Práctica 1: Web scraping-F1.

Miembros:

La actividad se ha realizado en grupo por Mario Ubierna San Mamés y Moreyba García Cedrés.

Contexto:

Esta práctica se ha realizado en el lenguaje de programación Python. En ella se aplican técnicas de web scraping para extraer datos de la web https://www.f1-fansite.com. Esta web está especializada en formula 1 (F1), el objetivo de esta web es poner a disposición de los aficionados todo el mundo, de este deporte, la mayor cantidad de información gratuita y de alta calidad posible, sobre el mismo. Con el fin de ser el mejor sitio web de F1. Además de información sobre pilotos, equipos y carreras, presta otros servicios como las noticias más actuales sobre F1, venta de boletos y productos de formula 1.

El sito web antes mencionado, contiene datos de archivo de los campeonatos de F1, desde 1950 hasta la actualidad, como son los resultados de las carreras, los tiempos de clasificación, las posiciones del campeonato, así como las escuderías que han participado, etc.

Toda esta información está desglosada por años, es decir, que podemos consultar todas las carreras disputadas durante un año concreto, pero no la evolución de una determinada carrera a lo largo de los años. Es por ello, que como objetivo de este proyecto se pretende es extraer un dataset, donde se agregue toda la información por año, consiguiendo así una serie temporal de la información de cada una de las carreras desde 1950 hasta hoy.

La importancia de este proyecto reside en la construcción de un dataset que permita hacer una análisis de la evolución de este deporte, más adelante desarrollaremos este punto con más detalle.

Título del dataset:

“Estadísticas de las carreras de formula 1 desde sus inicios hasta nuestros días”.

Descripción del dataset:

El dataset de los datos extraído se corresponde a la información de todas las carreras de formula 1 que han tenido lugar desde 1950 hasta hoy. Gracias a ello podemos responder preguntas como por ejemplo: ¿Cómo han afectado los cambios tecnológicos a la mejora de los tiempos por vuelta?, ¿cuándo puede empezar a competir por un título una nueva escudería o piloto?, ¿qué piloto o escudería va a ganar la próxima carrera o campeonato?, ¿cuál será el ritmo por vuelta de un gran premio?...

Como podemos apreciar son numerosas las preguntas que podemos responder con este conjunto de datos, y eso lo conseguimos gracias a que cada registro contiene información sobre la carrera, nombre, año y también información sobre el piloto ganador de la carrera, nombre, nacionalidad y tiempo que tardó en completar el circuito. Además de la escudería a la que pertenece el piloto y la nacionalidad de la misma.

Cabe destacar que en el conjunto de datos original contenía el tiempo que se había tardado en completar la carrera, pero no venia una estimación del tiempo por vuelta. Es por ello, que hemos realizado ese cálculo para así poder realizar un estudio más minucioso.

En la siguiente tabla podemos observar, cómo se presenta la información a través de la página web que hemos extraído los datos:

RACE	British Grand Prix
DATE	May 13
WINNER	Nino Farina
TEAM	Alfa Romeo
LAPS	70
TIME	02:13:23.600

Observando la anterior tabla, vemos que los datos de los que se compone el dataset son:

- El nombre del gran premio, es una variable de tipo texto.
- La fecha en la que se corrió la carreara, variable de tipo fecha.
- La nacionalidad del piloto que ganó la carrera, variable de tipo texto.
- El nombre del piloto que ganó la carrera, variable de tipo texto.
- La nacionalidad de la escudería que ganó la carrera, variable de tipo texto.
- El nombre de la escudería que ganó la carrera, variable de tipo texto.
- Las vuelta que se dieron en la carrera, variable de tipo numérico.
- El tiempo que se tardó en completar la carrera, variable de tipo texto.
- La tiempo por vuelta de la carrera, variable de tipo texto.

Una vez que se ha realizado el web scraping sobre los datos, obtenemos la siguiente información, la cual será añadida al dataset:

RACE	British Grand Prix
DATE	13/05/1950
NATIONALITY_WINNER	Italy
WINNER	Nino Farina
NATIONALITY_TEAM	Switzerland
TEAM	Alfa Romeo
LAPS	70
TIME	02:13:23.600
TIME_PER_LAP	1:54.337

Representación gráfica:

Esta imagen representa visualmente el tema que hemos escogido para nuestro proyecto de web scraping.

Esta imagen se ha extraído mediante técnicas de web scraping.

En la siguiente ilustración podemos ver una representación del ciclo de vida del proyecto, es decir, desde la fase para extraer los datos hasta su publicación en Zenodo:

1º/ Accedemos a F1-fansite (https://www.f1-fansite.com/f1-results/)

2/ Accedemos a cada uno de los años de forma ordenada

3/ Extraemos los valores de cada Gran Premio para ese año

4/ Almacenamos la información extraída en un fichero CSV

5/ Publicamos el fichero CSV en Zenodo

Tal y como podemos apreciar en la anterior ilustración, el objetivo principal de este proyecto es el de realizar web scraping sobre una página (en nuestro caso, https://www.f1-fansite.com/f1-results/), extraer dichos valores y almacenarlos en un fichero CSV, para así finalmente publicar ese dataset en Zenodo.

Contenido:

El conjunto de datos se corresponde a información de diferentes carreras de formula 1, cada registro se corresponde a una carrera con información sobre el piloto, la escudería y el año que tuvo lugar, etc. A continuación se detalla cada uno de los campos informados en el dataset:

RACE: Nombre de la carrera de formula 1.

DATE: Fecha en la que tuvo lugar la carrera con formato dd/mm/aaaa.

NATIONALITY_WINNER: Nacionalidad del piloto que ganó la carrera.

WINNER: Nombre del piloto que ganó la carrera.

NATIONALITY_TEAM: Nacionalidad de la escudería a la que pertenece el piloto.

TEAM: Nombre de la escudería.

LAPS: Número de vueltas que comprenden la carrera en formato de números enteros.

TIME: Tiempo que tardó el ganador en completar todas las vueltas de la carrera en formato hh:mm:ss.ff.

TIME_PER_LAP: Campo calculado, el cual se ha obtenido del dividir el TIME entre LAPS, los que nos da el tiempo promedio que el piloto ha tardado en recorrer una vuelta.

La web https://www.f1-fansite.com de la cual se han extraído los datos fue creada en 2001 y recoge datos de carreras de formula 1 desde el año 1950 hasta hoy.

Para recopilar los datos de dicha web se ha utilizado como lenguaje de programación Python y como se pudo ver el diagrama del apartado anterior, para la realizar el web scraping, en primer lugar, hemos accedido a la web https://www.f1-fansite.com/f1-results/, conectando con el servidor utilizamos la librería CloudScraper, posteriormente parseamos el contenido de la web a un objeto BeatifulSoup con el que navegar por el dominio. Recorremos cada uno de los años de forma ordenada extrayendo los campos descritos para cada gran premio y obtenemos la tabla de nuestro dataset, por último guardamos la información extraída en un archivo csv.

Agradecimientos.

Si hacemos un Whois de la web seleccionada podemos comprobar que el propietario de los datos es f1-fansite.com que está registrada a nombre de Cloudflare, Inc.

Por otro lado, tenemos que dar nuestro agradecimiento a fansite-f1 por la gran labor que ha realizado, recopilando todos los datos relacionados con la formula 1 desde comenzó este deporte.

Por último, los análisis similares que hemos utilizado como referencia, han sido análisis sobre resultados deportivos de otros deportes como:

Los resultados de futbol que refleja la siguiente web:
https://www.football-data.co.uk/data.php

También se ha utilizado esta otra web sobre estadísticas de baloncesto:
https://www.nba.com/stats/players/traditional/?PerMode=Totals&sort=PTS&dir=-1.

Tras nuestra investigación previa pudimos comprobar que había gran cantidad de análisis sobre estos dos deportes, por lo que decidimos escoger otro deporte sobre el que pudiéramos aportar alguna innovación. Este fue uno de los motivos por el que escogimos la formula 1, pero la principal razón fue que era un deporte que nos interesaba a ambos.

Inspiración:

La principal motivación de este análisis es dar respuesta a las preguntas planteadas en el apartado descripción del dataset:
¿Cómo han afectado los cambios tecnológicos a la mejora de los tiempos por vuelta?, ¿cuándo puede empezar a competir por un título una nueva escudería o piloto?, ¿qué piloto o escudería va a ganar la próxima carrera o campeonato?, ¿cuál será el ritmo por vuelta de un gran premio?

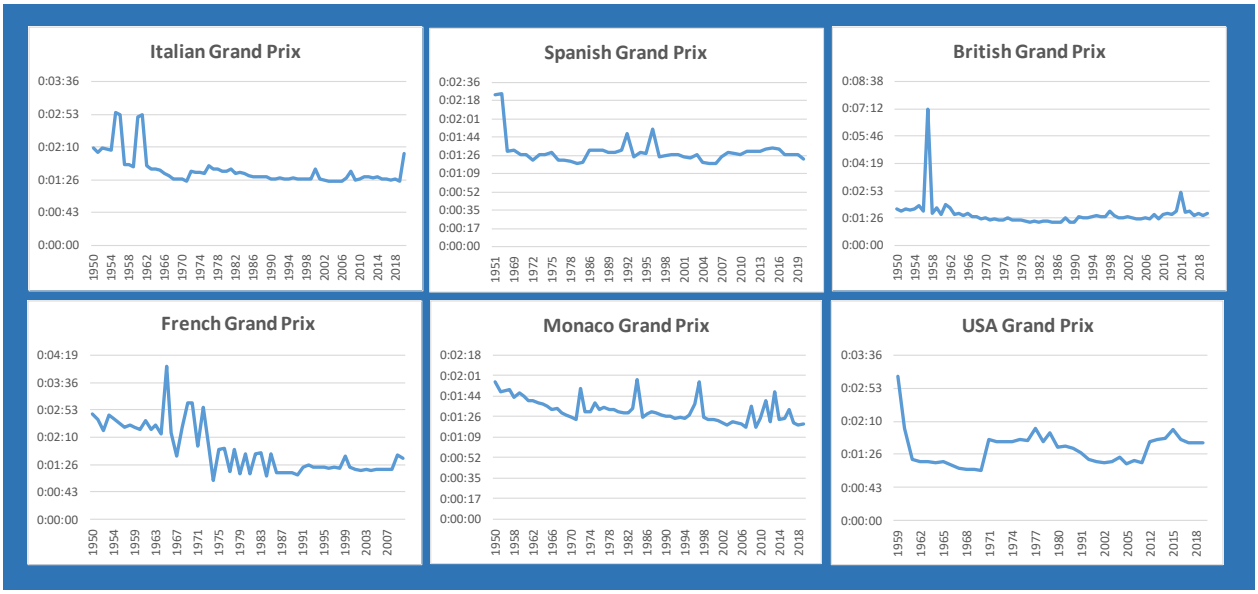
Creemos que el dataset que hemos construido aplicando técnicas de web scraping pude dar respuesta a estas preguntas. Ya que el conjunto de datos obtenido es una serie temporal que permite ver la evolución los tiempos por vuelta, los pilotos y escuderías ganadoras por carrera y año, analizar qué escuderías y pilotos tienen más títulos, comprobando si existen diferencias tecnológicas con respecto a otras escuderías, etc.

Como ejemplo de cómo puede dar respuesta a lo anterior, hemos realizado un pequeño análisis gráfico a partir de los datos obtenidos en nuestro dataset sobre formula 1, hemos elaborado unas representaciones gráficas de las carreras con más recorrido en la historia de la formula 1.

Cada una de las gráficas representa un circuito distinto y en ellas se representa la evolución de los tiempos promedios por vuelta de los ganadores de estas carreras a lo largo de los años.

Como podemos ver la tendencia general, en todos los circuitos, es la de reducir sus tiempos promedios debido a la mejora técnicas y aerodinámica de los monoplazas hasta 2006, temporada en la que la FIA introduce cambios en los motores con el fin de reducir las velocidades en la formula 1[1].

[1] Información obtenida de: https://es.wikipedia.org/wiki/Reglas_de_la_Fórmula_1



Además, de lo antes mencionado, este conjunto de datos es interesante por diversos motivos. El primero de ellos es que recopila los datos de todas las carreras de la historia de la formula 1 en su totalidad, lo cual es muy interesante para cualquier aficionado de este deporte. También resulta de interés para cualquier revista deportiva que quiera haces una retrospectiva sobre la F1.

Por otro lado, se podría considerar la construcción de un modelo de aprendizaje automático predictivo que usara los tiempos de calificación y los resultados de carreras anteriores para hacer predicciones sobre el posible resultado de una carrera.

En lo que respecta a la relación con los análisis de referencia, del Resultados históricos de fútbol y datos de probabilidades de apuestas, nos hemos inspirado para la parte predictiva del análisis y el análisis de las Estadísticas avanzadas de la NBA lo hemos tomado de referencia para obtener la tendencia a lo largo de los años de los tiempos promedios por vuelta.

Licencia:

Tras revisar las distintas licencias creative comomons (CC), creemos que la más adecuada para la publicación de nuestro dataset es la Creative Commons Attribution 4.0 International (CC BY 4.0).

Esta licencia solo tiene una restricción, para usar una obra en cualquier tipo de medio es indispensable citar al autor de la misma.

Esto nos da gran flexibilidad ya que con esta licencia permite que los datos puedan usarse de la forma que se desee, es decir, modifícalos y/o comercializarlos, siempre que se haya citado al autor para ello.

Código:

Para el desarrollo de la práctica el lenguaje de programación elegido ha sido Python. Se ha organizado en el proyecto de la siguiente manera:

- src/scraper: se encarga de hacer el scraping, desde la página inicial, va buscando por la web y obtiene los datos.

- src/csvmodule: se define la clase de creación del CSV. A partir de los datos que consigue el scraper, los introduce en un fichero.

- src/main: es el programa principal que inicia el proceso de scraping y generación del fichero csv con los datos resultantes del scraping.

- src/img: este un proceso independiente a los anteriores se encarga de hacer scraping buscando una imagen concreta y haciendo la descarga de la misma.

Cabe destacar que se han seguido las mejores prácticas para el desarrollo del web scraping de esta práctica:

- Se comprobó si existía una API, y sí que existen. Sin embargo, para la página y los datos que hemos obtenido no, es decir, las API que se encuentran por internet proporcionan información sobre los pilotos, escuderías, circuitos… Sin embargo, no proporcionan un histórico como los datos que obtenemos nosotros.

- No parseamos el HTML de forma manual, sino que hacemos uso de la librería BeutifulSoup.

- No saturamos las peticiones al servidor, nuestro dataset no es muy grande, por lo que no hacemos demasiadas peticiones al mismo. Cabe destacar que para hacer la conexión con el servidor, tuvimos que usar la librería CloudScraper, ya que la librería Request no nos permitía acceder a la web.

- Comprobamos si existía el fichero Robots.txt, pero no existe tal fichero en la página web. Es por ello que hemos considerado lícito el hacer uso de web scraping sobre la misma.

- Por último, hemos tenido en consideración la calidad de los datos, y éstos sí que cumplen con las seis dimensiones definidas por Data Management Association.

Dataset:

El DOI obtenido en la publicación del CSV ha sido:

10.5281/zenodo.4662943

url: https://zenodo.org/record/4662943

Tabla de contribuciones al trabajo:

Contribuciones	Firma
Investigación previa	MUSM, MGC
Redacción de las respuestas	MUSM, MGC
Desarrollo código	MUSM, MGC

MUSM: Mario Ubierna San Mamés.
MGC: Moreyba García Cedrés.