Práctica 1: Web scraping

UOC - Tipología y ciclo de vida de los datos

Miguel Santos Pérez y Alejandro Arzola García

14 de abril de 2020

Índice

1 Contexto

El auge del valor del dato y su aplicación a cualquier ámbito de la sociedad es claro desde su explosión unos años atrás. En este sentido, cada vez más tecnologías basadas en el poder del dato se emplean con eficiencia en distintos deportes como el fútbol, para diversas tareas tales como el fichaje de jugadores, campañas de marketing o estrategias de partidos. Técnicas similares son aplicadas al tenis.

En cuanto al deporte del que nos ocuparemos en este trabajo, concretamente el baloncesto vivimos en un mundo bipolar. Mientras en América, las poderosas franquicias NBA cada vez lo utilizan más en su día a día con grandes equipos de Data Scientist, en Europa el gasto en la explotación del dato se mantiene en segundo plano. Así pues, la inspiración de este estudio es desarrollar el webscrapping sobre el conjunto de datos de baloncesto en Europa, como primera aproximación para luego utilizar estos datos y buscarles su valor. En este primer trabajo, por tanto, se tratará de acceder a los partidos de Euroliga y almacernarlos en datasets que sean analizables.

Todos estos resultados se encuentran en la web oficial de la competición, www.euroleague.net

2 Componentes del grupo

- Miguel Santos Pérez (miguel8santos@uoc.edu)
- Alejandro Arzola García (aarzola@uoc.edu)

3 Repositorio Github

Para la realización de esta práctica se ha creado un repositorio en *GitHub* para trabajar de manera colaborativa y tener un control de versiones sobre el código fuente. Se puede acceder a este repositorio a través del siguiente enlace:

- https://github.com/aarzola-uoc/practica1-tycvd

4 Título para el dataset

Euroliga 2019-2020. Marcadores y estadísticas.

5 Descripción del dataset

Se propone para este trabajo la obtención de dos datasets, el primero con todos los marcadores, partidos y enlaces a estadísticas y, el segundo, con el detalle de las propias estadísticas para su posterior análisis.

6 Representación gráfica

7 Contenido

En el presente trabajo, se ha desarrollado código para crear dos datasets. El primero, por todos los resultados de la Euroliga, en el que se incluye campo, horario, equipo local, equipo visitante y enlace a las estadísticas.

En el segundo, se presentan en sí las propias estadísticas desagregadas de cada partido a nivel jugador y equipo.

El periodo de recolección de los datos ha sido desde el inicio de la competición (3 de octubre) hasta la última jornada (5 de marzo, debido a la cancelación por la crisis del Coronaviurs).

8 Publicación del dataset

Se han publicados los dos *datasets* obtenidos en el repositorio web **Zenodo** y se pueden acceder a través del siguiente enlace:

- https://zenodo.org/record/3740661#.XoouwVNKgnU

El DOI (*Digital Object Identifier*) asignado ha sido el siguiente:

- 10.5281/zenodo.3740661

9 Agradecimientos

Los datos, recogidos de la web oficial de la competición, son propiedad de © Euroleague Ventures SA. Para ello se ha hecho uso del lenguaje de programación Python y de técnicas de WebScrapping para extraer la información alojada en las páginas.

10 Inspiración

De la mano de lo explicado en la contextualización, se hace interesante la disponibilidad de los datos de una manera sencilla de analizar como primer paso de cara a un posterior uso en el planteamiento de los partidos o de las semanas de entrenamiento. La mayor ambición es la realización de un análisis profundo que permita a los equipos disponer de una herramienta que les proporcione en los datos un valor añadido. A partir de las estadísticas recogidas por Euroliga, se podrían crear nuevas estadísticas más precisas que recojan otro tipo de datos en función de los ya existentes.

11 Licencia

12 Contribuciones al trabajo

Contribuciones	**Firma**
Investigación previa	MSP, AAG
Redacción de las respuestas	MSP, AAG
Creación del repositorio GitHub	MSP, AAG
Desarrollo código	MSP, AAG
Publicación de los datasets	MSP, AAG