PRÀCTICA 1

WEB SCRAPING

L'objectiu d'aquesta activitat serà la creació d'un data set a partir de les dades contingudes al web. Heu d'indicar les següents característiques del data set general:

1. **Títol del data set. Cal que poseu un títol que sigui descriptiu.**

Estadístiques jugadors NBA.

2. **Subtítol del data set. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.**

Estadístiques per temporades de tots els partits de jugadors en actiu i retirats.

3. **Imatge. Agregueu una imatge que identifiqui l vostre data set visualment.**

4. **Context. Quina és la matèria del conjunt de dades?**

La pàgina web que analitzaré https://www.basketball-reference.com és una àmplia base de dades relacionada amb la lliga de bàsquet professional dels EEUU: NBA. L'estudi actual es centrarà en una petita part d'aquest bast repositori de dades ja que únicament s'analitzaran les estadístiques dels els jugadors; tant els que estan en actiu com els que ja estan retirats.

5. **Contingut. Quins camps inclou? Quin és el període de temps de les dades i com s'ha escollit?**

Donat que el número de jugadors és molt gran he considerat que l'usuari que executi l'scraper pugui triar quin **jugador** vol estudiar (per exemple: Pau Gasol). A més, també haurà d'escollir la **temporada** en que es centrarà l'anàlisi (per exemple: 2018). Així doncs, a partir de dues variables d'entrada, es realitzarà l'scraper.

Estarem recollint dades a nivell de temporada regular de la NBA (no s'inclouen els playoff). Cada partit s'emmagatzemarà en una línia de l'arxiu csv i els seus camps contindran la següent informació:

- Date: Data en que es va jugar el partir.
- Opp: L'oponent al que es va enfrontar el jugador.
- W/L: El resultat del partit en forma binària: L (perdut) o W (guanyat).
- Time: Minuts que ha disputat el jugador.
- 2P: Tirs de 3 punts encertats.
- 3P: Tirs de 2 punts encertats.
- 1P: Tirs lliures encertats.
- Off Reb: Rebots ofensius.
- Def Reb: Rebots defensius.
- Assist: Assistències realitzades.
- Steal: Pilotes robades.
- Foul: Faltes realitzades als oponents.
- Points: Total de punts realitzats.
- Game: ScoreValoració final del jugador.

Comentar que en el supòsit de triar la temporada actual (2019), l'arxiu generat tindrà tantes línies com partits hagi jugat aquell jugador fins al moment.

Un exemple d'execució seria el següent:

```
Please enter player name (i.e. Pau Gasol): LeBron James
Please enter season (i.e. 2018): 2017

  Executing scrapper

Dataset LeBron_James_season_2017.csv created ! !
```

6. Agraïments. Qui és el propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.

Les dades, són extretes de Sports Reference LLC, propietària d'un conjunt de webs amb estadístiques de baseball, soccer, football, hockeys,... i bàsquet. Tot i que la web es nodreix den el nostre cas de la lliga professional NBA, el propietari de les dades és Sports Reference LLC.

Agraïments a Sport Reference LLC i al seu portal https://www.basketball-reference.com per mostrar les seves dades sense cap ànim de lucre d'una forma bàsica i de fàcil comprensió.

7. Inspiració. Per què és interessant aquest conjunt de dades? Quines preguntes li agradaria respondre la comunitat?

La motivació principal a l'hora de realitzar aquest treball ha sigut l'anàlisi individual de cada jugador per temporada. Aquest fet pot respondre preguntes a àmbits professionals com el propi de la NBA o un directament periodístic o informatiu.

Podem imaginar la figura del manager d'un dels equips de la NBA, el qual ha d'analitzar la temporada que està realitzant un dels seus jugadors; o bé obtenir informació d'altres jugadors per tal d'estudiar futurs enfrontaments o afrontar futurs fitxatges.

Per una altra banda, cal tenir en compte que també es pot recollir informació de jugadors ja retirats. Aquesta pot ser una eina interessant per a l'àmbit periodístic o d'investigació.

8. **Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir perquè l'heu seleccionada:**
    - **Released Under CC0: Public Domain License.**
    - **Released Under CC BY-NC-SA 4.0 License.**
    - **Released Under CC BY-SA 4.0 License.**
    - **Database released under Ipen Database License, individual contents under Database Contents License.**
    - **Other (specified above).**
    - **Unknown License.**

En el cas de posar el codi en producció i generar datasets, considero que una llicència Creative Commons BY-SA 4.0 seria la més adequada, ja que el fet de poder extreure'n un benefici econòmic, facilitaria el seu us en àmbits com el periodístic o en l'intercanvi d'informació empresa - club sobre jugadors.

9. **Codi: Cal adjuntar el codi amb el que heu generat el data set, preferiblement amb R o Python, que us ha ajudat a generar el data set.**

Com es pot observar s'han generat dos arxius:

- Script en python scraper_nba_players.py
- Arxiu Notebook de python3 scraper_nba_players.ipynb

El script scraper_nba_players.py és l'arxiu a executat. El notebook s'ha pujat a nivell informatiu ja que ha sigut l'eina a l'hora de crear el codi.

El codi consta de 5 grans blocs els quals s'explicaran breument a continuació.

1- Importació de les llibreries necessàries.

S'importen les llibreries que s'utilitzaran.

2- Funció de creació del link a analitzar.

Es parteix d'una estructura comuna:
https://www.basketball-reference.com/players
per a finalment obtenir un link de l'estil:
https://www.basketball-reference.com/players/g/gasolpa01/gamelog/2004/

Els jugadors es classifiquen, **generalment**, a partir de la inicial del seu cognom; per tant, es recorren tots els links de la lletra 'a' a la 'z'. En cada un d'aquests 26 links, es compara el nom del jugador introduït amb tots els jugadors que es troben al web. Si es troba s'afegeix al link inicial la lletra + el link del jugador (sense l'extensió html). Finalment s'acaba afegint el terme /gamelog/'temporada introduïda per l'usuari'.

3- Funció per a extreure la informació del jugador.

Es crea un llista amb el nom dels 14 camps que s'hauran d'extreure. Donat que el web conté la informació estructurada en forma de taula, s'emmagatzemen els valors a partir de 14 cel·les concretes.

4- Funció per a crear el dataset CSV.

La llista creada es converteix en un arxiu csv separat per ';'. El nom de l'arxiu es crea a partir del nom del jugador i la temporada.

5- Sol·licitud de les variables d'entrada: jugador i temporada.

El codi demana el nom del jugador i la temporada que es vol analitzar. A patir d'aquí es criden les tres funcions explicades. El codi és capaç de retornar dos tipus d'errors: en cas de no trobar el jugador i en cas de no trobar la temporada.

8. Dataset: Dataset en format CSV

Donat que el dataset es genera a partir de dues variables d'entrada (jugador i temporada), les opcions a l'hora de crear un dataset són múltiples. A mode d'exemple pujaré tres datasets corresponents a una temporada actual, un jugador retirat, i un jugador en actiu amb informació de temporades anteriors:

- **Pau_Gasol_season_2019.csv**
  Variable jugador: Pau Gasol
  Variable temporada: 2019
- **Michael_Jordan_season_1995.csv**
  Variable jugador: Michael Jordan
  Variable temporada: 1985
- **LeBron_James_season_2017.csv**
  Variable jugador: LeBron James
  Variable temporada: 2017