Preguntes:

1.- El nostre projecte de web scrapping recull la informació de la taula de classificació de la NBA. Coincidint l'inici de la pràctica amb l'inici de la NBA vam trobar molt interessant aquesta informació per fer-la servir ja que en les primeres jornades és un constant puja baixa dels equips en la classificació per no haver jugat un gran número de partits i això ens permetria veure canvis en la informació extreta. La web triada ha sigut https://resultados.as.com/resultados/baloncesto/nba/clasificacion/ . Es troba dins de la secció de la NBA de la pàgina del diari esportiu AS.

2.- NBA\_classification

3.- Bàsicament el dataset és una extracció de la taula de la classificació de la NBA com la que podem trobar a la web en la que recollim la informació. Trobem els següents camps: Division, Position, Team, GamesPlayed, Won, Lost, PointsAgainst, PointsForth, DiffPoints. A més a més hem afegit les columnes Date i Time per saber en quin moment es va fer l'extracció de la informació que tenim i poder fer un anàlisi amb les demes extraccions fetes en altres moments.

4.- **TO DO: (FER UN ESQUEMA/ DIAGRAMA, HO PARLEM! -> Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit)**

5.- En la següent taula es descriu amb més detall cada camp:

Name Type Description

Date date Data del moment en que s'ha fet l'extracció

Time time Hora del moment en que s'ha fet l'extracció

Division text Divisió a la que pertany l'equip

Position number Posició de l'equip a la divisió

Team text Nom de l'equip

GamesPlayed number Partits jugats

Won number Partits guanyats

Lost number Partits perduts

PointsAgaints number Punts anotats per l'equip

PointsForth number Punts rebuts a l'equip

DiffPoints number Diferencia de punts (punts anotats menys els punts rebuts)

Respecte la periocitat en la que hem extret les dades ha sigut diaria, aixi doncs des del dia 23 d'octubre s'ha realitzat una extracció diaria en la que podem veure reflexada els resultats de la jornada de la nit anterior. Per altra banda, vam trobar molt interessant durant una nit fer una extracció cada hora ja que durant una mateixa nit poden haver-hi molts partits i a més a més Estats Units és un pais que té 6 franjes horaries. Aquesta extracció amb una periocitat més alta la vam fer durant la nit del 26 al 27 d'octubre.

6.- Un cop vam tenir decidit que volíem basar la pràctica en el context esportiu, ens vam decantar per la classificació de la NBA. Primer de tot vam voler extreure les dades de la pàgina oficial de la competició, però quan vam estreure el codi html amb Beautiful Soup vam veure que era molt complicat poder extreure’n la informació de manera automàtica amb python per la complexitat del codi html. Vam buscar alternatives i la que ens encaixava més era la del diari AS. Així doncs, les dades han sigut extretes de la pàgina del diari esportiu AS, dins de la secció de NBA.

No ens hem basat en altres projectes similiars per desenvolupar el codi Python, hem agafat de referència les pràctiques d’exemple que se’ns proporcionaven, les explicacions en la documentació teòrica de l’assignatura i casos molt concrets trobats a Internet.

Abans d’extreure les dades vam inspeccionar el fitxer robots.txt de la pàgina per assegurar que teníem permís per accedir i extreure l’informació de la pràctica.

Graphical user interface, text

Description automatically generated

Figura 1 Fitxer robots.txt de la pàgina as.com

Com podem veure a la figura 1, de totes les pàgines que no es permet l’ús de bots no apareix la que hem utilitzat nosaltres.

7.- A l’hora de poder analitzar aquestes dades, creiem que es pot enfocar desde dos punts de vista diferents: Un podria ser el macroscòpic que podria ser analitzar com la taula de classificació va canviant a mesura que van pasar les jornades. I desde un punt de vista microscòpic es podria fer un anàlisis de com evoluciona un determinat equip en la taula de classificació durant el transcurs de les jonades.

Algunes de les preguntes que podríem respondre amb aquestes dades serien les següents entre d’altres:

* L’equip que ha pujat/baixat més en la taula després d’una jornada o interval de temps.
* La mitja de punts en una jornada de tots els equips.
* Ratxa més gran de victories o derrotes i l’equip d’aquesta.

8.- Creiem que la llicència que s’adapta més al nostre dataset és la llicència Released Under CC0: Public Domain License. Al final l’extracció de les dades no l’hem fet amb fins de recreació econòmica, totalment al contrari, si algu vol utilitzar el codi o les dades per el seu ús particular ho podrà fer perquè aquesta llicència categoritza com a Open Source el dataset, és a dir, d’ús públic.

9.- El codi de l’aplicació el trobem a la carpeta src del repositori de github. En el fitxer readme.md s’explica breument que és cada fitxer.

10.- En el següent enllaç de Zenodo és troba publicat el dataset:

**TODO: pujat fitxers data a Zenodo**

Taula de contribucions

|  |  |
| --- | --- |
| Contribucions | Signatura |
| Investigació prèvia | OMM, GRR |
| Redacció de les respostes | OMM, GRR |
| Desenvolupament del codi | OMM, GRR |