Práctica 1: Web Scraping

Francisco Fernández Poyato y Javier Gallego Fernández

Tipología y ciclo de vida de los datos

1. Contexto

El baloncesto es uno de los deportes más seguidos en el mundo, y la NBA (National Basketball Association) de Estados Unidos es la liga más prestigiosa del mundo. Durante toda la temporada, se juegan a diario partidos entre los 30 equipos que componen la liga. Al ser una liga en la que se disputan muchos partidos por temporada (en torno a los 1230 partidos), es una liga que constantemente está generando datos y por lo tanto es una fuente interesante de la cuál nutrir nuestro ejercicio.

En nuestro caso, vamos a recoer la información de los partidos que se disputan en la NBA, tomando los datos del equipo local, el visitante, el resultado y otros campos que veremos más adelante.

La información la hemos obtenido de la página oficial de la NBA, lugar donde se registran las estadísticas a tiempo real de los partidos que se disputan.

2. Título

Para esta práctica, el título que escogeremos para el dataset será : **"Partidos y resultados de la temporada 2021/2022 en la NBA"**

3. Descripción del dataset.

El conjunto de datos que se ha extraído contiene información acerca de los partidos de la NBA que se han disputado a lo largo de la temporada 2021/2022.

En este fichero, vamos a obtener la fecha en la que se disputará el partido, los equipos que se enfrentarán (diferenciando entre local y visitante) y los puntos anotados por cada equipo.

Además, decidimos crear una columna en función de la puntuación en la que se obtuviera el equipo ganador del encuentro a partir de evaluar el marcador, como se mostrará en el código.

4. Representación gráfica

En la imagen que se muestra a continuación, observamos el ejemplo de la página de donde cogeremos la información de cada partido. Como se puede observar, tendremos el resultado, los equipos y la fecha del partido, que será nuestros dtos objetivo.

5. Contenido

Este dataset que hemos creado estará compuesto por 6 columnas, de las cuáles se tienen 209 observaciones. Pasaremos a continuación a describir los campos que componen nuestro dataset.

· **DATE**: En este campo se recoge el día de la semana, mes, día y año del partido que se ha disputado

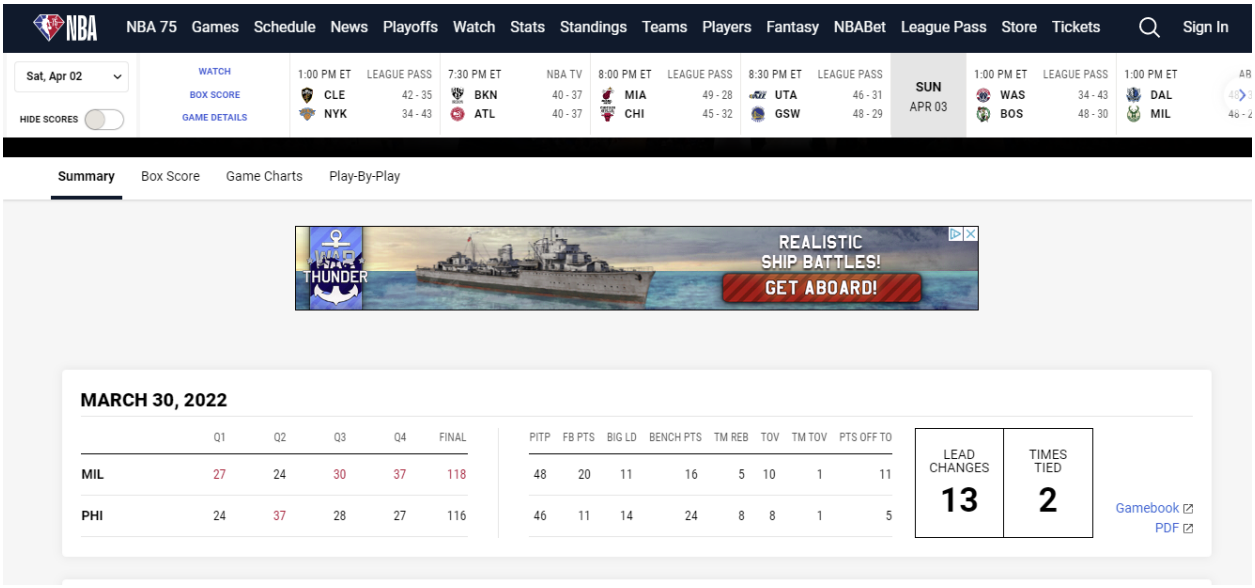

Figure 1: Ejemplo de página de donde se obtiene la información

· **TEAM_HOME**: En este campo se mostrará el equipo que jugará como local el encuentro.

· **TEAM_VISITOR**: En este campo se mostrará el equipo que jugará como visitante el encuentro

· **POINTS_HOME**: Aquí observaremos la puntuación del equipo local en el partido

· **POINTS_VISITOR**:Aquí observaremos la puntuación del equipo visitante en el partido

· **WINNER**: Por último, se muestra una columna en la que tendremos el equipo ganador del partido a partir de evaluar las columnas de los puntos de los equipos visitante y local.

6. Agradecimientos

El propietario de los datos es la National Basketball Association, ya que los datos provienen de la fuente oficial de la liga. Han sido muchos los diferentes proyectos que se han realizado de web scraping relacionados con el baloncesto, como por ejemplo este que adjuntamos, que se realiza a partir de la página Basketball Reference. Aquí se muestra un ejemplo de cómo puede ser un proyecto de scraping de datos de la NBA.

Para actuar de acuerdo con los principios éticos, hemos buscado información en la página web de origen si era posible utilizar el conjunto de datos sin tener ningún tipo de problema y que estaba permitido por la propia página.

Al buscar esta información, nos encontramos con el siguiente texto:

The Operator of this Site may make available on this Site statistics, including statistics generated and/or calculated by the Operator using proprietary calculations and analyses, relating to or arising out of the performance of players during or in connection with NBA, Women's National Basketball Association ("WNBA") NBA G League ("NBA G League") games, competitions or events (collectively, "NBA Statistics"). By using such NBA Statistics, you agree that: (1) any use, display or publication of the NBA Statistics shall include a prominent attribution to NBA.com in connection with such use, display or publication; (2) the NBA Statistics may only be used, displayed or published for legitimate news reporting or private, non-commercial purposes; (3) the NBA Statistics may not be used in connection with any sponsorship or commercial identification; (4) the NBA Statistics may not be used or referred to in connection with any gambling activity (including legal gambling activity); (5) the NBA Statistics may not be used in connection with any fantasy game or other commercial product or service; (6) the NBA Statistics may not be used in connection with any product or

service that presents a live, near-live or other real-time or archived play-by-play account or depiction of any NBA game; and (7) the NBA Statistics may not be used in connection with any web site, product or service that features a database (in any medium or format) of comprehensive, regularly updated statistics from NBA, WNBA or D-League games, competitions or events without the Operator's express prior consent.

De este documento, se obtiene que siempre y cuando no se haga uso comercial de los datos será lícito obtener los datos de esta página, por lo que no inferimos en ningún problema legal ni actuamos en contra de los principios éticos.

7. Inspiración

El conjunto de datos que hemos seleccionado, aunque pueda aparentar ser simple, a partir de él podremos realizar análisis bastante interesantes, como por ejemplo, contabilizar el número de victorias del equipo a lo largo de la temporada, hacer series temporales sobre cómo ha ido variando esa cifra de victorias con el paso de los meses, o estudiar si es diferencial el hecho de jugar como local.

8. Licencia

Para realizar esta práctica, utilizaremos la licencia **Released Under CC0: Public Domain License**, ya que esta licencia nos otorgará derechos de dominio público, al mismo tiempo que evitaremos la complejidad de la atribución de compatibilidad de licencias que se producen con otras.

9. Código

A continuación, adjuntamos el enlace a GitHub, donde tendremos el código y el dataset de esta práctica.

10. Dataset

Por último, adjuntamos el enlace a DOI, donde hemos subido el csv resultante del proceso de webscraping