

M2.951 - Pràctica 1: Web scraping

0. Introducció.

Aquest repositori conté la solució de la **Pràctica 1: Web scraping** de l'assignatura M2.951 - Tipologia i cicle de vida de les dades del Màster Universitari en Ciència de Dades de la Universitat Oberta de Catalunya.

El codi que conté aquest repositori ha estat desenvolupat pels alumnes **Alexandre Vidal de Palol** i **Adrián Alonso Gonzalo** (alexvidi i idriskameni en GitHub respectivament).

El codi que conté el repositori té com a objectiu descarregar una sèrie de 6 conjunts de dades (o *datasets*) de la pàgina oficial de la *National Basketball Association*, també coneguda com a *NBA*. Dos dels conjunts de dades contenen dades sobre la classificació (o *standings*) dels equips de les conferències est i oest en la temporada actual. Altres dos conjunts de dades contenen el llistat de jugadors que cadascun d'aquests equips té, així com algunes estadístiques biogràfiques d'aquests. I els dos datasets restants mostren una taula amb informació sobre estadístiques dels partits de cada jugador en cada conferència.

1. Context.

Els alumnes involucrats en la creació d'aquest repositori han recollit aquesta informació amb la intenció de mostrar com es podrien recollir una sèrie de conjunts de dades relacionats amb la *NBA* per després poder crear un **model de dades del tipus entitat-relació** que posteriorment pugui **ser explotat per analistes de dades** als quals els interessi fer anàlisis sobre aquest camp.

Tot i que en aquest repositori només ens hem centrat en els *standings* i en la llista de jugadors per equip, la web de la *NBA* també conté altres dades interessants com: estadístiques més detallades de cada jugador o resultats de partits de temporades passades. Seguint el mateix mètode aplicat en l'extracció dels conjunts proporcionats en aquest repositori, els interessats podrien seguir descarregant diferents conjunts de dades per completar i finalitzar el model de dades (nosaltres hem fet tres exemples com a demostració que això és possible).

La creació d'un model de dades del tipus entitat-relació permetria als analistes de dades d'empreses relacionades amb l'*NBA* prendre decisions basades en dades. Aquest model podria ajudar a identificar els millors jugadors per equips, els equips que poden tenir opcions a guanyar la lliga, etc.

En resum, el repositori proporciona conjunts de dades que tenen la intenció de ser útils en la possible creació d'un model de dades del tipus entitat-relació per la posterior explotació de les mateixes per l'extracció de coneixement (basat en dades) en l'àmbit de l'*NBA*.

2. Títols dels conjunts de dades.

En el repositori s'extreuen 6 conjunts de dades:

2.1. `teams_standing_eastern_conference/teams_standing_western_conference`

- **web-scraping-**
nba/docs/teams_standings/[teams standing eastern conference.csv](#): classificació dels equips de la conferència 'EAST' de l'*NBA* en el moment de l'execució del fitxer 'web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py'.
- **web-scraping-**
nba/docs/teams_standings/[teams standing western conference.csv](#): classificació dels equips de la conferència 'WEST' de l'*NBA* en el moment de l'execució del fitxer 'web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py'.

2.2. `players_stats_eastern_conference` i `players_stats_western_conference`

- **web-scraping-**
nba/docs/players_stats/[players stats eastern conference.csv](#): llistat i estadístiques de la temporada dels jugadors de cada equip de la conferència 'EAST' de l'*NBA* en el moment de l'execució del fitxer 'web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py'.
- **web-scraping-**
nba/docs/players_stats/[players stats western conference.csv](#): llistat i estadístiques de la temporada dels jugadors de cada equip de la conferència 'WEST' de l'*NBA* en el moment de l'execució del fitxer 'web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py'.

2.3. `players_list_eastern_conference` i `players_list_western_conference`

- **web-scraping-**
`nba/docs/players_list/players_list_eastern_conference.csv`: Llistat i informació biogràfica dels jugadors de cada equip de la conferència 'EAST' i de l'*NBA* en el moment de l'execució del fitxer '`web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py`'.
- **web-scraping-**
`nba/docs/players_list/players_list_western_conference.csv`: Llistat i informació biogràfica dels jugadors de cada equip de la conferència 'WEST' de l'*NBA* en el moment de l'execució del fitxer '`web-scraping-nba/src/web-scraping-nba/web_scraping_nba.py`'.

3. Descripció del dataset.

3.1. `teams_standing_eastern_conference`/`teams_standing_western_conference`

- Els datasets mostren les dades de la classificació dels 15 equips de la conferència 'EAST' i 'WEST' (respectivament) de la primera lliga americana de bàsquet de l'*NBA*. Les taules estan ordenades des de l'equip que té la màxima puntuació fins al que té menys punts. Altres dades que mostren les taules són per exemple els partits guanyats a casa o a fora, entre d'altres.

3.2. `players_stats_eastern_conference` i `players_stats_western_conference`

- Els conjunts de dades llisten cadascun dels jugadors de cada equip de la conferència 'EAST' i 'WEST' (respectivament) de l'*NBA* on en les diverses columnes s'observen estadístiques tècniques de la temporada actual.

3.3. `players_list_eastern_conference` i `players_list_western_conference`

- Els conjunts de dades llisten cadascun dels jugadors de cada equip de la conferència 'EAST' i 'WEST' (respectivament) de l'*NBA* on en les diverses columnes mostren dades biogràfiques com altura, pes, data de naixement, entre d'altres.

4. Representació gràfica.

4.1.

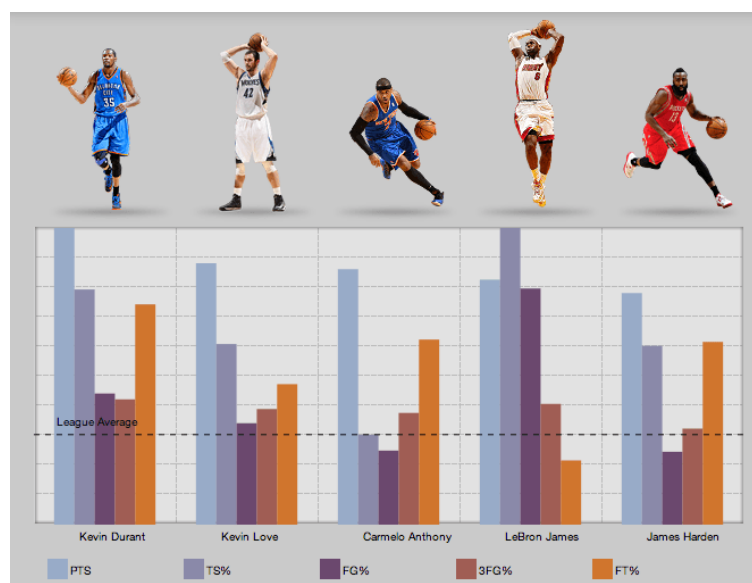
teams_standing_eastern_conference/teams_standing_western_conference

Exemple visual dels datasets `teams_standing_eastern_conference.csv` i `teams_standing_western_conference.csv` on mostra el mapa dels EEUU dividit entre les dues conferències de la lliga NBA amb la ubicació de cada equip




















4.2. players_stats_eastern_conference i players_stats_western_conference

Exemple gràfic dels datasets `players_stats_eastern_conference` i `players_stats_western_conference` que representa una comparativa sobre diverses variables estadístiques de diferents jugadors.



4.3. players_list_eastern_conference i players_list_western_conference

Exemple gràfic dels datasets players_list_eastern_conference i players_list_western_conference on es mostra informació visual sobre dades biogràfiques dels jugadors mostra informació visual sobre dades biogràfiques dels jugadors.

 Kostas Antetokounmpo Los Angeles Lakers #37 Forward 6' 10" Greece School: University of Dayton Born: Nov 20, 1997 Age: 22 2018 Draft: Round 2 Pick 5	 Avery Bradley Los Angeles Lakers #11 Guard 6' 3" United States School: Texas Born: Nov 26, 1990 Age: 29 2010 Draft: Round 1 Pick 19	 Devontae Cacok Los Angeles Lakers #12 Forward 6' 8" United States School: North Carolina-Wilm... Born: Oct 08, 1996 Age: 23 Undrafted	 Kentavious Caldwell-Poole Los Angeles Lakers #1 Guard 6' 5" United States School: Georgia Born: Feb 18, 1993 Age: 26 2013 Draft: Round 1 Pick 8	 Alex Caruso Los Angeles Lakers #4 Guard 6' 5" United States School: Texas A&M Born: Feb 28, 1994 Age: 25 Undrafted	 Quinn Cook Los Angeles Lakers #2 Guard 6' 1" United States School: Duke Born: Mar 23, 1993 Age: 26 Undrafted
 DeMarcus Cousins Los Angeles Lakers #15 Center 6' 10" United States School: Kentucky Born: Aug 13, 1990 Age: 29 2010 Draft: Round 1 Pick 5	 Troy Daniels Los Angeles Lakers #30 Guard 6' 4" United States School: Virginia Commonwe... Born: Jul 15, 1991 Age: 28 Undrafted	 Anthony Davis Los Angeles Lakers #3 Forward-Center 6' 10" United States School: Kentucky Born: Mar 11, 1993 Age: 26 2012 Draft: Round 1 Pick 1	 Jared Dudley Los Angeles Lakers #10 Forward 6' 6" United States School: Boston College Born: Jul 10, 1985 Age: 34 2007 Draft: Round 1 Pick 22	 Danny Green Los Angeles Lakers #14 Guard 6' 6" United States School: North Carolina Born: Jun 22, 1987 Age: 32 2009 Draft: Round 2 Pick 46	 Talen Horton-Tucker Los Angeles Lakers #5 Guard 6' 4" United States School: Iowa State Born: Nov 25, 2000 Age: 19 2019 Draft: Round 2 Pick 46
 Dwight Howard Los Angeles Lakers #39 Center-Forward 6' 10" United States School: SW Atlanta Christia... Born: Dec 08, 1985 Age: 34 2004 Draft: Round 1 Pick 1	 LeBron James Los Angeles Lakers #23 Forward 6' 9" United States School: St. Vincent-St. Mary... Born: Dec 30, 1984 Age: 35 2003 Draft: Round 1 Pick 1	 Kyle Kuzma Los Angeles Lakers #0 Forward 6' 8" United States School: Utah Born: Jul 24, 1995 Age: 24 2017 Draft: Round 1 Pick 27	 JaVale McGee Los Angeles Lakers #7 Center-Forward 7' 0" United States School: Nevada-Reno Born: Jan 19, 1988 Age: 31 2008 Draft: Round 1 Pick 18	 Rajon Rondo Los Angeles Lakers #9 Guard 6' 1" United States School: Kentucky Born: Feb 22, 1986 Age: 33 2006 Draft: Round 1 Pick 21	

5. Contingut.

5.1.teams_standing_eastern_conference/teams_standing_western_conference

Camps i descripció:

- Clasif - Numero en el ranking de la lliga
- Nombre - Nom de l'equip
- Nombre.1 - Abreviació nom de l'equip
- V - victories
- D - derrotes

- % - tant per cent de victòries sobre les derrotes
- DIF - diferència respecte a l'anterior
- CONF - valors respecte equips conferència
- DIV - valors estadístics
- Local - partits guanyats a casa
- Visitante - partits guanyats a fora
- 10 últimos - punts dels deu últims partits
- Racha - partits guanyats consecutius
- FP - valors estadístics
- PA - valors estadístics
- DIF.2 - diferència entre FP i PA
- Conference - a quina conferència pertany l'equip

Període de temps de les dades:

- Les dades són extretes de la temporada actual de la lliga de l'NBA 2021-2022. En concret des de l'inici de la temporada, l'octubre de 2021, fins a l'actualitat, abril de 2022.

5.2. players_stats_eastern_conference i players_stats_western_conference

Camps i descripció:

- B - punts totals anotats
- GS - jocs començats
- PPP - mitjana de punts per partit
- REPP - ratio de punts
- APP - ratio de passades per partit
- MPP - ratio assistències de punts per partit
- EFI - valor respecte a l'eficiència
- %TC - tant per cent tirs a canasta
- %3P - tant per cent de triples
- %TL - tant per cent de tirs lliure
- OFE - valor estadístic respecte ofensiva
- DEF - valor estadístic respecte a defensa

- ROPP - mitjana sobre valor estadístic entre ofensiva i defensiva
- TPP - mitjana estadística en quan a bloquejos
- PÉR - valors respecte torns de sortida al camp
- FP - mitjana de faltes per partit
- Team - Equip al qual juga
- Conference - Conferència que juga

Període de temps de les dades:

- Les dades són extretes de la temporada actual de la lliga de l'NBA 2021-2022. En concret des de l'inici de la temporada, l'octubre de 2021, fins a l'actualitat, abril de 2022.

5.3. players_list_eastern_conference i players_list_western_conference

Camps i descripció:

- Nombre - Nom del jugador
- Pos. - Posició que juga al camp
- Altura - Alçada del jugador
- Peso - Pes del jugador
- Numero - número de camiseta
- Fecha de nacimiento - Data de naixement
- Exp - Anys a l'NBA
- Antes de la NBA - Equip al qual jugava anterior a l'NBA
- País - Nacionalitat del jugador
- Team - Equip actual del jugador
- Conference - Conferència a la qual juga

Període de temps de les dades:

- Les dades són extretes de la temporada actual de la lliga de l'NBA 2021-2022. En concret des de l'inici de la temporada, l'octubre de 2021, fins a l'actualitat, abril de 2022.

Procés d'extracció de les dades

Per la gestió del procés en web scraping s'ha utilitzat la llibreria Selenium per automatitzar el navegador web, en concret Chrome, utilitzant com a enllaç de referència "<https://es.global.nba.com/statistics>" o també a la pestanya de "classificació" de la web principal "<https://www.sportingnews.com/es/nba?gr=www>". Tanmateix, amb la llibreria "BeautifulSoup" s'ha extret la informació del contingut en format HTML. Durant el procés d'extracció de les dades s'ha utilitzat la funció find_all bàsicament per buscar l'atribut (nba-stat-table) de tipus tag. Un cop identificada la taula hem fet un bucle per identificar les taules i seguidament transformar-les en un dataframe.

Un cop aconseguida els datasets de la classificació de les dues conferències hem repetit el procés per l'obtenció dels quatre datasets restants.

6. Agraïments.

El propietari del lloc web i del conjunt de dades de les quals s'ha dut a terme el projecte és l'empresa The Sporting News Holdings, que ha facilitat l'anàlisi i les estadístiques per al procés de web scraping i seguidament a l'obtenció dels datasets. El propietari dona com a contacte el seu mail info@sportingnewsholdings.com, juntament amb altres enllaços a través de les xarxes socials els quals són:

- www.linkedin.com/company/thesportingnews
- <https://www.instagram.com/sportingnews>
- <https://www.facebook.com/thesportingnews>
- <https://www.youtube.com/sportingnews>

Per altra banda, abans de fer web scraping a la web oficial de l'NBA, ens hem assegurat que totes les dades de stats.nba.com estan totalment disponibles per a ús públic. Tanmateix, volem comunicar que l'objectiu del projecte es basa en finalitats ètiques i educatives per a un ús totalment educatiu.

Anàlisis similars

Existeixen diversos anàlisis semblants que es poden trobar per internet els quals han utilitzat tècniques similars. La majoria d'aquests anàlisis són bàsicament educatius per a l'ús d'eines en web scraping. Alguns d'aquests projectes són:

- <https://towardsdatascience.com/web-scraping-nba-2k-data-d7fdd4c8898c>
- <https://blog.geetest.com/en/article/web-scraping-NBA-salary>

- <https://medium.com/@osanchez2323/web-scraping-nba-stats-4b4f8c525994>

7. Inspiració.

El bàsquet és un esport més complicat d'analitzar que el baseball, pel simple fet d'estar en moviment constant i tenir més factors que afecten els números. Al bàsquet, la principal evolució sorgeix en què els fins ara elements bàsics d'anàlisi (percentatge de tir, anotació, rebot... per citar-ne alguns) segueixen tenint significat, és clar, però ja està del tot estesa allà la idea que no capten el dibuix complet del que ha passat. Cal anar més enllà.

Als Estats Units, són els amants més grans de les estadístiques en l'esport i això es reflecteix en tots els seus grans esports. El bàsquet no és una excepció, ja l'NBA, les estadístiques es tracten a tots els nivells. I és que analitzant les estadístiques, pots arribar a treure'n moltes conclusions sobre com s'ha desenvolupat un determinat partit, i quins jugadors han estat més encertats i en quins aspectes del joc. Tanmateix, els cossos tècnics dels equips de l'NBA i de la Lliga Endesa tenen especialistes en estadístiques i en obtenir informació valuosa dels números.

Un altre punt molt rellevant és respecte el negoci de les apostes, les quals també tenen en l'estadística la millor font d'informació per realitzar els seus pronòstics a la lliga nord-americana de bàsquet NBA i poder aspirar a anticipar algun resultat llegint de manera adequada dades numèriques.

Comparant amb anàlisis anteriors mencionats al punt 6 podem observar l'ús habitual i la importància de llibreries com Selenium i BeautifulSoup que ajuden en el procés de localització d'enllaços i en l'extracció de contingut de les pròpies webs. Són mètodes fàcils d'aplicar que formen part del procés de web scraping.

8. Llicència.

La llicència escollida per aquest repositori ha sigut la llicència Creative Commons Zero v1.0 Universal.

Hem escollit la llicència CC0 perquè permet als científics, educadors, artistes i altres creadors i propietaris de continguts amb drets d'autor a renunciar-hi perquè altres puguin utilitzar lliurement per a qualsevol propòsit sense restricció.

9. Codi.

Els datasets s'han generat executant la comanda "python web_scraping_nba.py" un cop la nostra terminal es troba en la carpeta "/src/web-scraping-nba/".

10. Dataset (Zenodo).

10.1.teams_standing_eastern_conference/teams_standing_western_conference

- <https://zenodo.org/record/6419210>

10.2. players_stats_eastern_conference i players_stats_western_conference

- <https://zenodo.org/record/6419205>

10.3. players_list_eastern_conference i players_list_western_conference

- <https://zenodo.org/record/6419199>

Taula de contribucions al treball

Signatures:

Alexandre Vidal de Palol – A.V.P

Adrián Alonso Gonzalo – A.A.G

Contribucions	Signatura
Investigació prèvia	A.A.G/A.V.P
Redacció de les respostes	A.A.G/A.V.P
Desenvolupament del codi	A.A.G/A.V.P