

Màster en Ciència de Dades

Tipologia i Cicle de Vida de les Dades

Pràctica 1 - Web Scraping

Autors: Carlos Gómez, Carlos Molina

Abril 2020

- 1. Context
- 2. Definir un títol pel dataset
- 3. Descripció del dataset
- 4. Representació gràfica
- 5. Contingut
- 6. Agraïments
- 7. Inspiració
- 8. Llicència
- 9. Codi
- 10. Dataset
- 11. Lliurar
- 12. Contribucions al treball

1. Context

Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

S'ha decidit fer aquesta pràctica de Web Scraping sobre Stadia, una nova plataforma de videojocs en streaming de Google.

La idea d'agafar aquesta plataforma de videojocs va néixer en veure que no existia cap lloc web on informes sobre l'evolució dels preus dels jocs dins d'aquesta plataforma, cosa que seria interessant per veure si en el moment de comprar un joc és una bona oferta o no. Ni la mateixa empresa de Google dona aquesta informació en obert (sense estar registrat), ni mitjançant la seva API. Suposem que en tractar-se d'una plataforma bastant nova, no han implementat encara aquesta API.

A part de fer un seguiment dels preus dels jocs, també es interessant poder extreure el llistat complet dels jocs que ja hi han a la plataforma i els que estan confirmats que vindran en un futur. En l'àmbit dels videojocs hi han moltes preguntes sobre la plataforma i quins jocs són els que hi han. En tractar-se d'una plataforma tancada als subscriptors "PRO", els usuaris normals només es poden informar d'un llistat de jocs que hi ha a la web oficial de

Stadia, però hem vist que no son tots els que hi han actualment dins la plataforma. Això ho hem pogut saber perquè un dels membres de l'equip té accés a la plataforma desde dins, cosa que volem aprofitar per extreure la informació dels jocs existents i fer un dataset amb aquesta informació.

Fins al dia 8 d'Abril de 2020, era una plataforma tancada als subscriptors de pagament, on només els primers inscrits, al mes de novembre, tenien accés. Ara mateix, s'ha obert aquest servei per a tothom amb un compte de Google. Pel que es pot utilitzar qualsevol compte per fer funcionar el script de scraping creat.

Per tant, en una primera fase ens hem marcat l'objectiu de:

- Obtenir un llistat amb tots els jocs disponibles i futurs, obtenint la informació de pàgines webs oficials.
- Anar a buscar dades sobre aquests videojocs a altres webs per complementar la informació disponible i donar més valor al dataset. Dades com la puntuació a altres plataformes i dates de llançament.

2. Definir un títol pel dataset

Triar un títol que sigui descriptiu.

Stadia Games Info

Es tracta d'un llistat, on cada registre es un videojoc, o complement, i s'amplia amb els atributs que donen informació sobre aquest videojoc.

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 triat).

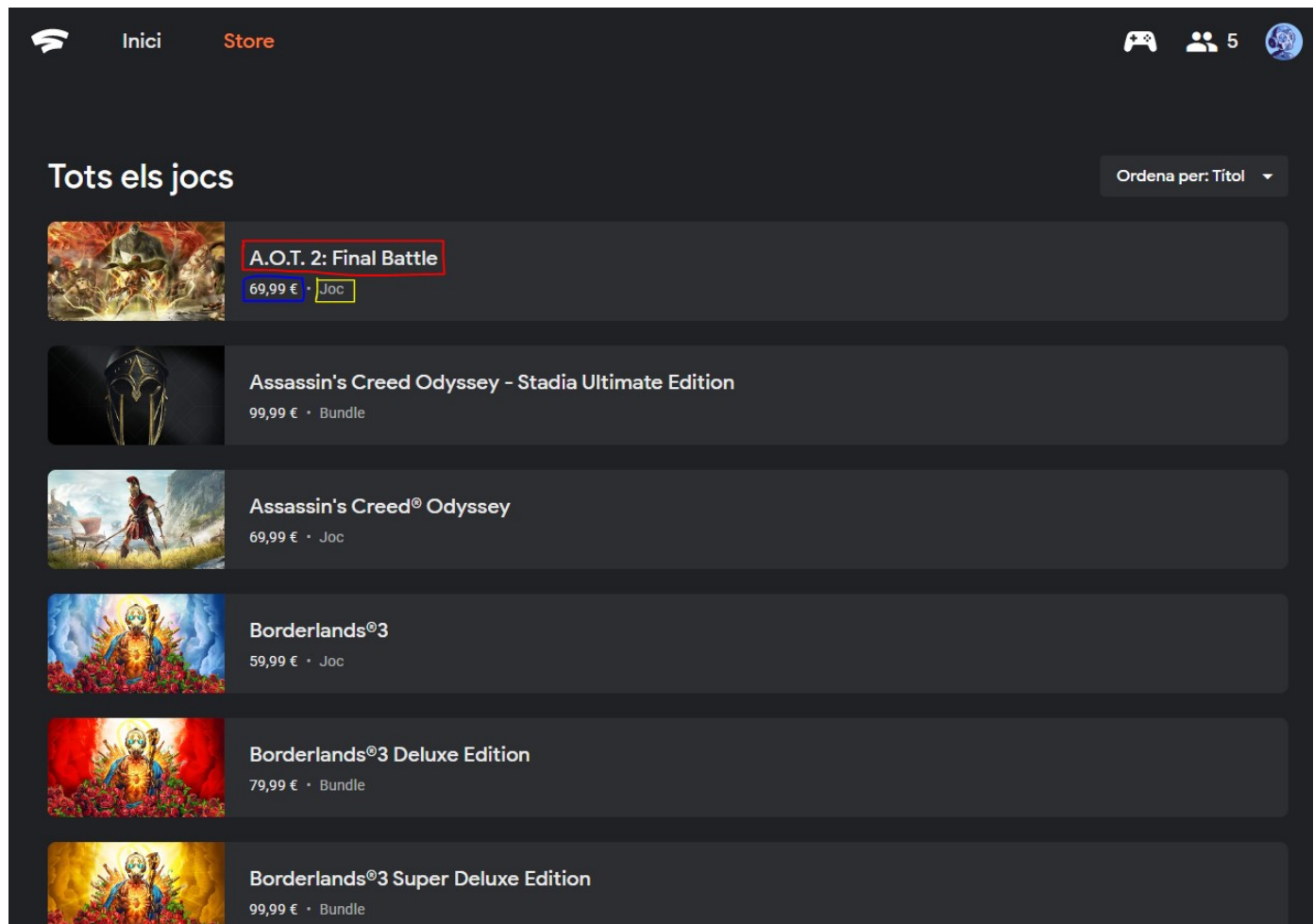
Aquest dataset conté dades de tres fonts:

- **Dades de obertes de la web pública de Stadia** https://store.google.com/es/product/stadia_games (https://store.google.com/es/product/stadia_games): Les dades d'aquesta font contenen els títols dels videojocs de la plataforma Stadia disponibles ara o en el futur.
- **Puntuacions dels jocs per a altres plataformes**: Per a cada joc de la plataforma Stadia, s'ha buscat la puntuació de revistes especialitzades i dels usuaris per a les plataformes PC, PlayStation 4 i XboxOne. Les dades s'han extret de la pàgina web www.metacritic.com (www.metacritic.com)
- **Preu dels jocs a la plataforma Stadia**: Aquestes dades s'han extret de la pàgina de la plataforma Stadia. Per a accedir a aquestes dades, s'ha d'estar registrat a la plataforma.

4. Representació gràfica

Presentar una imatge o esquema que identifiqui el dataset visualment

La primera imatge es una captura de la web interna de Stadia on pots comprar els videojocs. Es pot observar que d'aquesta web podem obtenir el títol, preu i tipus de joc:



La segona web on anar a buscar informació és la pàgina on pots inscriure't a la plataforma. En una de les pestanyes de la web pots veure un llistat de jocs que anuncien: https://store.google.com/es/product/stadia_games (https://store.google.com/es/product/stadia_games)



Assassin's Creed Odyssey



Attack on Titan 2: Final Battle



Baldur's Gate 3



Borderlands 3



Cyberpunk 2077



Darksiders Genesis



Destiny 2: The Collection



Destroy All Humans!



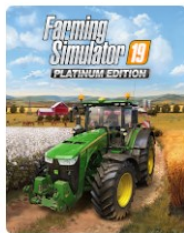
DOOM



DOOM Eternal



DRAGON BALL XENOVERSE 2



Farming Simulator 19



FINAL FANTASY XV



Football Manager 2020



Get Packed

5. Contingut

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

El dataset inclou els següents camps.

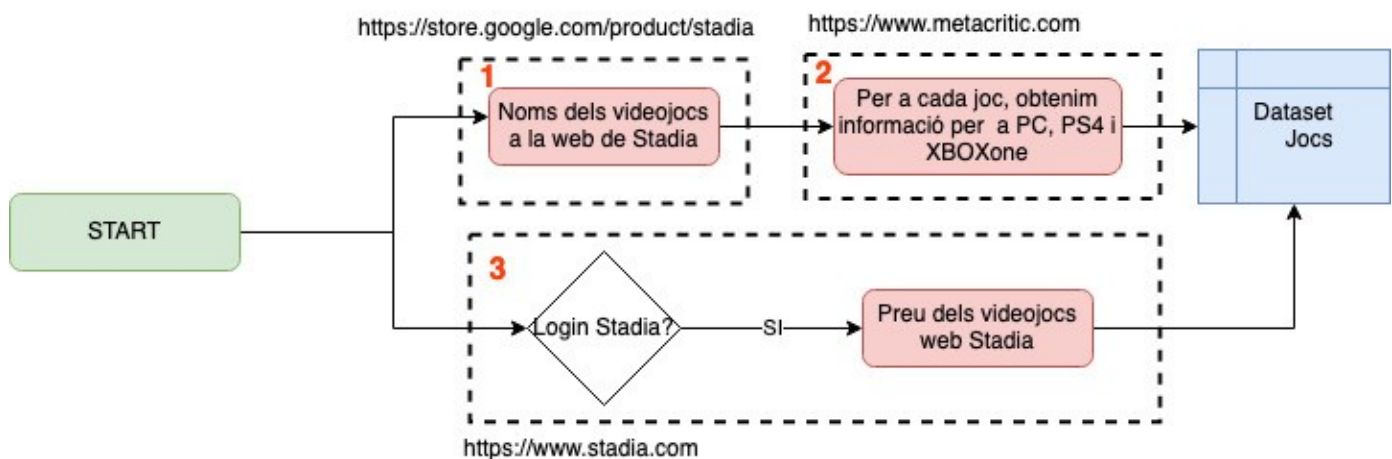
- **title:** Tipus CHAR. Descriu el títol del videojoc.
- **pc_genere:** Tipus CHAR. Descriu el gènere el videojoc per a PC segons la web metacritic.
- **pc_release:** Tipus DATE. Data de llançament per a PC segons la web metacritic.
- **pc_metascore:** Tipus FLOAT. Conté la mitjana de la nota de revistes especialitzades per al joc en PC.
- **pc_n_reviews:** Tipus INT. Conté el número de reviews de revistes especialitzades per al joc en PC.
- **pc_user_score:** Tipus FLOAT. Conté la mitjana de la nota donada pels usuaris per al joc en PC.
- **pc_n_user_score:** Tipus INT. Conté el número de reviews d'usuaris per al joc en PC.
- **pc_developer:** Tipus CHAR. Desenvolupador del joc per a PC.
- **pc_n_online_multiplayer:** Tipus CHAR. Número de jugadors online per al joc en PC.
- **ps4_genere:** Tipus CHAR. Descriu el gènere el videojoc per a ps4 segons la web metacritic.
- **ps4_release:** Tipus DATE. Data de llançament per a PC segons la web metacritic.
- **ps4_metascore:** Tipus FLOAT. Conté la mitjana de la nota de revistes especialitzades per al joc en PC.
- **ps4_n_reviews:** Tipus INT. Conté el número de reviews de revistes especialitzades per al joc en PC.
- **ps4_user_score:** Tipus FLOAT. Conté la mitjana de la nota donada pels usuaris per al joc en PC.
- **ps4_n_user_score:** Tipus INT. Conté el número de reviews d'usuaris per al joc en PC.
- **ps4_developer:** Tipus CHAR. Desenvolupador del joc per a PC.

- **ps4_n_online_multiplayer**: Tipus CHAR. Número de jugadors online per al joc en PC.
- **xboxone_genre**: Tipus CHAR. Descriu el gènere el videojoc per a xboxone segons la web metacritic.
- **xboxone_release**: Tipus DATE. Data de llançament per a PC segons la web metacritic.
- **xboxone metascore**: Tipus FLOAT. Conté la mitjana de la nota de revistes especialitzades per al joc en PC.
- **xboxone_n_reviews**: Tipus INT. Conté el número de reviews de revistes especialitzades per al joc en PC.
- **xboxone_user_score**: Tipus FLOAT. Conté la mitjana de la nota donada pels usuaris per al joc en PC.
- **xboxone_n_user_score**: Tipus INT. Conté el número de reviews d'usuaris per al joc en PC.
- **xboxone_developer**: Tipus CHAR. Desenvolupador del joc per a PC.
- **xboxone_n_online_multiplayer**: Tipus CHAR. Número de jugadors online per al joc en PC.
- **price_stadia**: Tipus FLOAT. Preu indicat a la web d'Stadia.

S'ha de tenir en compte que molts jocs només disposen d'informació disponible per a cap, una o dues plataformes. En el primer cas, és degut a que els jocs encara no han soprtit al mercat. En el segon i tercer cas, és degut a que aquests jocs no s'han llançat per a totes les plataformes.

El període de temps de les dades correspon a les dades de cada joc des de la seva data de llançament fins al dia en que s'ha executat el scrapper per última vegada. Si el scrapper s'executa de nou en el futur, les dades de jocs nous apareixeran actualitzades.

El metode per a recollir les dades es resumeix en el següent diagrama:



En el cas dels noms dels videojocs (número 1 en el diagrama), aquests s'han extret directament de la pàgina pública de Stadia https://store.google.com/es/product/stadia_games (https://store.google.com/es/product/stadia_games).

Les dades videojocs de la pàgina de Metacrític (número 2 en el diagrama) s'han extret en dos passos. En el primer pas, s'ha extret el títol i el link de **tots** els jocs de PC, PS4 i XBOXone. Tot i que aquest procés pot arribar a durar força minuts, s'ha escollit aquest mètode degut a que a l'arxiu `robots.txt` de la pàgina www.metacritic.com (www.metacritic.com) s'indica explícitament que no s'utilitzi l'opció de cerca quan es fa scraping. En el segon pas, un cop s'ha extret la llista de tots els jocs, aquesta s'ha filtrat per a fer coincidir els títols dels jocs de la web Metacritic amb els títols de la llista de videojocs extrets de la web pública de Stadia. Amb la llista filtrada, s'ha accedit a la pàgina de cada joc, obtenint-ne les dades d'interès.

Pel que fa a les dades de preus de la web de Stadia (número 3 en el diagrama), primer s'ha de fer login amb una compta de Google. Per a fer això, l'única manera que hem trobat per a fer-ho és fent login a Google desde una pàgina externa (en el nostre cas *StackOverflow*), per a després accedir a la web de Stadia. Aquest pas només s'executa en el cas de tenir accés a la web de Stadia.

6. Agraïments

Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Les dades presentades en aquesta pràctica relatives a puntuacions dels jocs per a altres plataformes pertanyen a la web www.metacritic.com.

7. Inspiració

Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Existeixen webs semblants, que extreuen un llistat dels jocs disponibles a la plataforma. Però, aquests portals estan mantinguts a mà pel seu creador i no informen sobre el preu dels jocs:

- Stadia Game DB —> <https://stadiagamedb.com/> (<https://stadiagamedb.com/>)
- The Stadia DB —> thestadiadb.com (thestadiadb.com)

La idea d'aquesta pràctica es poder tenir un script automàtic que obtingui tota aquesta informació en el moment d'executar-lo.

Una de les utilitats que pot tenir aquest script és la de extreure la informació diàriament, programant aquest script perquè s'executi un cop al dia. D'aquesta manera es podria fer un seguiment dels preus dels videojocs a nivell individual i poder arribar a fer les visualitzacions oportunes sobre aquesta informació.

8. Llicència

Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

Per aquest conjunt de dades hem escollit una llicència de Cultura Lliure: **Reconeixement-Compartir Igual 4.0 Internacional de Creative Commons**



CC BY-SA 4.0 License

- **BY:** Amb aquest tipus de llicència es dona llibertat per copiar, reproduir o modificar el codi, sempre que es reconegui l'autoria del projecte original, per donar valor a la feina feta en aquesta pràctica.
- **SA:** També afegim que els projectes derivats d'aquests, es comparteixin de la mateixa manera, per evitar que un producte derivat d'aquest tingui una llicència que prohibeixi l'ús comercial, limitant així les possibilitats de qualsevol persona que vulgui desenvolupar

9. Codi

Adjuntar el codi amb el qual s'ha generat el dataset,

preferiblement en Python o, alternativament, en R.

El codi es troba en la carpeta `src` a l'arrel del repositori de github.com/KRLSMolina/ScrapingStadia (<https://github.com/KRLSMolina/ScrapingStadia>).

A l'arxiu `scraper_external.py` es troba el codi necessari per a obtenir els títols dels jocs i les dades de la web Metacritic. A l'arxiu `scraper_stadia.py` es troba el codi per a fer el login a la web de Stadia i obtenir les dades referents als preus dels jocs.

10. Dataset

Publicar el dataset en format CSV a Zenodo amb una xicoteta descripció.

El DOI del dataset a Zenodo: <https://doi.org/10.5281/zenodo.3750425> (<https://doi.org/10.5281/zenodo.3750425>)

11. Lliurar

Presentar el treball amb el DOI del dataset a Github

Tota la informació es pot trobar al repository de Github: <https://github.com/KRLSMolina/ScrapingStadia> (<https://github.com/KRLSMolina/ScrapingStadia>)

12. Contribucions al treball

Contribucions	Signa
Recerca prèvia	CM, CG
Redacció de les respostes	CM, CG
Desenvolupament codi	CM, CG