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

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

Componentes del grupo

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

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

Título para el dataset

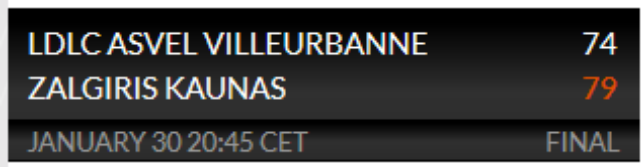
Euroliga 2019-2020. Marcadores y estadísticas por partido.

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.

Representación gráfica

El proyecto representa, para cada partido, en un dataset los siguientes datos. El primer dataset, recoge los resultados y es el siguiente:



Dataset 1

El segundo dataset, recoge las estadísticas individuales dentro de un partido y representa:

LDLC ASVEL VILLEURBANNE																	
							Rebounds						Blocks		Fouls		
#	Player	Min	Pts	2FG	3FG	FT	O	D	T	As	St	To	Fv	Ag	Cm	Rv	PIR
1	JACKSON, EDWIN	6:31			0/1							2					-3
2	**TAYLOR, JORDAN**	28:28	15	2/4	3/6	2/2	1	1	2	2	1	2			4	2	11
3	**JEKIRI, TONYE**	28:49	11	4/6		3/5	3	4	7	2		2	1		4	4	15
6	**MALEDON, THEO**	21:30	13	2/4	1/1	6/6		1	1	5					5	3	15
9	LOMAZS, RIHARDS	13:24								1					2	2	1
11	**GALLIOU, CHARLES**	14:31			0/1					1					3		-3
17	**JEAN-CHARLES, LIVIO**	24:41	12	3/6	1/3	3/3	1	3	4		1	1			1	2	12
19	DIOT, ANTOINE	1:48			0/1										1		-2
21	BAKO, ISMAEL	8:51	6	2/4		2/2	3	2	5				1		1	1	10
23	LIGHTY, DAVID	25:20	14	4/8	1/5	3/3	2	2	4	3	1	1			1	5	17
32	STRAZEL, MATTHEW	8:28		0/2	0/2					1						2	-1
33	PAYNE, ADREIAN	17:39	3	0/2	0/4	3/4	3	3	6				1			2	5
	Team						**1**	**3**	**4**			**2**					**2**
	Totals	**200:00**	**74**	**17/36**	**6/24**	**22/25**	**14**	**19**	**33**	**15**	**3**	**10**	**3**	**0**	**22**	**23**	**79**
				47.2%	**25%**	**88%**											

ZALGIRIS KAUNAS																	
							Rebounds						Blocks		Fouls		
#	Player	Min	Pts	2FG	3FG	FT	O	D	T	As	St	To	Fv	Ag	Cm	Rv	PIR
0	**WALKUP, THOMAS**	29:23	11	2/4	1/1	4/4	1	5	6	4		2			1	5	21
1	RIVERS, KC	12:35			0/1			1	1								
4	LEKAVICIUS, LUKAS	27:26	21	7/11	2/3	1/2				1	1	1			2	1	15
10	HAYES, NIGEL	15:24	3		1/5										4	1	-4
13	**JANKUNAS, PAULIUS**	20:21	5	2/3	0/1	1/2	1	3	4		1				5	1	3
16	LUKOSIUNAS, KAROLIS	DNP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	**MILAKNIS, ARTURAS**	21:56	12	3/4	2/3	0/1		1	1		1	2			2	1	8
23	GEBEN, MARTINAS	DNP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
31	JOKUBAITIS, ROKAS	3:18													1		-1
32	LEDAY, ZACH	18:00	9	3/9		3/5	4	2	6	1		2		3		5	8
34	**LANDALE, JOCK**	22:00	12	5/6		2/2	1	3	4	1		2			4	2	12
92	**ULANOVAS, EDGARAS**	29:37	6	2/3	0/1	2/2		3	3	3					4	6	12
	Team							**4**	**4**			**1**					**3**
	Totals	**200:00**	**79**	**24/40**	**6/15**	**13/18**	**7**	**22**	**29**	**10**	**3**	**10**	**0**	**3**	**23**	**22**	**77**
				60%	**40%**	**72.2%**											

Contenido

Así, definimos la estructura de los dataset como:

- **Euroleague_Scoreboards**: Dicho dataset recoge los resultados por partido y se compone de siete campos:
  - MatchId: identificador único del partido.
  - Date: fecha del partido.
  - HomeTeam: nombre del equipo local.
  - HomeScore: puntos anotados por el equipo local
  - VisitingTeam: nombre del equipo visitante.
  - VisitingScore: puntos anotados por el equipo visitante.
  - Link: link de estadísticas.
- **Euroleague_Stats_per_Game**: Dicho dataset recoge las estadísticas a nivel jugador por partido y se compone de:
  - MatchId: identificador único de partido.
  - Team: equipo al que pertenece el jugador.
  - PlayerNumber: número del jugador.
  - PlayerName: nombre del jugador.
  - Min: minutos jugados.
  - Pts: puntos anotados.
  - 2FG: tiros de dos (metidos-anotados).
  - 3FG: tiros de tres (metidos-anotados).
  - FT: tiros libres (metidos-anotados).
  - O: rebotes ofensivos.

- D: rebotes defensivos.
- T: rebotes totales.
- As: asistencias.
- St: recuperaciones.
- To: pérdidas.
- Fv: tapones a favor.
- Ag: tapones en contra.
- Cm: faltas cometidas.
- Rv: faltas recibidas.
- PIR: valoración del jugador.

El periodo de recolección de los datos es esta temporada, desde los partidos qeu empezaron el 3 de octubre hasta los terminados el 5 de marzo.

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

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.

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.

Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido **CC BY-SA 4.0 License**. A continuación detallamos los motivos principales qeu han llevado a escoger esta licencia: - *Se debe proveer proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado*. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en realidad con el trabajo original. - *Se permite un uso comercial*. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento a los autores. - *Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma*. Eso hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó.

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

Código

El proceso de extracción ha sido bastante sencillo. En primer lugar, accedimos al fichero *robots.txt* para la web de la Euroliga, disponible en el siguiente enlace:

- https://www.euroleague.net/robots.txt

El contenido de este fichero se muestra a continuaciÃ³n:

```
User-agent:                                                    Sosospider
Disallow:                                                               /

User-agent:                                                        Yandex
Disallow:                                                               /

User-agent:                                                   Baiduspider
Disallow:                                                               /

User-agent:                                                             *
Crawl-delay: 15
```

Podemos ver que para las herramientas `Sosospider`, `Yandex` y `Baiduspider` está bloqueado todo el contenido, pero dado que no usamos ninguna de ellas no nos suponía un problema. La última instrucciónn indica que para cualquier *"user-agent"* se aplica la política `Crawl-delay` con un valor de 15 segundos.

Ninguna de estas restricciones afectaba al desarrollo de nuestro código *Python* por lo que comenzamos con su desarrollo. Importamos la librería *BeautifulSoup* y probamos a obtener todo el código HTML de la página. Después de revisar que los datos que queríamos *scrapear* se encontraban en el HTML ya comenzamos a utilizar las funciones de esta librería como `find()` o `find_all()`. Cuando ya teníamos todos los datos de los partidos, decidimos obtener el segundo dataset para las estadísticas individuales de los jugadores. Finalmente realizamos una refactorización del código para pasar de un paradigma de programación estructurada a un paradigma de programación orientada a objetos con la creación de nuestra propia clase `EuroligaScraper`.

Bibliografía consultada

1. Lawson, R. (2015). Web Scrapping with Python. Packt Publishing Ltd. Chapter 2. Scraping the data.
2. Mitchel, R. (2015). Web Scrapping with Python: Collecting data from the moder web. O'Reilly Media, Inc Chapter 1. Your first web scrapper.