

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> y generar un dataset con datos con todas las carreras de formula 1 realizadas desde 1950 hasta hoy. En este sitio web podemos encontrar estadísticas, eventos por temporada y noticias de formula 1.

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 nos podemos responder con este conjunto de datos, y eso lo conseguimos 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 sí que venía el tiempo que se había tardado en completar la carrera, pero no venía una estimación del tiempo por vuelta. Es por ello, que nosotros hemos realizado esa cálculo para así poder realizar un mejor estudio.

En la siguiente tabla podemos observar el 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 carrera, 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 vueltas 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/>)

Seasons Championship Overview				
Formula One				
1950	1951	1952	1953	1954
1955	1956	1957	1958	1959
1960	1961	1962	1963	1964
1965	1966	1967	1968	1969
1970	1971	1972	1973	1974
1975	1976	1977	1978	1979
1980	1981	1982	1983	1984
1985	1986	1987	1988	1989
1990	1991	1992	1993	1994
1995	1996	1997	1998	1999
2000	2001	2002	2003	2004
2005	2006	2007	2008	2009
2010	2011	2012	2013	2014
2015	2016	2017	2018	2019
2020	2021			

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

2020 F1 World Championship					
RACE	DATE	WINNER	TEAM	LAPs	TIME
Australian Grand Prix	July 5	Lewis Hamilton	Mercedes	71	01:30:55.739
Spanish Grand Prix	July 12	Lewis Hamilton	Mercedes	71	01:22:50.683
Mongolian Grand Prix	July 19	Lewis Hamilton	Mercedes	70	01:36:12.473
British Grand Prix	August 2	Lewis Hamilton	Mercedes	52	01:28:01.203
50th Anniversary GP	August 9	Max Verstappen	Red Bull	52	01:19:41.993
Spanish Grand Prix	August 16	Lewis Hamilton	Mercedes	66	01:31:45.779
Singapore Grand Prix	August 30	Lewis Hamilton	Mercedes	44	01:24:08.761
Italian Grand Prix	September 6	Pierre Gasly	AlphaTauri	53	01:47:06.056
French Grand Prix	September 13	Lewis Hamilton	Mercedes	68	02:19:35.060
Russian Grand Prix	September 27	Lewis Hamilton	Mercedes	53	01:34:00.364
Eifel Grand Prix	October 11	Lewis Hamilton	Mercedes	60	01:35:49.641
Portuguese Grand Prix	October 25	Lewis Hamilton	Mercedes	66	01:29:56.828
Emilia Romagna Grand Prix	November 1	Lewis Hamilton	Mercedes	63	01:28:32.430
Abu Dhabi Grand Prix	November 15	Lewis Hamilton	Mercedes	58	01:42:19.313
Qatar Grand Prix	November 29	Lewis Hamilton	Mercedes	57	02:59:47.515
Saudi Grand Prix	December 6	George Russell	Mercedes	87	01:31:15.114
Abu Dhabi Grand Prix	December 13	Max Verstappen	Red Bull	55	01:36:28.645

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

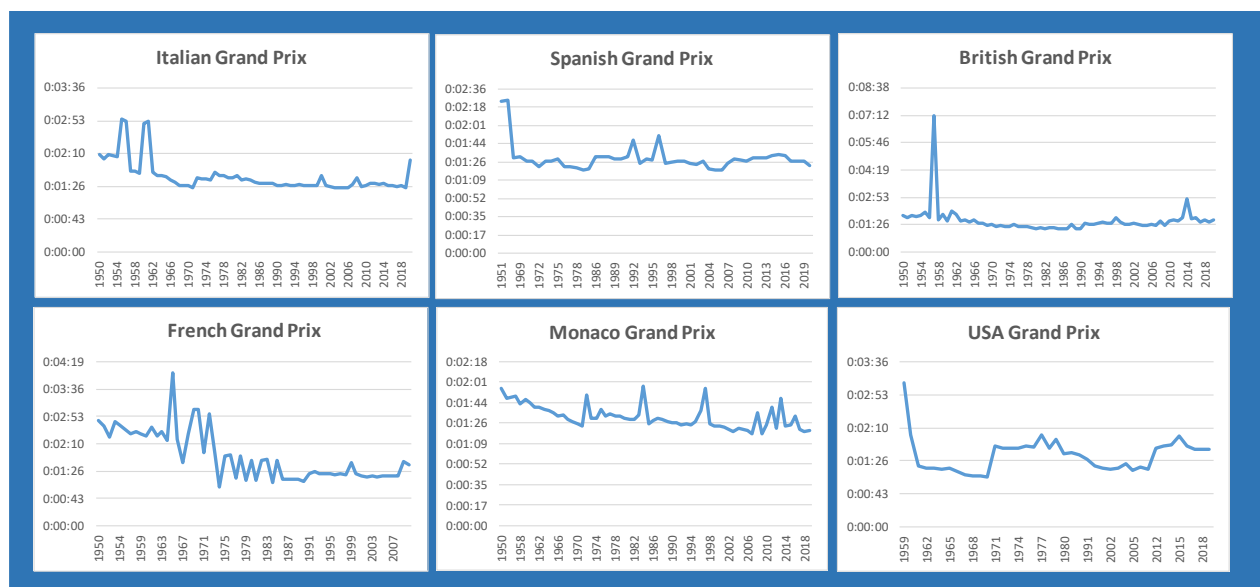


Tal y como podemos apreciar en la anterior ilustración, el objetivo principal de este proyecto es el de realizar web scraping sobre una página (en nuestro caso, <https://www.f1-fansite.com/f1-results/>), extraer dichos valores y almacenarlos en un fichero CSV, para así finalmente publicar ese dataset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

Por otro lado, 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 1¹.



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.

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

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 se han utilizado técnicas de web scraping para la extracción de los datos.

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:

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 hacer una retrospectiva sobre la F1.

Por último y la principal motivación de este análisis es dar respuesta cómo ha afectado los cambios tecnológicos y los cambios de normativa a los tiempos por vuelta de cada carrera a lo largo de los años. Para ello se podría analizar qué escuderías y pilotos tienen

más títulos, comprobando si existen diferencias tecnológicas con respecto a otras escuderías.

Por otro lado, si se análisis como ha evolucionado de los tiempos que han tardado en completar un circuito determinado los ganadores de la carrera, estudiando las causas de la obtención de los mismos, este análisis nos puede permitir predecir cual puede ser el futuro de un gran premio, es decir, la posible escudería y piloto ganador de dicho gran premio.

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, modificarlos 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 BeautifulSoup.
- 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	
Redacción de las respuestas	
Desarrollo código	