Tipologia i Cicle de Vida de les Dades. Pràctica 1

Autor: Marta Montclus, Jose Montufo
Assignatura: Tipologia i cicle de vida de les dades
Codi d'assignatura: M2.951
Aula: 1
Curs: 2020-21, semestre 2
Professor: Xavier Vivancos Garcia

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

L'objectiu de la pràctica consisteix en elaborar un catàleg de jocs de taula amb les seves característiques com ara el nombre de jugadors, el temps estimat de cada partida o el tipus de joc.

Concretament s'extraurà la informació de la web de l'empresa Zacatrus, que comercialitza jocs de taula de forma online i alhora també és editora de jocs propis i accessoris.

Un fet interessant que ens fa fixar-nos en aquesta web i no d'altres és que proporciona l'identificador del joc dins de la base de dades de BGG (Board Game Geek).

BGG és la base de dade més gran de jocs de taula que existeix a la xarxa, i proporciona, per cada joc, gran quantitat d'informació, des de les dades més basiques (com les que trobaríem a Zacatrus.es) fins a indicadors de complexitat, popularitat, etc. En aquest cas, BGG proporciona una API que permet accedir aquesta informació de manera massiva.

La nostra pràctica, doncs, consisteix a obtenir el catàleg de productes de zacatrus.es mitjançant tecniques de *web scraping*.

El projecte es podrà complementar amb dos subprojectes que milloraran la utilitat de les dades de cara a fer-ne anàlisis posteriors:

- Incorporació d'atributs descarregats de la BGG com ara valoració dels usuaris o complexitat
- Incorporació d'una visió historica de preus.

Ambdós subprojectes s'implementaran per aquesta practica en funció de la disponibilitat de temps.

2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

zaca_games dataset ???????

Hem triat aquest nom perquè es identificatiu de la tipologia de productes que inclou (games) i alhora de la principal font on l'hem obtingut (Zacatrus.es).

Definim el títol del dataset com "**Catàleg de jocs de taula de la botiga online de Zacatrus.es**"

Hem triat aquest nom perquè defineix de forma concisa però acurada el contingut del dataset. Determina tant el contingut de les dades (jocs de taula) com l'origen de les dades (botiga online de Zacatrus.es)

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).

El dataset zaca_games conté una imatge estatica dels jocs de taula comercialitzats per l'empresa Zacatrus.

Cadascuna de les línies del dataset representa un dels jocs de taula que es troben en venda a la botiga online de Zacatrus en un moment determinat en el qual s'ha realitzat un scraping de la botiga. Per cadascun dels jocs, s'obtindran tres tipus d'atributs, representats a les columnes del dataset:

- Atributs relacionats amb les dades de venda del joc de taula a Zacatrus, com per exemple la seva disponibilitat, i el seu preu.
- Característiques pròpies del joc que proporciona Zacatrus a la seva botiga, com per exemple el nombre de jugadors, el temps de joc estimat, etcètera.
- Característiques addicionals pròpies del joc obtingudes de la BGG, com per exemple (completar quan les obtinguem)
- Valoració del joc obtinguda de la BGG, juntament amb la seva classificació.

4. Representació gràfica. Presentar 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.

Variable	Descripció	Font	Periode de temps
name	VARCHAR. Nom amb el qual es comercialitza el joc	Scrapping Zacatrus.es	Abril 2021
price	INT. Preu en el moment de la descarrega de dades	Scrapping Zacatrus.es	Abril 2021

availability	BOOL. Disponible / No disponible en el moment de la descarrega a Zacatrus.es	Scrapping Zacatrus.es	Abril 2021
autor	VARCHAR. Autor del joc	Scrapping Zacatrus.es	Abril 2021
BGG	INT. Codi del joc a BGG	Scrapping Zacatrus.es	Abril 2021
tematica	VARCHAR. Aspectes sobre els quals tracta el joc (ex. cultura, investigació, comerç)	Scrapping Zacatrus.es	Abril 2021
sibuscas	VARCHAR. Paraules claus per buscar el joc relacionades amb el context de joc (ex. familia, festa)	Scrapping Zacatrus.es	Abril 2021
edad	VARCHAR. Trams d'edat pels quals es recomana el joc.	Scrapping Zacatrus.es	Abril 2021
num_jugadores	VARCHAR. Nombre de jugadors que poden participar	Scrapping Zacatrus.es	Abril 2021
tiempo	VARCHAR. Temps que dura una partida de mitjana	Scrapping Zacatrus.es	Abril 2021
medidas	VARCHAR. Dimensions de l'embalatge	Scrapping Zacatrus.es	Abril 2021
complejidad	VARCHAR. Dificultat del joc (Facil, Medio, Dificil)	Scrapping Zacatrus.es	Abril 2021
editorial	VARCHAR. Editor del joc	Scrapping Zacatrus.es	Abril 2021
dependencia_idioma	VARCHAR. Descriu si és necessari conèixer l'idioma del joc per poder jugar (Ex. No necessari, Només instruccions, Alta).	Scrapping Zacatrus.es	Abril 2021
mecanica	Mecanisme del joc (Ex. Pregunta/resposta, gestió de recursos, crawler)	Scrapping Zacatrus.es	Abril 2021
idioma	Idioma en el qual es comercialitza el joc	Scrapping Zacatrus.es	Abril 2021

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.

Les dades principals s'han extret de la web de l'empresa Zacatrus, que ven i edita jocs de taula.

Com que les dades s'han obtingut amb un objectiu purament acadèmic, no s'ha considerat necessari demanar permís a Zacatrus per a l'obtenció de les dades de la seva botiga. A més a més, s'ha respectat en tot moment el contingut del fitxer robots.txt de la botiga. També s'han seguit les bones pràctiques recomanades per efectuar web scraping de forma respectuosa amb el lloc web escanejat, com per exemple l'espaiament temporal entre crides HTTP al servidor de la botiga per obtenir els diferents HTMLs de la pàgina.

No s'han trobat projectes previs d''scraping de la web de Zacatrus. Aquest tipus de projectes d'obtenció del catàleg d'una botiga online son habitualment realitzats per part de la competència, i per tant no solen ser publicats. En canvi, sí que s'han trobat numerosos projectes d'scraping de la BGG i/o de pàgines similars, que construeixen un llistat de jocs de taula amb les seves característiques i puntuacions, però que no disposen com el nostre del preu i la disponibilitat del joc en un distribuidor determinat. Els enllaços a alguns d'aquests projectes son els següents:

Web Scraping Board Game Descriptions with Python

Board games data scraping and processing from BoardGameGeek and more!

Scraping Popular Board Games

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.

Hem volgut realitzar un conjunt de dades força versàtil, que recull força informació per als diferents jocs de taula de diferents fonts. A diferència dels anàlisis anteriors, que disposaven d'un catàleg més o menys complet de jocs enlloc del catàleg d'una botiga en concret, aquests no disposaven del preu dels jocs. D'aquesta manera, el nostre dataset podria ser utilitzat per diversos agents diferents, aportant a cadascun d'ells respostes a diferents preguntes que es poden plantejar:

- Per a una botiga online de la **competència**, aquest conjunt de dades els podria servir per analitzar la política de preus i d'estocs de Zacatrus, i contrarestar-la amb les polítiques pròpies.
- Per a un posible **comprador**, el conjunt de dades el pot servir per detectar ofertes entre els jocs que s'adapten als seus gustos en funció de les característiques dels jocs. La puntuació a la BGG també pot ser un factor interessant per a aquest perfil.
- Per a una **pàgina de comparació de preus** enfocada als jocs de taula, es podria utilitzar juntament amb altres datasets realitzats a partir dels preus a altres botigues per afegir els preus de Zacatrus a la seva comparació. En aquest cas la falta d'un identificador únic universal dels jocs (per exemple, el codi EAN) pot suposar un problema, però en part es pot solventar amb l'identificador dels jocs a la BGG.

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

○ 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 Open Database License, individual contents under Database Contents License

○ Other (specified above)

○ Unknown License

9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

10. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

Taula de Contribucions:

Contribucions	Signa
Recerca previa	
Redacció de les respostes	
Desenvolupament del codi	

Recursos

- Materials de web scraping de l'assignatura