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.

Por otro lado, 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.

Representación gráfica:

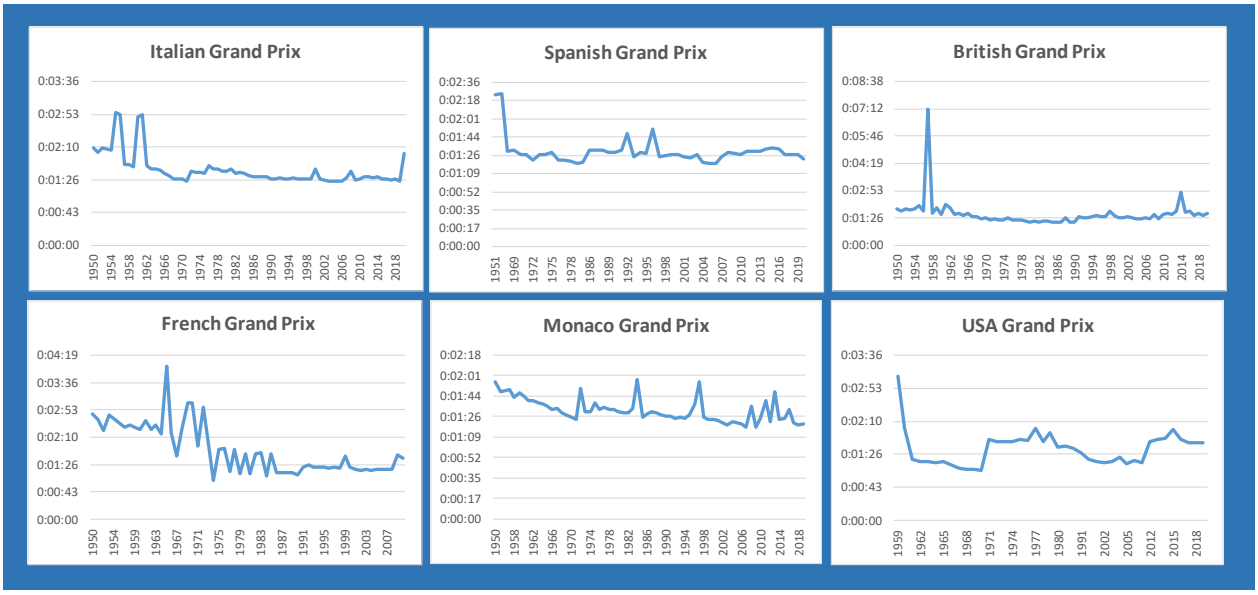
Esta imagen representa visualmente el tema que hemos escogido para nuestro proyecto de web scraping.

Esta imagen se ha extraído mediante técnicas de web scraping.

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 1[1].



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.

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

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 haces 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 tardo 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, 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.

Dataset:

El DOI obtenido en la publicación del CSV ha sido:

10.5281/zenodo.4647563

url: https://doi.org/10.5281/zenodo.4647563

Tabla de contribuciones al trabajo:

Contribuciones	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo código	