Pràctica 1. Tipologia i cicle de vida de les dades

Autors:

- Álvaro Díaz
- David Leiva

Respostes:

1 Context

Per tal de poder estudiar les trajectòries dels jugadors de bàsquet així com la composició de cada equip que participa a la Euroleague a la temporada 2020/21, hem decidit extreure la informació de la web oficial d'aquesta competició (https://www.euroleague.net/). Aquesta web proporciona totes les estadístiques tant del jugadors, com dels equips així com del diferents partits jugats.

Aquesta web no disposa d'una API ni es pot descarregar fitxers amb aquesta informació, pel que per disposar de les dades de la màxima competició europea de bàsquet s'ha de recórrer a fer web scrapping.

2 Títol del dataset

Es generen dos datasets diferents:

- **euroleaguePlayers_season**
- **euroleaguePlayers_average**

3 Descripció del dataset

A ambdós datasets s'extreu la informació de les estadístiques dels jugadors en actiu la temporada 2020/21 a l'Euroleague de Basket.

- **euroleaguePlayers_season**: S'extreu les estadístiques de cada temporada que ha participat en l'Euroleague cada jugador.

- **euroleaguePlayers_average:** S'extreu les dades mitjanes proporcionades per a cada jugador.

4 Representació gràfica

5 Contingut

Los datasets conté la informació dels jugadors en actiu de la temporada actual i de les temporades anteriors jugades per cada jugador. Les dades disponibles son de la temporada 01-02 fins a la 20-21. .

L'arxiu robots.txt permet l'ús de la gran majoria de rastrejadors excepte Sosopider, Yandex i Baiduspider. La resta *spiders* han de tenir un *delay* de 15 segons entre consulta i consulta.

El camps que contenen els respectius datasets són:

- **euroleaguePlayers_season**
  - **Season:** Temporada
  - **Team:** Equip
  - **G:** Partits jugats.
  - **Pts:** Punts anotats.
  - **Avg:** Mitjana de punts per partit.
  - **2FG:** Tirs de 2 punts en format "encerts/llançaments"
  - **2FG_%:** Percentatge de tirs de 2 punts anotats
  - **3FG:** Tirs de 3 punts en format "encerts/llançaments"
  - **3FG_%:** Percentatge de tirs de 3 punts anotats
  - **FT:** Tirs lliures en format "encerts/llançaments"
  - **FT_%:** Percentatge de tirs lliures anotats
  - **Reb:** Rebots totals per temporada
  - **St:** Robatoris totals per temporada (steals)

    - **As:** Assistències totals per temporada
    - **Bl:** Taps totals per temporada
    - **Full_name:** Nom complet del jugador en format “Cognom, Nom”

- **euroleaguePlayers_average** (s’omet la descripció de les variables comunes als dos datasets)
    - **Full_name**
    - **Surname:** Cognom del jugador
    - **Name:** Nom del jugador
    - **Club:** Equip on juga la temporada actual
    - **Dorsal:** Número que porta el jugador
    - **Position:** Posició que ocupa sent [“Guard”, Center”,”Forward”]
    - **Height:** Altura en metres
    - **Born:** Data de naixement en format “dd Month, Year”
    - **Nationality:** Nacionalidad
    - **G**
    - **Pts**
    - **Avg**
    - **2FG**
    - **2FG_%**
    - **3FG**
    - **3FG_%**
    - **FT**
    - **FT_%**
    - **Reb** Mitjana de rebots per partit de totes les temporades.
    - **St** Mitjana de robatoris per partit de totes les temporades.
    - **As** Mitjana d’assistències per partit de totes les temporades.
    - **Bl** Mitjana de taps per partit de totes les temporades.

6 Agraïments

Agrair la informació al propietari de les dades, el propietari del domini ‘EUROLEAGUE.NET’ , l’organització ‘Euroleague Properties S.A’ amb seu a Luxemburg.

Altres estudis utilitzant les dades de la web és:

- Vinué, Guillermo. “A Web Application for Interactive Visualization of European Basketball Data.” *Big Data*, vol. 8, no. 1, Jan. 2020, pp. 70–86. *liebertpub.com (Atypon)*, doi:10.1089/big.2018.0124.

- Horvat, Tomislav, et al. “Prediction of Euroleague Games Based on Supervised Classification Algorithm K-Nearest Neighbours:” *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support*, SCITEPRESS - Science and Technology Publications, 2018, pp. 203–07. *DOI.org (Crossref)*, doi:10.5220/0006893502030207.

7 Inspiració

L’estadística cada dia té més influència en el món de l’esport en general i al bàsquet en concret. Cada dia es prenen més decisions basades en les dades i n’hi ha més persones treballant en els departaments d’analítica, sobretot a EEUU. A Europa aquesta transformació va més lenta però serà imprescindible i marcarà diferències de rendiment.

Les dades que s’han extret possibiliten un estudi de rendiment dels jugadors en la seva carrera podent determinar en quins equips ha sigut més favorable, determinar si està en el pic de forma o potser aquestes dades són any rere any menys atractives.

Es poden fer rànquings estadístics per edat per tal de poder veure la tendència dels jugadors joves en els següents anys de la seva carrera si evolucionen com ho han fet els altres jugadors en temporades anteriors.

També es poden fer agregacions estadístiques per equips i poder comparar la tendència entre els diferents clubs. Aquestes dades donen una visió general de cada equipo o jugador per temporada i creiem que és un bon punt de partida per a realitzar estudis més exhaustius.

La premsa pot fer servir aquestes dades per a fer reportatges, ja que són dades molt entenedores per al públic general.

Per últim, cal destacar que aquestes dades només són la punta de l’iceberg del conjunt total de dades que es tindran en els pròxims 5-10 anys a Europa. En poc temps arribarà el reconeixement per imatge de les accions dels jugadors a entrenaments i partits i la quantitat de dades es disparará. Aquestes dades ja les disposen a l’NBA i les possibilitats d’extreure coneixements es multipliquen o s’eleven a l’exponent.

8 Llicència

Les directrius que hem decidit que volem per les dades generades son:

- Que siguin obertes, amb llibertat per ser usades sense gaires limitacions, tant per copiar-les com per transformar-les o ser usades per construir nou material. Estem d'acord que no ens importa que el nostre treball sigui usat amb fins comercials, així que no hem limitar aquesta possibilitat.
- Volíem que si les dades son usades, s'ha de fer referència a la nostra autoria.
- Si es fa un remix, transforma o construeixen noves dades a partir de les originals, s'ha de mantenir la mateixa llicència.

Entre les diferents llicències existents, la que s'adapta als nostres requisits és ***Released Under CC BY-SA 4.0 License***

Amb aquesta llicència, a més, ens assegurem que si les dades són emprades per construir un altre material, aquest haurà de mantenir el mateix tipus de llicència.

9 Codi

El codi el podeu trobar clickant al link

10 Dataset

Les dades es poden trobar publicades a Zenodo clickan al link

La llicència és la mateixa que la del repositori de Github.