Práctica 1 Inteligencia Artificial
Python - Web Scraping

Mario Uceda Yeves
Ignacio Gil Garzón

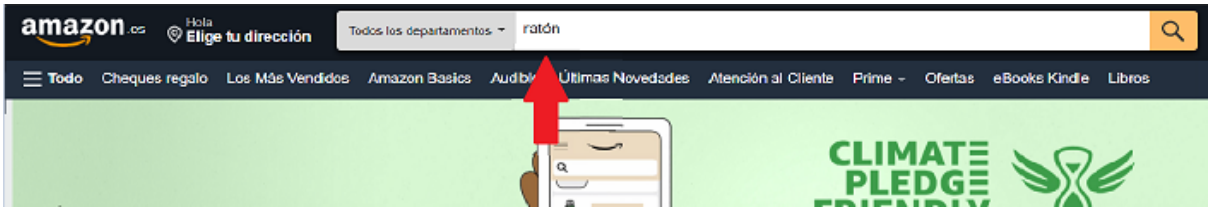
1. Selenium

Usaremos la página web de Amazon.es para realizar diferentes acciones con Selenium:

- Realizar click en algún botón:

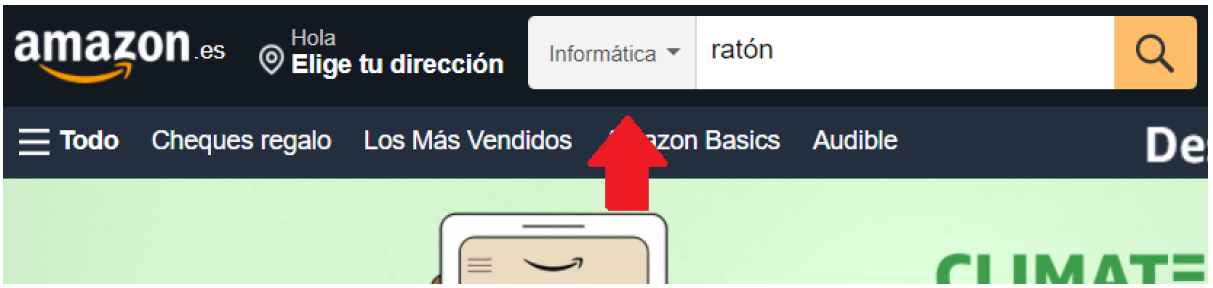

```
1. #Aceptar cookies
2. cookie=driver.find_element_by_xpath('//*[@id="sp-cc-accept"]'
   )
3. cookie.click()
```

- Escribir texto en alguna caja de texto:




```
1. #Escribir texto en un buscador
2. buscador =
   driver.find_element_by_xpath('//*[@id="twotabsearchtextbox"]')
3. texto = "ratón"
4. buscador.send_keys(texto)()
```

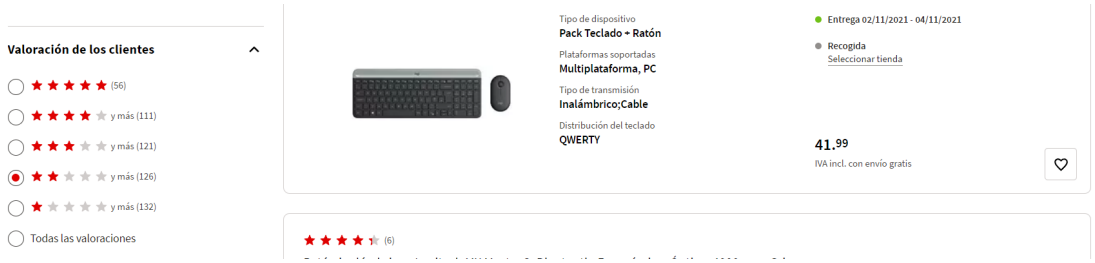
- Seleccionar opciones de menús desplegables:




```
1. #Seleccionar opción en un menú desplegable
2. desplegable =
   driver.find_element_by_xpath('//*[@id="searchDropdownBox"]')
3. elemento_selec = Select(desplegable)
4. elemento_selec.select_by_value("search-alias=computers")
```

- Seleccionar radiobutton:

No hemos encontrado radiobutton en Amazon, pero sí en la página de MediaMarkt:




```
1. #Seleccionar radiobutton
2. radio =
   driver.find_element_by_xpath('//*[@id="facet-Valoración de
   los clientes"]/div[2]/div/div/div/label[4]')
3. radio.click()
```

Otras acciones mediante esta librería:

```
1. #volver a la página anterior
2. driver.back()
3. #ir a la página siguiente
4. driver.forward()
5. #cerrar la página
6. driver.close()
```

2. BeautifulSoup

- Parsear información:

```
1. from bs4 import BeautifulSoup
2. import requests
3.
4. URL = "https://www.marca.com/"
5.
6. req = requests.get(URL)
7.
8. # Comprobamos que la petición nos devuelve un Status Code= 200
9. status_code = req.status_code
10.if status_code == 200:
11.
12.     # Pasamos el contenido HTML de la web a un objeto BeautifulSoup()
13.     html = BeautifulSoup(req.text, "html.parser")
```

- Formatear la información deseada:

```
-       # Obtenemos todos los li donde están las noticias
-       noticias = html.find_all('li', {'class': 'flex__item'})
-       listanoticias = []
-       # Recorremos todas las noticias para extraer el título y autor
-       for i, noticia in enumerate(noticias):
-
-           titulo = noticia.find('a', {'class':
    'flex-article__heading-link'})
-           if(titulo is not None):
-               titulo = titulo.getText()
-           else:
-               titulo = "None"
-
-           autor = noticia.find('ul', {'class': 'mod-author'})
-           if(autor is not None):
-               autor = autor.find('span').getText()
-           else:
-               autor = "None"
-           numero = i+1
-
-           # Imprimo el Título y Autor de las noticias
-           print (numero, " -> ",titulo, " - ",autor)
-
-           noticiaDatos = [numero, titulo, autor]
-           listanoticias.append(noticiaDatos)
-
-   else:
-       print ("Status Code ",status_code)
```

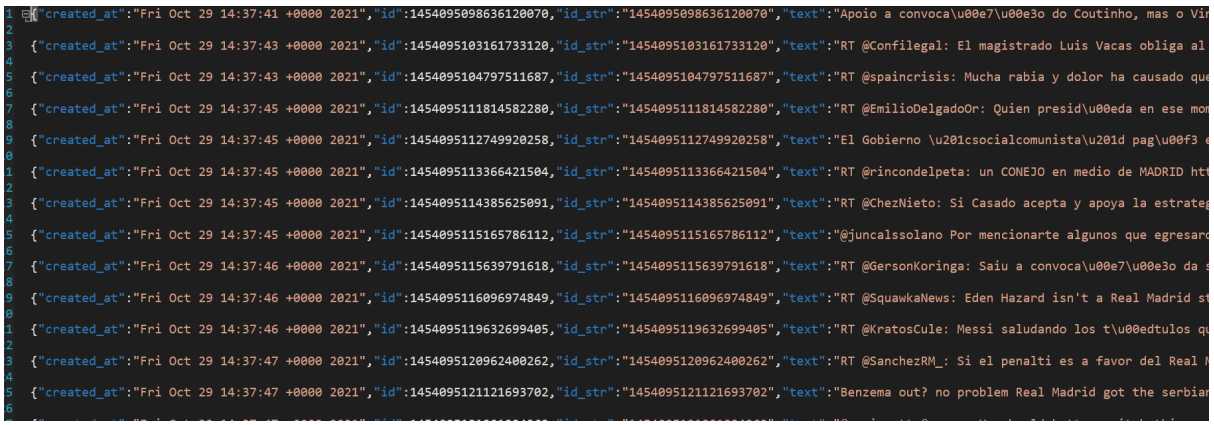
3. Guardar la información

```
1. f = open ('../Datos/noticias.txt','w')
2. for noticia in listanoticias:
3.     linea = (str(noticia[0])) + " -> " + noticia[1] + " - " +
   noticia[2]+"\n"
4.     f.write(linea)
5. f.close()
```

4. Tweepy

- Escucha y guardado de la información en Twitter:

```
1. import tweepy
2.
3. from tweepy import Stream
4. from tweepy.streaming import StreamListener
5. from tweepy import OAuthHandler
6.
7. class MyListener (StreamListener):
8.
9.     def on_data(self, data):
10.         try:
11.             with open ("C:/Users/nacho/IA/Tweets2.json", 'a') as f:
12.                 f.write(data)
13.                 return True
14.         except BaseException as e:
15.             print("Error en el dato: %s" % str(e))
16.             return True
17.     def on_error(self, status):
18.         print(status)
19.         return True
20.
21.#Credenciales del Twitter API
22.consumer_key = "S8HBEm4evnbfxLBryV44Qz1po"
23.consumer_secret = "iZUWayc5dn2IY0LtAZYh2lGyZcCcDAqYDoBkJlvlJxWLXRC8hH"
24.access_token = "1441480563169312771-kYQajbkkxxV612VQGafNsgr3OxPulv"
25.access_secret = "k87QEfgMmklo1RODhLskKTuORjQubowvDC4A1QVi18unO"
26.
27.auth=OAuthHandler(consumer_key, consumer_secret)
28.auth.set_access_token(access_token, access_secret)
29.
30.api = tweepy.API(auth)
31.
32.twitter_stream = Stream (auth, MyListener())
33.twitter_stream.filter(track=['madrid'])
```


[{"created_at":"Fri Oct 29 14:37:41 +0000 2021","id":1454095098636120070,"id_str":"1454095098636120070","text":"Apoio a convoca\u00e7\u00e3o do Coutinho, mas o Vi
{"created_at":"Fri Oct 29 14:37:43 +0000 2021","id":1454095103161733120,"id_str":"1454095103161733120","text":"RT @Confilegal: El magistrado Luis Vacas obliga al
{"created_at":"Fri Oct 29 14:37:43 +0000 2021","id":1454095104797511687,"id_str":"1454095104797511687","text":"RT @spaincrisis: Mucha rabia y dolor ha causado que
{"created_at":"Fri Oct 29 14:37:45 +0000 2021","id":1454095111814582280,"id_str":"1454095111814582280","text":"RT @EmilioDelgadoOr: Quien presid\u00eda en ese mon
{"created_at":"Fri Oct 29 14:37:45 +0000 2021","id":1454095112749920258,"id_str":"1454095112749920258","text":"El Gobierno \u201csocialcomunista\u201d pag\u00f3 e
{"created_at":"Fri Oct 29 14:37:45 +0000 2021","id":1454095113366421504,"id_str":"1454095113366421504","text":"RT @rincondelpeta: un CONEJO en medio de MADRID htt
{"created_at":"Fri Oct 29 14:37:45 +0000 2021","id":1454095114385625091,"id_str":"1454095114385625091","text":"RT @ChezNieto: Si Casado acepta y apoya la estrateg
{"created_at":"Fri Oct 29 14:37:45 +0000 2021","id":1454095115165786112,"id_str":"1454095115165786112","text":"@juncalssolano Por mencionarte algunos que egresaro
{"created_at":"Fri Oct 29 14:37:46 +0000 2021","id":1454095115639791618,"id_str":"1454095115639791618","text":"RT @GersonKoringa: Saiu a convoca\u00e7\u00e3o da s
{"created_at":"Fri Oct 29 14:37:46 +0000 2021","id":1454095116096974849,"id_str":"1454095116096974849","text":"RT @SquawkaNews: Eden Hazard isn't a Real Madrid st
{"created_at":"Fri Oct 29 14:37:46 +0000 2021","id":1454095119632699405,"id_str":"1454095119632699405","text":"RT @KratosCule: Messi saludando los t\u00edtulos qu
{"created_at":"Fri Oct 29 14:37:47 +0000 2021","id":1454095120962400262,"id_str":"1454095120962400262","text":"RT @SanchezRM_: Si el penalti es a favor del Real M
{"created_at":"Fri Oct 29 14:37:47 +0000 2021","id":1454095121121693702,"id_str":"1454095121121693702","text":"Benzema out? no problem Real Madrid got the serbian

5. Pandas

- Realizar un cuadro de mando:

Cuadro de Mandos utilizando Pandas y madrid.json obtenido con Tweepy

1. Importar librerías y json

[156]:
```
import pandas as pd
import json
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

def parseMultipleJSON3(lines):
    lines = ''.join(lines).split('}{')
    data = [json.loads('%s}' % line)
        if idx == 0 else json.loads('{%s' % line)
        if idx == len(lines)-1
        else json.loads('{%s}' % line)
        for idx, line in enumerate(lines)]
    return data

with open('madrid.json','r') as json_file:
    lines = json_file.readlines()
    lines = [line.strip("\n") for line in lines]
    data = parseMultipleJSON3(lines)

df = pd.DataFrame(data)
```

2. Vemos qué columnas tenemos disponibles para seleccionar nuestros datos

[157]:
```
df.columns
```

[157]:
```
Index(['created_at', 'id', 'id_str', 'text', 'source', 'truncated',
       'in_reply_to_status_id', 'in_reply_to_status_id_str',
       'in_reply_to_user_id', 'in_reply_to_user_id_str',
       'in_reply_to_screen_name', 'user', 'geo', 'coordinates', 'place',
       'contributors', 'retweeted_status', 'quoted_status_id',
       'quoted_status_id_str', 'quoted_status', 'quoted_status_permalink',
       'is_quote_status', 'quote_count', 'reply_count', 'retweet_count',
       'favorite_count', 'entities', 'favorited', 'retweeted', 'filter_level',
       'lang', 'timestamp_ms', 'extended_entities', 'possibly_sensitive',
       'display_text_range', 'extended_tweet'],
      dtype='object')
```

3. Crearemos un nuevo datasheet en el que almacenaremos el dato que queremos representar, en nuestro caso, el idioma del tweet(lang)

[158]:
```
data2 = np.unique(df.lang, return_counts = True)
```

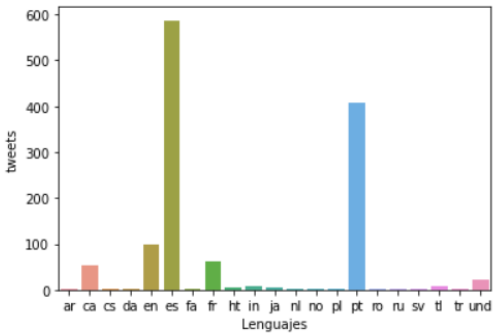
[159]:
```
data2
```

[159]:
```
(array(['ar', 'ca', 'cs', 'da', 'en', 'es', 'fa', 'fr', 'ht', 'in', 'ja',
        'nl', 'no', 'pl', 'pt', 'ro', 'ru', 'sv', 'tl', 'tr', 'und'],
       dtype=object),
 array([  1,  54,   1,   1,  98, 587,   1,  62,   6,   9,   5,   2,   1,
          1, 407,   1,   1,   1,   9,   3,  23], dtype=int64))
```

4.1 Representaremos el dato seleccionado en un gráfico de barras, por ejemplo

[160]:
```
plt = sns.barplot(x = data2[0], y = data2[1])
plt.set(xlabel = "Lenguajes", ylabel = "tweets")
```

[160]:
```
[Text(0.5, 0, 'Lenguajes'), Text(0, 0.5, 'tweets')]
```



4.2 Otra forma de representarlo es con gráfico de "quesito" (pie)

[161]:
```
import matplotlib.pyplot as plt
# He tenido que importar la librería de nuevo aquí porque daba fallo

fig1, ax1 = plt.subplots()
ax1.pie(data2[1],
        labels=data2[0],
        colors = sns.color_palette("hls",21),
        autopct='%1.1f%%',
        shadow=True,
        # Intenté hacer el gráfico más grande cambiando el radio pero no funciona
        radius = 100,
        startangle=90)
ax1.axis('equal')

plt.show()
```