53', '04/07/2022', '11:06:35', 'FERROVIAL', '24,7
700', '2,27', '24,7900', '24,4100', '79.456', '1.956,18', '04/07/2022', '11:06:33', 'FLUIDRA', '18,6200', '-3,6
2', '19,5100', '18,4700', '206.584', '3.886,29', '04/07/2022', '11:06:15', 'GRIFOLS CL.A', '18,5800', '-0,30',
'18,6350', '18,0950', '171.955', '3.159,57', '04/07/2022', '11:07:11', 'IAG', '1,2450', '-1,31', '1,2865', '1,2
420', '3.621.780', '4.570,90', '04/07/2022', '11:07:11', 'IBERDROLA', '10,3800', '0,39', '10,4350', '10,3350',
'799.015', '8.296,61', '04/07/2022', '11:06:34', 'INDITEX', '21,9900', '0,09', '22,1700', '21,9300', '202.961',
'4.474,03', '04/07/2022', '11:05:02', 'INDRA A', '8,9400', '0,11', '9,0400', '8,8900', '164.998', '1.477,39',
'04/07/2022', '11:06:30', 'INM.COLONIAL', '5,9950', '-1,56', '6,1600', '5,9950', '328.890', '1.997,64', '04/07/
2022', '11:05:21', 'MAPFRE', '1,7100', '0,12', '1,7130', '1,6960', '373.202', '635,85', '04/07/2022', '11:06:2
3', 'MELIA HOTELS', '5,9750', '0,17', '6,1150', '5,9300', '239.637', '1.440,86', '04/07/2022', '11:01:24', 'MER
LIN', '9,1500', '-1,13', '9,3200', '9,1200', '81.694', '756,38', '04/07/2022', '11:06:26', 'NATURGY', '28,150
0', '1,30', '28,1900', '27,7700', '32.858', '920,69', '04/07/2022', '11:07:04', 'PHARMA MAR', '68,0800', '-0,6
1', '68,9200', '66,8400', '4.785', '324,90', '04/07/2022', '11:07:03', 'R.E.C.', '18,4300', '-0,14', '18,5600',
'18,4000', '2.128.244', '39.308,86', '04/07/2022', '11:06:27', 'REPSOL', '13,8250', '2,56', '13,9500', '13,730
0', '796.656', '11.022,88', '04/07/2022', '11:06:43', 'ROVI', '58,3500', '1,21', '58,9000', '57,3000', '20.63
8', '1.202,18', '04/07/2022', '11:05:18', 'SACYR', '2,3240', '-0,43', '2,3640', '2,3240', '222.597', '520,05',
'04/07/2022', '10:58:315', 'SIEMENS GAME', '17,8650', '0,14', '17,8700', '17,8350', '614.433', '10.966,32', '04/
07/2022', '11:06:39', 'SOLARIA', '20,9300', '1,21', '21,0800', '20,6500', '104.216', '2.178,63', '04/07/2022',
'11:06:15', 'TELEFONICA', '4,9390', '1,35', '4,9570', '4,8550', '2.951.814', '14.550,23', '04/07/2022', '11:07:
12']
```

In [222…
```
# se añaden los valores al array mediante un diccionario que asigna los valores
# a la clave correspondiente
fill_stocks ={ }
for i in range(9):
    fill_stocks[array[i]]= stocks [i::9]
```

In [223…
```
# creación del DF con el listado iterable del IBEX 35
df = pd.DataFrame(fill_stocks)
print(df)
```

```
          Nombre      Últ. % Dif.      Máx.      Mín.      Volumen  \
0        ACCIONA  180,1000   0,61  181,1000  178,6000       12.102
1   ACCIONA ENER   38,0600   1,28   38,3000   37,8400       29.402
2      ACERINOX    8,6020  -1,38    8,8600    8,5840      346.897
3           ACS   21,7300   0,65   21,8400   21,5800      163.729
4        AENA   127,1500   2,66  127,1500  124,5500       21.286
5      AMADEUS   53,0600   0,19   53,7400   53,0200       49.611
6    ARCELORMIT.   21,7000   0,51   22,0000   21,5750       57.286
7    B.SANTANDER    2,6710  -0,48    2,6980    2,6615  2.867.087
8    BA.SABADELL    0,7378  -2,97    0,7640    0,7312  12.846.357
9      BANKINTER    5,7640  -2,64    5,9480    5,7020      637.686
10          BBVA    4,3385   0,58    4,3680    4,3170    2.606.043
11     CAIXABANK    3,2530  -1,81    3,3220    3,2200    3.466.757
12       CELLNEX   38,1600   0,00   38,4800   37,8900      110.997
13       ENAGAS   21,5800   1,46   21,6000   21,3600      181.390
14       ENDESA   18,6100   0,32   18,7350   18,4850    1.136.910
15     FERROVIAL   24,7700   2,27   24,7900   24,4100       79.456
16       FLUIDRA   18,6200  -3,62   19,5100   18,4700      206.584
17  GRIFOLS CL.A   18,5800  -0,30   18,6350   18,0950      171.955
18           IAG    1,2450  -1,31    1,2865    1,2420    3.621.780
19      IBERDROLA   10,3800   0,39   10,4350   10,3350      799.015
20       INDITEX   21,9900   0,09   22,1700   21,9300      202.961
21       INDRA A    8,9400   0,11    9,0400    8,8900      164.998
22  INM.COLONIAL    5,9950  -1,56    6,1600    5,9950      328.890
23        MAPFRE    1,7000   0,12    1,7130    1,6960      373.202
24  MELIA HOTELS    5,9750   0,17    6,1150    5,9300      239.637
25        MERLIN    9,1500  -1,13    9,3200    9,1200       81.694
26       NATURGY   28,1500   1,30   28,1900   27,7700       32.858
27   PHARMA MAR   68,0800  -0,61   68,9200   66,8400        4.785
28        R.E.C.   18,4300  -0,14   18,5600   18,4000    2.128.244
29        REPSOL   13,8250   2,56   13,9500   13,7300      796.656
30          ROVI   58,3500   1,21   58,9000   57,3000       20.638
31         SACYR    2,3240  -0,43    2,3640    2,3240      222.597
32  SIEMENS GAME   17,8650   0,14   17,8700   17,8350      614.433
33       SOLARIA   20,9300   1,21   21,0800   20,6500      104.216
34     TELEFONICA    4,9390   1,35    4,9570    4,8550    2.951.814

   Efectivo (miles €)       Fecha      Hora
0            2.178,40  04/07/2022  11:06:30
1            1.119,37  04/07/2022  11:06:17
2            3.020,63  04/07/2022  11:06:26
3            3.552,45  04/07/2022  11:06:33
4            2.684,12  04/07/2022  11:06:30
5            2.644,70  04/07/2022  11:06:46
6            1.247,10  04/07/2022  11:06:25
7            7.681,44  04/07/2022  11:06:46
8            9.577,56  04/07/2022  11:06:29
9            3.704,97  04/07/2022  11:06:08
10          11.312,47  04/07/2022  11:06:15
11          11.304,29  04/07/2022  11:05:51
12           4.237,92  04/07/2022  11:04:40
13           3.897,30  04/07/2022  11:07:00
14          21.162,53  04/07/2022  11:06:35
15           1.956,18  04/07/2022  11:06:33
16           3.886,29  04/07/2022  11:06:15
17           3.159,57  04/07/2022  11:07:11
18           4.570,90  04/07/2022  11:07:11
19           8.296,61  04/07/2022  11:06:34
20           4.474,03  04/07/2022  11:05:02
21           1.477,39  04/07/2022  11:06:30
22           1.997,64  04/07/2022  11:05:21
23             635,85  04/07/2022  11:06:23
24           1.440,86  04/07/2022  11:01:24
25             756,38  04/07/2022  11:06:26
26             920,69  04/07/2022  11:07:04
27             324,90  04/07/2022  11:07:03
28          39.308,86  04/07/2022  11:06:27
29          11.022,88  04/07/2022  11:06:43
30           1.202,18  04/07/2022  11:05:18
31             520,05  04/07/2022  10:58:31
32          10.966,32  04/07/2022  11:06:39
33           2.178,63  04/07/2022  11:06:15
34          14.550,23  04/07/2022  11:07:12
```

A continuanción creo un programa que permita extraer el valor de las acciones de una empresa específica en el momento de la búsqueda

In [224…
```
import time
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.common.by import By

driver_path = Service('/Applications/chromedriver')
driver = webdriver.Chrome(service = driver_path)
options = webdriver.ChromeOptions()
options.add_argument('--start-maximized')
driver.get('https://www.bolsamadrid.es')

# selección de la casilla de búsqueda mediante xpath
search = driver.find_element(By.XPATH, '//*[@id="formBusq"]/input')
#introducción te texto. Aquí se puede cambiar la empresa.
search.send_keys('telefonica')

#confirmación de la empresa en el listado de bisqueda
enter = driver.find_element(By.XPATH, '//*[@id="formBusq"]/span/a/span')
enter.click()
acciona = driver.find_element(By.XPATH, '//*[@id="ctl00_Contenido_tblEmisoras"]/tbody/tr[2]/td/a')
acciona.click()

#obtención de valor de acciones.
date = driver.find_element(By.XPATH, '//*[@id="ctl00_Contenido_tblPrecios"]/tbody/tr[2]/td[1]')
time = driver.find_element(By.XPATH, '//*[@id="ctl00_Contenido_tblPrecios"]/tbody/tr[2]/td[2]')
closing = driver.find_element(By.XPATH, '//*[@id="ctl00_Contenido_tblPrecios"]/tbody/tr[2]/td[6]')
print("Valor Telefonica: ", closing.text)
print("Fecha: ", date.text)
print("Hora: ", time.text)
driver.quit()

#driver.find_elements_by_xpath('//*[@id="formBusq"]/input')
```

```
Valor Telefonica:  4,9200
Fecha:  04/07/2022
Hora:  11:42:20
```

Nivell 2

- Exercici 2: Documenta en un word el teu conjunt de dades generat amb la informació que tenen els diferents arxius de Kaggle.

Nivell 3

- Exercici 3 Tria una pàgina web que tu vulguis i realitza web scraping mitjançant la llibreria Scrapy.