

Práctica 1 – Web Scraping

Contenido

Contexto	2
Definir un título para el dataset	2
Descripción del dataset	2
Representación gráfica	3
Contenido	4
Descripción de los campos	4
Obtención de los datos	5
Período de tiempo de los datos	5
Agradecimientos	5
Inspiración	5
Licencia	6
Código	6
Dataset	6
Tabla de contribuciones	6
Referencias	7

Contexto

La información con la que se ha construido el *dataset* se ha recolectado mediante técnicas de *web scraping* de la página web Basketball-Reference.com [1], un sitio web dedicado a la estadística e historia del baloncesto.

Se ha escogido esta referencia porque es una compilación de datos actualizada de numerosas ligas (vigentes e históricas), así como los equipos y jugadores que han participado en ellas. Otro de los motivos es por la facilidad de navegación a través del *web scraping*, ya que las tablas de datos poseen enlaces a más referencias dentro de la misma página. Es decir, cuando se nombra a un jugador como, por ejemplo MVP de una temporada, este aparece como enlace a su página de referencia. Esto hace posible la recopilación de datos personales de dicho jugador que no aparecían en la tabla origen. Podríamos decir que se asemeja a la navegación dentro de un modelo de datos relacional.

Definir un título para el dataset

Talento en la NBA: Jugadores más notables a lo largo de la historia del baloncesto americano.

Descripción del dataset

El conjunto de datos extraído contiene información, por temporada y liga, de los jugadores (y sus equipos) que destacaron más en distintos aspectos del juego (puntos, rebotes, asistencias, contribución, etc.), así como los premiados en las categorías MVP (*Most Valuable Player*) y ROTY (*Rookie of the Year*). A continuación se adjunta un resumen del *dataframe* correspondiente al *dataset* resultante (usando la función `summary` en R):

Season Length:83 Class :character Mode :character	League Length:83 Class :character Mode :character	Champion Length:83 Class :character Mode :character	MVP Length:83 Class :character Mode :character	MVP.Age Min. :21.00 1st Qu.:26.00 Median :28.00 Mean :27.81 3rd Qu.:30.00 Max. :36.00 NA's :9	MVP.Team Length:83 Class :character Mode :character	MVP.Country Length:83 Class :character Mode :character	ROTY Length:83 Class :character Mode :character
ROTY.Age Min. :20.00 1st Qu.:22.00 Median :23.00 Mean :22.67 3rd Qu.:23.00 Max. :26.00 NA's :2	ROTY.Team Length:83 Class :character Mode :character	ROTY.Country Length:83 Class :character Mode :character	Points.Leader Length:83 Class :character Mode :character	Points Min. :1007 1st Qu.:2189 Median :2376 Mean :2356 3rd Qu.:2554 Max. :4029	Pts.Age Min. :21.00 1st Qu.:24.50 Median :26.00 Mean :26.67 3rd Qu.:29.00 Max. :35.00	Pts.Team Length:83 Class :character Mode :character	Pts.Country Length:83 Class :character Mode :character
Rebounds.Leader Length:83 Class :character Mode :character	Rebounds Min. : 610 1st Qu.:1081 Median :1188 Mean :1288 3rd Qu.:1476 Max. :2149 NA's :4	Rb.Age Min. :21.00 1st Qu.:24.50 Median :27.00 Mean :27.33 3rd Qu.:29.50 Max. :37.00 NA's :4	Rb.Team Length:83 Class :character Mode :character	Rb.Country Length:83 Class :character Mode :character	Assists.Leader Length:83 Class :character Mode :character	Assists Min. : 120.0 1st Qu.: 655.5 Median : 766.0 Mean : 757.6 3rd Qu.: 888.0 Max. :1164.0	Ast.Age Min. :23.00 1st Qu.:26.00 Median :28.00 Mean :28.54 3rd Qu.:31.00 Max. :38.00
Ast.Team Length:83 Class :character Mode :character	Ast.Country Length:83 Class :character Mode :character	Win.Shares.Leader Length:83 Class :character Mode :character	Win.Shares Min. : 9.00 1st Qu.:15.00 Median :17.00 Mean :17.01 3rd Qu.:19.00 Max. :25.00	WS.Age Min. :21.00 1st Qu.:26.00 Median :27.00 Mean :27.65 3rd Qu.:29.00 Max. :36.00	WS.Team Length:83 Class :character Mode :character	WS.Country Length:83 Class :character Mode :character	

Representación gráfica

Las siguientes imágenes representan, de manera gráfica, el proceso de obtención del *dataset*:

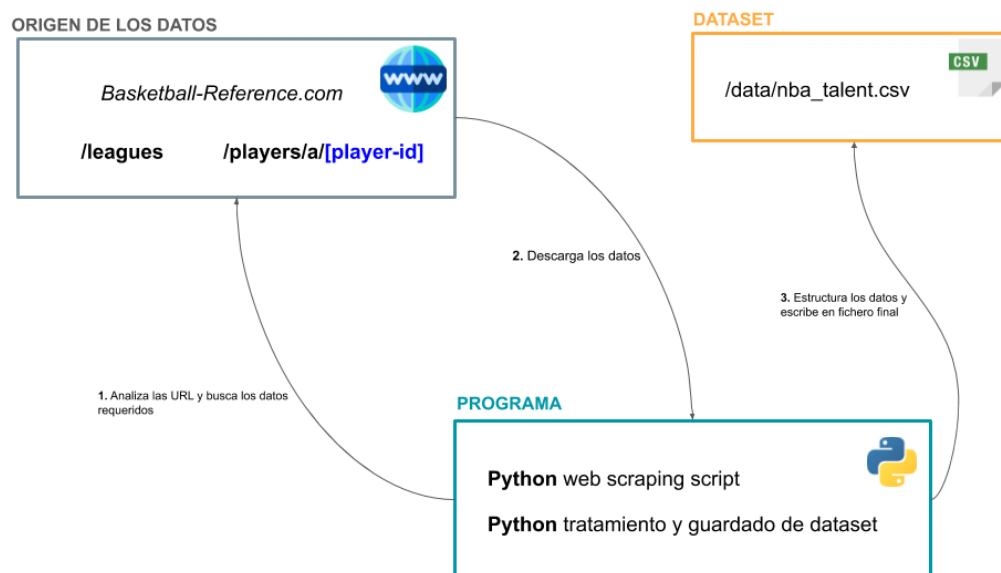


Ilustración 1: Vista general del proceso de obtención del dataset

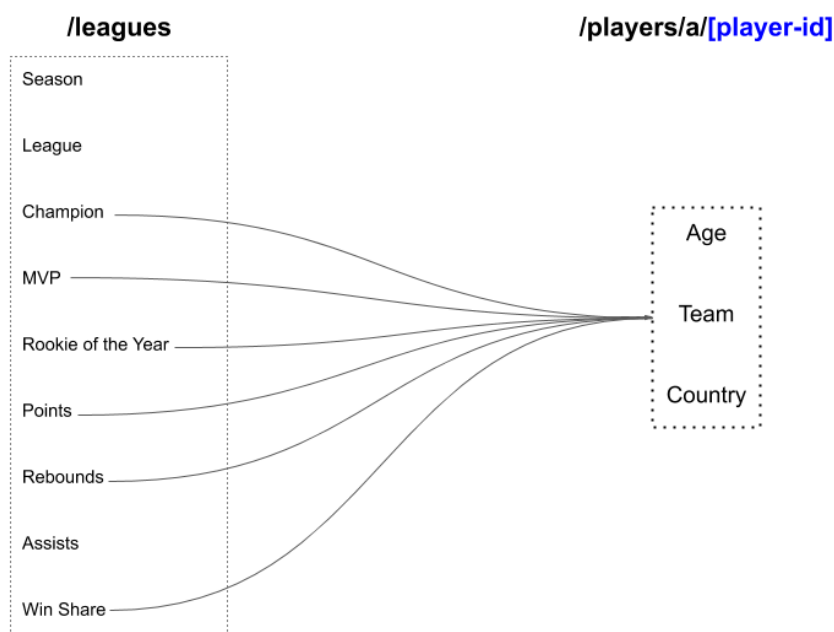


Ilustración 2: Vista general de la construcción de los campos del dataset

Contenido

Descripción de los campos

- **Season:** temporada a la que hace referencia el registro, en el formato YYYY-YY.
- **League:** liga, pudiendo ser NBA, ABA o BAA
- **Champion:** equipo que quedó campeón
- **MVP:** nombre del jugador mejor valorado de la temporada regular
- **MVP Age:** edad del jugador mejor valorado de la temporada regular (en el momento de obtener el galardón)
- **MVP Team:** equipo del jugador mejor valorado de la temporada regular
- **MVP Country:** país de origen del jugador mejor valorado de la temporada regular
- **ROTY:** nombre del novato¹ del año
- **ROTY Age:** edad del novato del año (en el momento de obtener el galardón)
- **ROTY Team:** equipo del novato del año
- **ROTY Country:** país de origen del novato del año
- **Points Leader:** nombre del máximo anotador de la temporada regular
- **Points:** puntos anotados por el máximo anotador de la temporada regular
- **Pts Age:** edad del máximo anotador de la temporada regular (en el momento de obtener el galardón)
- **Pts Team:** equipo del máximo anotador de la temporada regular
- **Pts Country:** país de origen del máximo anotador de la temporada regular
- **Rebounds Leader:** nombre del máximo reboteador de la temporada regular
- **Rebounds:** número de rebotes totales del máximo reboteador de la temporada regular
- **Rb Age:** edad del máximo reboteador de la temporada regular (en el momento de obtener el galardón)
- **Rb Team:** equipo del máximo reboteador de la temporada regular
- **Rb Country:** país de origen del máximo reboteador de la temporada regular
- **Assists Leader:** nombre del jugador líder en asistencias de la temporada regular
- **Assists:** número de asistencias del jugador líder en asistencias de la temporada regular
- **Ast Age:** edad del jugador líder en asistencias de la temporada regular (en el momento de obtener el galardón)
- **Ast Team:** equipo del jugador líder en asistencias de la temporada regular
- **Ast Country:** país de origen del jugador líder en asistencias de la temporada regular
- **Win Shares Leader:** nombre del jugador con mayor *Win Share* de la temporada regular
- **Win Shares:** valor del jugador con mayor *Win Share* de la temporada regular
- **WS Age:** edad del jugador con mayor *Win Share* de la temporada regular (en el momento de obtener el galardón)
- **WS Team:** equipo del jugador con mayor *Win Share* de la temporada regular
- **WS Country:** país de origen del jugador con mayor *Win Share* de la temporada regular

¹ Jugador en sus dos primeros años en la liga

Obtención de los datos

Los datos que se recogen en los campos descritos anteriormente se han recopilado mediante *web scraping* de la página Basketball-Reference.com. Se han utilizado dos de sus rutas o *endpoints* para combinar los datos y enriquecer el conjunto. De [NBA & ABA League Index](#) se ha obtenido la información básica referente a cada temporada y los jugadores mejor valorados, así como los ganadores de los galardones de la temporada. Esta comprende los campos *Season, League, Champion, MVP, ROTY, Points Leader, Points, Rebounds Leader, Rebounds, Assists Leader, Assists, Win Shares Leader, Win Shares*. De la página de cada jugador ([ejemplo](#)) se han extraído los datos para los campos restantes (o los necesarios para realizar los cálculos). Estos campos son *MVP Age, MVP Team, MVP Country, ROTY Age, ROTY Team, ROTY Country, Pts Age, Pts Team, Pts Country, Rb Age, Rb Team, Rb Country, Ast Age, Ast Team, Ast Country, WS Age, WS Team, WS Country*.

Período de tiempo de los datos

El *dataset* recopilado se compone de datos referentes a las últimas 73 temporadas de las principales ligas de baloncesto estadounidenses. Este intervalo va desde la temporada 1946-1947 hasta la terminada este octubre, la 2019-2020.

Agradecimientos

Como se ha indicado anteriormente, la totalidad de los datos extraídos para la construcción del *dataset* se ha extraído combinando varios recursos de la página web de estadística e historia del baloncesto Basketball-Reference.com. En esta se pueden encontrar prácticamente todos los datos y estadísticas relativos a los jugadores y equipos (activos e inactivos) de las ligas de baloncesto estadounidenses: datos personales, estadísticas, récords personales, récords de equipo, trayectorias profesionales, etc.

Inspiración

Este conjunto de datos es interesante por varios motivos. El primero es que no existía previamente como tal, sino que es la combinación de varios orígenes de datos que pueden ayudar a transmitir más información de manera conjunta que por separado. El segundo es la cantidad de preguntas que se pueden plantear (y responder) gracias al *dataset*. Al contener datos de los jugadores más notables de cada temporada en las principales ligas de baloncesto de EE.UU., podemos ver cómo ha variado la edad en la que se consiguen dichos méritos, la distribución de los países de origen, en qué equipos suelen destacar más estos jugadores, cuáles han conseguido destacar en más categorías diferentes, hay o ha habido algún período dominado por algún jugador/equipo en concreto... Y muchas más. Sobre todo puede servir también para contrastar algunas creencias populares con datos que las avalen o contradigan. Una de estas creencias podría ser, por ejemplo, si hay motivos, basándonos en estos datos, para que Kobe Bryant esté en discusiones sobre el *GOAT*² (*Greatest Of All Times*). En menor medida, se podría ver también si ha habido cambios en la normativa que han provocado cambios en las magnitudes o patrones de los puntos, rebotes o asistencias.

² GOAT, término atribuido al mejor jugador de un deporte en concreto, independientemente de su período de actividad.

Licencia

La licencia elegida para el *dataset* es **Released Under CC0: Public Domain License**. De esta manera se permite a cualquier persona “copiar, modificar, distribuir e interpretar la obra, incluso para propósitos comerciales, sin pedir permiso” [2]. Se ha elegido esta licencia porque no soy propietario de los datos extraídos (ya son de dominio público, simplemente se han compilado en un formato más usable) y porque considero que serán más útiles de esta manera. Cualquier persona podría realizar sus experimentos de *data science* con este sencillo *dataset* sin tener la necesidad de volver a hacer *web scraping*.




Código

El código mediante el que se ha generado el conjunto de datos se encuentra en [este repositorio](#) y, además, se adjunta junto a este documento. También en el repositorio se incluye el *dataset* actualizado.

Dataset

El dataset ha sido publicado en Zenodo y le corresponde el **DOI 10.5281/zenodo.4261333**. Puede ser consultado y descargado [aquí](#).

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	 Daniel Sarmiento Rocha
Redacción de las respuestas	 Daniel Sarmiento Rocha
Desarrollo código	 Daniel Sarmiento Rocha

Referencias

- [1] Sports Reference LLC, «Basketball-Reference.com - Basketball Statistics and History,» 2020. [En línea]. Available: <https://www.basketball-reference.com>. [Último acceso: Octubre 2020].
- [2] Creative Commons, «CC0 1.0 Universal (CC0 1.0),» 2020. [En línea]. Available: <https://creativecommons.org/publicdomain/zero/1.0/deed.es>. [Último acceso: Noviembre 2020].