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 a 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 dades 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 bàsiques (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 tècniques de web scraping.

El projecte es complementa amb un subprojecte que millorarà la utilitat de les dades de cara a fer-ne anàlisis posteriors, la incorporació de la valoració realitzada per part dels usuaris dels jocs a la Board Game Geek.

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

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).*

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.

- Valoració del joc obtinguda de la BGG.

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
Num. Id	INT. Codi del joc al nostre dataset. Serveix per identificar el directori on es troben les seves imatges	Generat pel codi	Abril 2021
name	VARCHAR. Nom amb el qual es comercialitza el joc	Scraping Zacatrus.es	Abril 2021
price	INT. Preu en el moment de la descarrega de dades	Scraping Zacatrus.es	Abril 2021
availability	VARCHAR. Disponible / No disponible en el moment de la descarrega a Zacatrus.es	Scraping Zacatrus.es	Abril 2021
autor	VARCHAR. Autor del joc	Scraping Zacatrus.es	Abril 2021

BGG	INT. Codi del joc a BGG	Scraping Zacatrus.es	Abril 2021
tematica	VARCHAR. Aspectes sobre els quals tracta el joc (ex. cultura, investigació, comerç)	Scraping Zacatrus.es	Abril 2021
sibuscas	VARCHAR. Paraules claus per buscar el joc relacionades amb el context de joc (ex. familia, festa)	Scraping Zacatrus.es	Abril 2021
edad	VARCHAR. Trams d'edat pels quals es recomana el joc.	Scraping Zacatrus.es	Abril 2021
num_jugadores	VARCHAR. Nombre de jugadors que poden participar	Scraping Zacatrus.es	Abril 2021
tiempo	VARCHAR. Temps que dura una partida de mitjana	Scraping Zacatrus.es	Abril 2021
medidas	VARCHAR. Dimensions de l'embalatge	Scraping Zacatrus.es	Abril 2021
complejidad	VARCHAR. Dificultat del joc (Facil, Medio, Dificil)	Scraping Zacatrus.es	Abril 2021
editorial	VARCHAR. Editor del joc	Scraping 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).	Scraping Zacatrus.es	Abril 2021
mecanica	Mecanisme del joc (Ex. Pregunta/resposta, gestió de recursos, crawler)	Scraping Zacatrus.es	Abril 2021
idioma	Idioma en el qual es comercialitza el joc	Scraping Zacatrus.es	Abril 2021
ranking	Valoració del joc per part dels usuaris de la BGG	API de la BGG	Abril 2021

Distingim quatre orígens diferents de les dades:

- L'id del joc al nostre dataset, es genera en el moment de realitzar el dataset. Aquest identificador s'ha afegit amb l'objectiu d'identificar el directori on es troben les imatges de cadascun dels jocs.
- Els camps name, price i availability s'obtenen de les metadades que es troben al codi html de les pàgines dels jocs de taula a Zacatrus. Concretament, dels camps de metadada "og:title" i "product:price:amount" respectivament.
- Per obtenir el camp availability, la obtenim directament del "div" que s'utilitza per mostrar-la a la pàgina corresponent del joc. La resta dels camps que s'obtenen de fer

scraping a Zacatrus (tota la resta menys ranking) s'obtenen de la "table" que al html de les pàgines de Zacatrus té el id "product-attribute-specs-table". D'aquesta taula, es recorren totes les seves files ("tr") i s'obtenen els atributs disponibles per al joc.
- El ranking, s'obté en un procés posterior, consultant el seu valor a la BGG a partir de la API que proporcionen.

A més a més d'aquests atributs, també s'obtenen les imatges que publica Zacatrus a la pàgina de cadascun dels jocs. Per obtenir-les, s'han obtingut de l'html dels jocs els scripts, i es filtra l'script que controla el funcionament de la galeria d'imatges (és aquell que conté un json amb un element "mage/gallery/gallery"). Una vegada trobat aquest script, s'ha parsejat el json que conté per tal d'extreure les url de les imatges.

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 i a la qual agraïm la possibilitat de poder-les utilitzar.

Com que les dades s'han obtingut amb un objectiu purament acadèmic, no s'ha considerat necessari demanar permís formalment 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, que no posa cap mena de limitació a la realització del web scraping del lloc web (només proporciona els enllaços al sitemap del lloc). 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 són habitualment realitzats per part de la competència, i per tant no solen ser publicats. En canvi, sí que s'han trobat nombrosos 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 distribuïdor determinat. Els enllaços a alguns d'aquests projectes són 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 diverses fonts. A diferència de les anàlisis anteriors, que disposaven d'un catàleg més o menys complet de jocs en lloc 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 possible 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 resoldre 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:

Donat que per elaborar el dataset utilitzem dades d'una empresa privada, seleccionem per a la seva publicació la llicència CC BY-NC-SA 4.0, ja que limita el possible ús comercial que es faci de les dades. Aquesta llicència implica el següent:

- Reconeixement: Qualsevol que utilitzi el dataset té l'obligació de reconèixer l'autoria. També s'ha d'indicar si s'han realitzat canvis, i afegir un enllaç a la font original.
- No Comercial: No es pot utilitzar aquest dataset amb finalitats lucratives.
- Compartir Igual: Encara que es realitzin modificacions al dataset, aquest dataset modificat només es pot compartir amb la mateixa llicència de l'original.

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

Tant el codi com el dataset són accessibles a través del següent enllaç:
https://github.com/jmontufo/zacatrusScraper

Per a la realització de la pràctica s'han creat tres fitxers python:

- zacatrusScraper.py --> Codi que realitza la tasca de scraping al lloc web de Zacatrus.es per realitzar un dataset amb el seu catàleg.
- get_avg_rating.py --> Codi que afegeix una columna al dataset amb la valoració dels usuaris a la base de dades de la BoardGameGeek, si la valoració es troba disponible.
- Throttle.py --> Classe que permet afegir una espera entre crides a un mateix domini, per tal de no saturar el lloc web al qual es fa web scraping.

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

El dataset es pot trobar a l'enllaç: https://doi.org/10.5281/zenodo.4679367

Dins el dataset es poden trobar dos fitxers csv, i un zip amb les imatges dels jocs descarregades de Zacatrus.es.

El primer dels csv, games.csv, conté el resultat d'executar el codi de zacatrusScraper.py. Per tant, contindrà el llistat de jocs del catàleg de Zacatrus, amb totes les seves dades definides al punt 5 tret del rating que s'obté de la BGG. El format del fitxer és un csv amb les dades separades per punts i coma (;), amb una fila inicial per als noms dels atributs.

L'altre csv, games_with_rating.csv, conté el resultat d'executar el codi de get_avg_rating.py. Per tant, el seu contingut serà el mateix que el de games.csv, però amb una columna adicional per a l'atribut rating. El format del fitxer és el mateix que el de games.csv.

Finalment, també es pot trobar un directori comprimit Pictures.zip, que conté una carpeta per cadascun dels jocs del dataset. El nom de cadascuna de les carpetes es correspon amb l'identificador que hem assignat als jocs al nostre dataset. El contingut de cadascuna de les carpetes són les imatges que s'han obtingut per al joc en qüestió.

Taula de Contribucions:

Contribucions	Signa
Recerca previa	JM, MM
Redacció de les respostes	JM, MM
Desenvolupament del codi	JM, MM

Recursos

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2010). El lenguaje Python. Editorial UOC.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.