

1. Context. Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.

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. Títol. Definir un títol que sigui descriptiu pel dataset.

NBA_classification

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

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. Representació gràfica. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

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 per un equip contrari
DiffPoints	number	Diferència de punts (punts anotats menys punts rebuts)

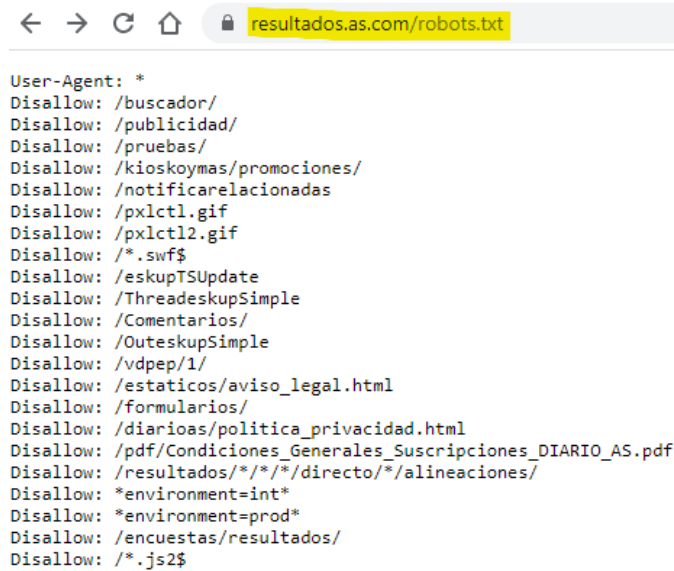
Respecte la perioditzat en la que hem extret les dades ha sigut diària, així doncs des del dia 23 d'octubre s'ha realitzat una extracció diària en la que podem veure reflexada els resultats de la jornada de la nit anterior. Per altra banda, vam trobar molt interessant durant una nit fer una extracció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

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

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 extreure el codi html amb BeautifulSoup 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 similars 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 la informació de la pràctica.



```
User-Agent: *
Disallow: /buscador/
Disallow: /publicidad/
Disallow: /pruebas/
Disallow: /kioskoymas/promociones/
Disallow: /notificarelacionadas
Disallow: /pxlctl.gif
Disallow: /pxlctl2.gif
Disallow: /*.swf$
Disallow: /eskupTSUpdate
Disallow: /ThreadeskupSimple
Disallow: /Comentarios/
Disallow: /OuteskupSimple
Disallow: /vdpep/1/
Disallow: /estaticos/aviso_legal.html
Disallow: /formularios/
Disallow: /diarioas/politica_privacidad.html
Disallow: /pdf/Condiciones_Generales_Suscripciones_DIARIO_AS.pdf
Disallow: /resultados/*/*/*directo/*/alineaciones/
Disallow: *environment=int*
Disallow: *environment=prod*
Disallow: /encuestas/resultados/
Disallow: /*.js2$
```

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. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

A l'hora de poder analitzar aquestes dades, creiem que es pot enfocar des de 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 passar les jornades. Per altra banda, des de un punt de vista microscòpic es podria fer un anàlisi de com evoluciona un determinat equip en la taula de classificació durant el transcurs de les jornades.

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 victòries o derrotes i l'equip d'aquesta.

8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:

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 algú 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. Codi. Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

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. Dataset. Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

En el següent enllaç de Zenodo és troba publicat el dataset: <https://doi.org/10.5281/zenodo.5627357>

Taula de contribucions

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