Práctica 1: Web scraping

Datos sobre los resultados de la Fórmula 1 a lo largo de la historia.

1. Contexto

Se ha optado por crear un set de datos que tenga la información de los resultados de la Fórmula 1 desde sus inicios. Según los organizadores este tipo de eventos las primera carreras oficiales son del 1950, y se realizan anualmente en diferentes circuitos del mundo.

Los datos han sido recopilados desde la web oficial de la Fórmula 1. http://www.formula1.com. Con un poco más de detalle, la página que muestra los resultados de cada carrera a lo largo de los años está en el siguiente enlace: https://www.formula1.com/en/results.html.

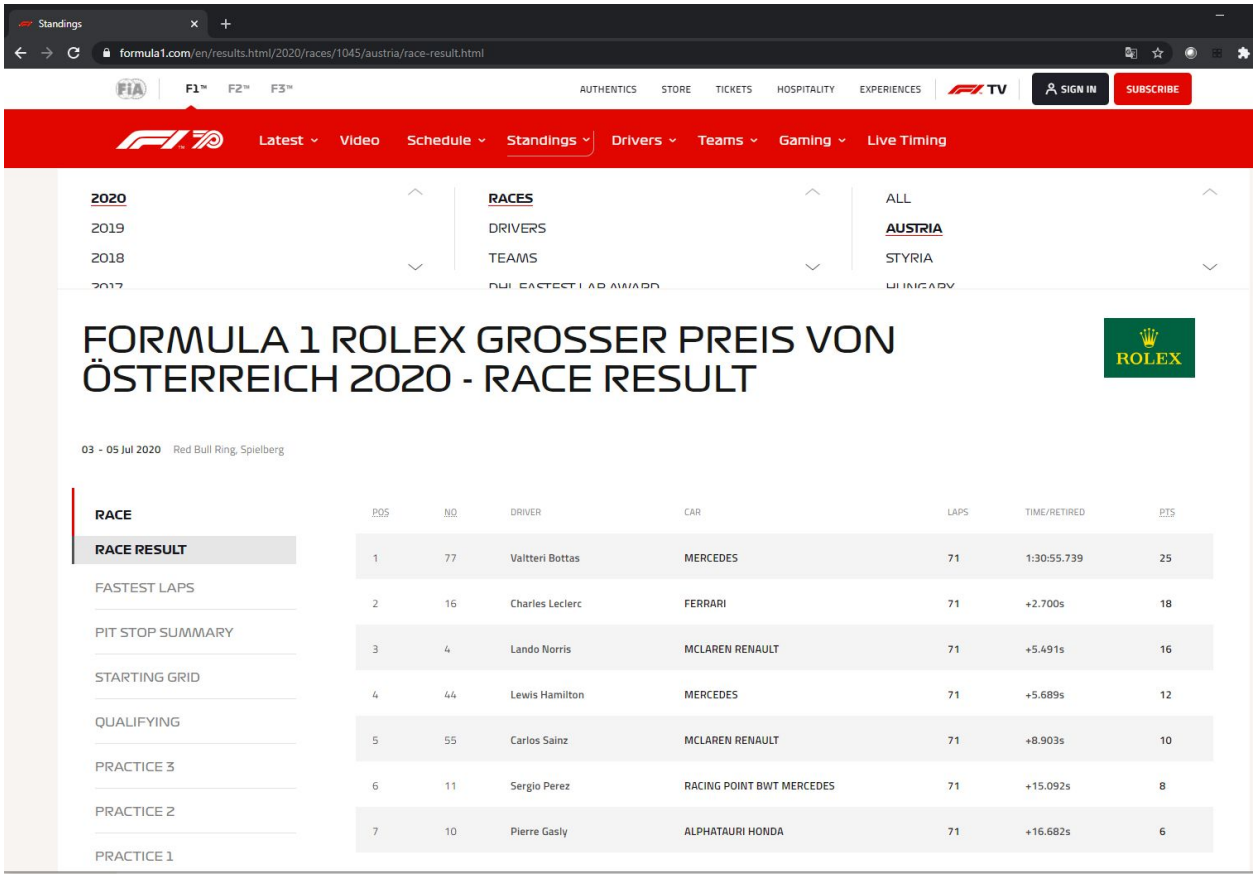
La página permite navegar por los diferentes años, y escoger un año concreto, cargando la información específica de ese año se puede escoger por tipología de información, en nuestro caso serían las carreras (Races), finalmente podemos seleccionar la información específica de cada circuito donde se ha realizado la carrera. Por defecto carga un resumen de quién ha ganado dicha competición, pero en nuestro caso queremos la información completa de como han quedado todos los pilotos que han participado en el evento deportivo.

Una vez escogido el circuito, observamos en la parte izquierda que hay la información de las fechas del evento, el nombre comercial del circuito y la localidad donde se celebra. Además justo debajo observamos diferentes selectores sobre la información a visualizar, en nuestro caso la que se obtiene por defecto, Race Result, es la que contiene la tabla con los resultados que queremos extraer.

Observamos también la tabla de resultados de la carrera. Cada línea es uno de los participantes, y para cada uno se facilitan varios datos, la posición que ha terminado en la carrera, el número del vehículo, el nombre y apellidos del conductor, la escudería o equipo, el número de vueltas que ha dado al circuito, el tiempo de la carrera, que como se observa es el tiempo del primero, el resto que ha entrado en la misma vuelta tiene la diferencia con el primero, y el resto de vehículos doblados indica cuantas vueltas respecto al primero. Y finalmente el número de puntos obtenidos según la posición en la que han finalizado la carrera.

Además, se ha observado que hay información que se visualiza en función del dispositivos, en especial hay el acrónimo o alias de los pilotos, y que también recopilaremos.

A modo de ejemplo, esta es la vista de los resultados del primer gran premio de este año 2020.



Finalmente comentar en esta introducción que se ha revisado el contenido del fichero robots.txt, y se no se han observado restricciones.

https://www.formula1.com/robots.txt
Sitemap: https://www.formula1.com/content/fom-website/en.sitemap-index.xml
User-Agent: *
Disallow:
Allow: /

2. Definir un título para el dataset

En este caso, un título descriptivo podría ser el de "***Formula_1_results.csv***".

3. Descripción del dataset

Para cada uno de los años disponibles se ha obtenido el listado de de las carreras realizadas, y para cada una se han extraído los datos de contexto de la carrera, fecha, nombre del circuito y país donde se celebra el evento, y los relativos al resultado de la carrera.

De cada carrera se obtiene la posición que ha quedado cada participante, el número del vehículo, el nombre, apellido y alias del piloto, el equipo o escudería, el número de vueltas realizado, el tiempo que se ha tardado en realizar el circuito, y finalmente el número de puntos que se le han otorgado según la posición en la carrera.

4. Representación gráfica.

A modo visual, se muestran las primeras y las últimas entradas del dataset.

	date	country	circuit	position	car_num	name	surname	alias	team	laps	duration	points
0	2020-07-05	Austria	Red Bull Ring, Spielberg	1	77	Valtteri	Bottas	BOT	Mercedes	71	1:30:55.739	25
1	2020-07-05	Austria	Red Bull Ring, Spielberg	2	16	Charles	Leclerc	LEC	Ferrari	71	+2.700s	18
2	2020-07-05	Austria	Red Bull Ring, Spielberg	3	4	Lando	Norris	NOR	McLaren Renault	71	+5.491s	16
3	2020-07-05	Austria	Red Bull Ring, Spielberg	4	44	Lewis	Hamilton	HAM	Mercedes	71	+5.689s	12
4	2020-07-05	Austria	Red Bull Ring, Spielberg	5	55	Carlos	Sainz	SAI	McLaren Renault	71	+8.903s	10

	date	country	circuit	position	car_num	name	surname	alias	team	laps	duration	points
23093	1950-09-03	Italy	Autodromo Nazionale Monza, Italy	NC	42	Maurice	Trintignant	TRI	Simca-Gordini	13	DNF	0
23094	1950-09-03	Italy	Autodromo Nazionale Monza, Italy	NC	46	Consalvo	Sanesi	SAN	Alfa Romeo	11	DNF	0
23095	1950-09-03	Italy	Autodromo Nazionale Monza, Italy	NC	44	Robert	Manzon	MAN	Simca-Gordini	7	DNF	0
23096	1950-09-03	Italy	Autodromo Nazionale Monza, Italy	NC	30	Prince	Bira	BIR	Maserati	1	DNF	0
23097	1950-09-03	Italy	Autodromo Nazionale Monza, Italy	NC	28	Paul	Pietsch	PIE	Maserati	0	DNF	0

5. Contenido.

El conjunto de datos está compuesto por 12 atributos, que son:

Date:	Fecha del evento.
Country:	País donde se realiza el evento.
Circuit:	Nombre comercial del circuito y población donde se localiza.
Position:	Resultado de la carrera.
Car_num:	Número que tiene asignado el vehículo.
Name:	Nombre del piloto.

Surname:	Apellido del piloto.
Alias:	Alias corto del piloto.
Team:	Equipo o escudería.
Laps:	Número de vueltas que se ha dado al circuito.
Duration:	Tiempo requerido por el ganador.
Points:	Puntos ganados en función de la posición.

Los registros van desde 1950, año de la primera carrera oficial de Fórmula 1, hasta la actualidad. La ejecución del código descarga toda la información hasta la última carrera que tenga los datos subidos en la web.

El proceso de descarga empieza con una primera extracción de todos los años que existen datos, y genera las url de cada año. Esta extracción se guarda en una lista.

Con dicha información, se realiza una nueva conexión por cada año y se extraen los datos de todas las carreras que ha habido en dicho año, generando una url para cada uno de los eventos. Esta información también se guarda en una lista.

Finalmente, se realiza una conexión por cada una de las url generada para cada gran premio, y se extrae la información de la carrera, con todo los atributos identificados. Con el objetivo de evitar una sobrecarga sobre la página, se ha incorporado un retraso de 1 segundo al finalizar este último paso.

El tiempo de ejecución es elevado ya que en total hay poco más de 1030 carreras realizadas, por lo que tarda más de 30 minutos.

6. Agradecimientos.

Los datos se han extraído de la página http://www.formula1.com, gestionada por “**Formula One Digital Media Limited**”.

Se agradece a dicha organización el esfuerzo que han realizado para habilitar la posibilidad de utilizar cierta información de su web adaptando los derechos de autor para fines educativos, permitiendo a los estudiantes descargar la información para propósitos educativos y de investigación, siempre que no se altere la misma. Dichos términos pueden ser consultados en esta página web: https://www.formula1.com/en/toolbar/legal-notices.html.

7. Inspiración.

La Fórmula 1 es un deporte muy competitivo, donde los equipos el éxito de un campeonato depende de muchos factores, tanto de la innovación tecnológica como de la estabilidad y consistencia del piloto.

Teniendo la información de los resultados de todas las carreras, se puede obtener información estadística sobre el rendimiento de los pilotos y las escuderías, detectar que circuitos se le dan mejor a que pilotos, o incluso se podría obtener una visión de la evolución de los pilotos durante su periódo profesional, y de las escuderías durante periodos de tiempo más extensos, permitiendo analizar la tendencia de los resultados.

8. Licencia.

Los datos de este dataset tienen ciertas restricciones de uso por parte del propietario de los mismos.

Es necesario referenciar al autor de los datos, en este caso www.formula1.com.
No se pueden utilizar para fines comerciales.
No se pueden cambiar los datos.
No se pueden cambiar el tipo de licenciamiento de los datos en su redistribución.

Tras revisar los diferentes tipos de licenciamiento disponibles, el que creo más se ajusta a dichos requisitos es el CC BY-NC-SA.

Attribution-NonCommercial-ShareAlike *This license lets others remix, adapt, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.* *Source:* https://creativecommons.org/licenses/	

Este licenciamiento permitiría a otros trabajar con el conjunto de datos, pero siempre referenciando al generador de los datos, no utilizándolos para fines comerciales, no cambiando los datos y no cambiando el tipo de licenciamiento tras su uso.

Para ver las diferencias se ha utilizado la información que se encuentra en la web de Creative Commons: https://creativecommons.org/licenses/

9. Código.

```
import requests
from bs4 import BeautifulSoup, NavigableString, Tag
import csv
from datetime import datetime
import pandas as pd
import time
from tqdm import tqdm

# Raíz de la página sobre la que trabajaremos
url_root = "http://www.formula1.com"
# Página específica donde iniciaremos nuestra extracción de información
url_start = "https://www.formula1.com/en/results.html"

# Creamos un dataset vacio con todas las columnas que vamos a obtener
Formula_1 = pd.DataFrame(columns=['date','country','circuit','position','car_num','name','surname','alias',
'team','laps','duration','points'])

# Función extracción url relativos a los diferentes años de los que hay datos.

def F1_year_extract(link):
    # Realizamos la petición a la web de la Fórmula 1.
    connection = requests.get(link)

    # Comprobamos que la petición nos devuelve un Status Code = 200
    statusCode = connection.status_code

    # Si obtenemos los datos, realizamos el proceso, en caso contrario, terminamos.
    if statusCode == 200:

        # Creamos una lista vacía para alojar las url de cada año
        years_url_list = []

        # Descargamos la página raíz de los resultados
        content = requests.get(link).text
        soup = BeautifulSoup(content, "lxml")

        # Acotamos al código donde se especifican las url de cada año
        cod_url_years = soup.find('div', {'class': 'resultsarchive-filter-container'})

        # Acotamos a cada división que contiene la url que buscamos
        data_year = cod_url_years.find_all('li', {'class': 'resultsarchive-filter-item'})

        # Extraemos el año y la url donde buscaremos los resultados y lo incorporamos a la lista creada
        for dy in data_year:
            year = dy.find("span").getText()
            if year.isdigit():
                link = dy.find("a").get("href")
                url = "%s%s" % (url_root, link)
                years_url_list.append(url)  # Lista "path" años.


        return years_url_list
    else:
        print("Error de carga en la página inicial:",link)
```

```
# Función extracción url relativos a cada gran premio de cada año.

def F1_prix_extract(link):
    # Realizamos la petición a la web de la Fórmula 1.
    connection = requests.get(link)

    # Comprobamos que la petición nos devuelve un Status Code = 200
    statusCode = connection.status_code

    # Si obtenemos los datos, realizamos el proceso, en caso contrario, terminamos.
    if statusCode == 200:

        # Creamos una lista vacía para almacenar los link a cada evento concreto
        prix_url_list = []

        # Descargamos la página de cada año
        content = requests.get(link).text
        soup = BeautifulSoup(content, "lxml")


        # Seleccionamos la división de donde extreremos los links de los eventos.
        table = soup.find('table', {'class':'resultsarchive-table'})

        table_body = table.find('tbody')
        rows = table_body.find_all('tr')
        for row in rows:
            link = row.find('a', {'class':'dark bold ArchiveLink'}).get("href")
            country = row.find('a', {'href':link}).getText(strip=True)
            date = row.find('td',{'class':'dark hide-for-mobile'}).getText(strip=True)
            url = "%s%s" % (url_root, link)
            prix_url_list.append((country,date,url))

        return(prix_url_list)

    else:
        print("Error de carga en la página de un año concreto:",link)


# Función extracción información de resultados para cada gran premio.

def F1_data_extract(link):

    # Realizamos la petición a la web de la Fórmula 1.
    connection = requests.get(link)

    # Comprobamos que la petición nos devuelve un Status Code = 200
    statusCode = connection.status_code

    # Si obtenemos los datos, realizamos el proceso, en caso contrario, terminamos.
    if statusCode == 200:

        # Descargamos la página de cada evento
        content = requests.get(link).text
        soup = BeautifulSoup(content, "lxml")

        # Acotamos al código donde se especifica el nombre del circuito
        circuit = soup.find('span', {'class': 'circuit-info'}).getText()
```

```
        table = soup.find('table', {'class':'resultsarchive-table'})
        table_body = table.find('tbody')
        rows = table_body.find_all('tr')

        records = []

        for row in rows:
            cols = row.find_all('td')
            pos = cols[1].getText()
            no = cols[2].getText()
            name = cols[3].find('span',{'class':'hide-for-tablet'}).getText()
            surname = cols[3].find('span',{'class':'hide-for-mobile'}).getText()
            alias = cols[3].find('span',{'class':'uppercase hide-for-desktop'}).getText()
            team = cols[4].getText()
            laps = cols[5].getText()
            duration = cols[6].getText()
            points = cols[7].getText()
            records.append((circuit,pos,no,name,surname,alias,team,laps,duration,points))

        # Incorporamos un retraso de 1 segundo después de cara obtención de datos
        time.sleep(1)
        return(records)

    else:
        print("Error de carga en la página de un evento concreto:",link)


# Obtenemos el listado de url de cada año.
F1_url_by_year = F1_year_extract(url_start)


# Obtenemos el listado de urls de cada gran premio según el año.
F1_event_url_list = []

# Vamos a extraer la información para cada url de cada año
for season in tqdm(F1_url_by_year):

    # Extraemos la lista de url de eventos por cada año
    url_event_list = F1_prix_extract(season)

    # Extraemos los datos de la lista para crear una única lista con la información que nos interesa.
    for event in url_event_list:
        country = event[0]
        date = event[1]
        link_e = event[2]
        reg = (country,date,link_e)
        # Esta lista contendrá todas las url de todos los eventos.
        F1_event_url_list.append(reg)


# Obtenemos los resultados de cada carrera

#for race in F1_event_url_list:
for race in tqdm(F1_event_url_list):

    # Identificamos la información que tenemos en cada elemento de la lista
    country = race[0]
    date = race[1]
```

```
    link = race[2]

    # Extraemos los datos de cada carrera
    data = F1_data_extract(link)

    #recorrer la lista data de los resultado de ese evento concreto
    for result in data:

        date_f = datetime.strptime(date, '%d %b %Y').date()

        # Incorporar en el dataset cada registro.
        new_reg = {'date':date_f,
                   'country':country,
                   'circuit':result[0],
                   'position':result[1],
                   'car_num':result[2],
                   'name':result[3],
                   'surname':result[4],
                   'alias':result[5],
                   'team':result[6],
                   'laps':result[7],
                   'duration':result[8],
                   'points':result[9]
                  }
        Formula_1.loc[len(Formula_1)] = new_reg


# Escribimos el dataset obtenido en un fichero csv
Formula_1.to_csv('Formula_1_results.csv', index=False)
```

10. Dataset.

Se ha creado una cuenta en Github. https://github.com/Idomingog donde se dejará la información del proyecto.

https://zenodo.org/

Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.