TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1: WEB SCRAPING

Joan Cueva Moreno

Universitat Oberta de Catalunya

12 d'abril de 22

ÍNDEX

CONTEXT 3
TÍTOL 3
DESCRIPCIÓ DEL DATASET 3
REPRESENTACIÓ GRÀFICA 3
CONTINGUT 4
AGRAÏMENTS 5
INSPIRACIÓ 6
LLICÈNCIA 6
CODI 7
DATASET 7
VÍDEO 7
REFERÈNCIES 7
CONTRIBUCIONS 7

CONTEXT

Per la realització d'aquesta pràctica, havíem de realitzar web scraping sobre una pàgina web per recopilar les seves dades en un dataset. Per afinitat, he realitzat la pràctica sobre una pàgina web que faig servir sovint, ECC Cómics (ecccomics.com).

Aquesta pàgina web és de l'editorial ECC Cómics, i és on mostren els comics i altres productes que tenen disponibles al seu catàleg. Per cada producte, ofereixen informació extra com: nom, descripció, preu, etc.

He agafat aquesta web perquè crec que pot sortir un dataset bastant complet amb tota la informació que tenen disponible i perquè, segurament, és un web scraper al que li puc donar un ús personal.

TÍTOL

Hem anomenat el nostre dataset "ecc_comics_products" ja que és essencialment això: un dataset que conté els productes disponibles per l'editorial ECC Cómics.

DESCRIPCIÓ DEL DATASET

Com ja hem comentat, el nostre dataset conté informació extreta de les fitxes dels productes publicats per ECC Cómics a la seva pàgina web.

Les dades inclouran informació com el nom del producte, la seva descripció, el seu format, el seu preu, etc. A més, tenim informació també dels diferents catàlegs, categories i sèries, informació que hem pogut extreure de la pàgina web i que ens servirà per classificar millor els productes.

REPRESENTACIÓ GRÀFICA

El nostre esquema (fig. 1) representa els models que hem fet servir al codi, que inclouen informació extra que no hem exportat al nostre dataset en csv, els enllaços a la informació.

Aquests enllaços els hem fet servir bàsicament per fer el scraping, i no creiem que aportin informació important al dataset. A més, els enllaços poden canviar i per tant aportarien més problemes que aspectes positius. Un altre motiu per no incloure'ls és que els usuaris facin servir la informació de cada producte del dataset sense anar a cercar-amb el link.

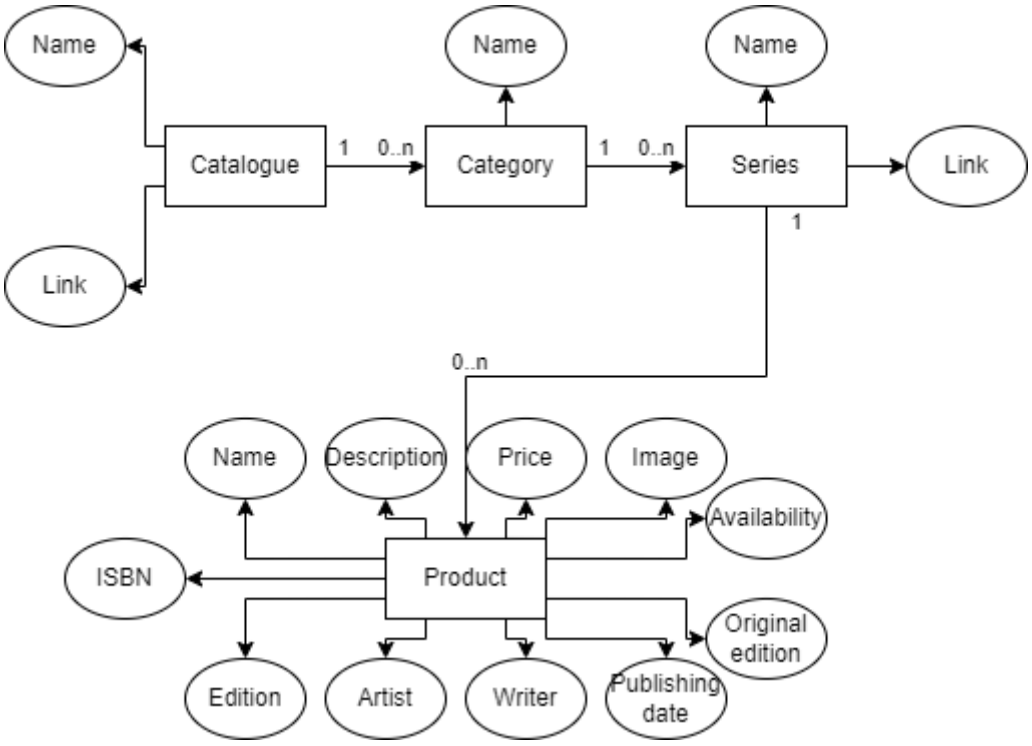

Fig 1: Esquema dels models de les dades.

CONTINGUT

El nostre dataset conté catorze paràmetres, que descrivim a continuació:

- **catalogue**: String. Nom del catàleg al que pertany. Exemple: "DC Deluxe".
- **category:** String. Nom de la categoria a la que pertany. Exemple: "Ediciones Deluxe – Universo DC".
- **series:** String. Nom de la sèrie a la que pertany. Exemple: "Batman".
- **product_name:** String. Nom del producte. Exemple: "Batman: Amor loco (Edición Deluxe)".
- **product_descrition:** String. Descripció del producte. Exemple: "Los genios responsables de la aclamada serie de animación de Batman, Paul Dini y Bruce Timm...".
- **product_price:** Float. Preu del producte, en euros (€). Exemple: "26.0".
- **product_image:** String. Enllaç a l'imatge del producte. Exemple: "https://www.ecccomics.com/content/productos/4480/cubierta_batman_amor_loco_deluxe_FORRO_WEB_156.jpg".
- **product_availability:** Boolean. Disponibilitat del producte. Exemple: "True".
- **product_original_edition:** String. Edició original del producte. Exemple: "The Batman Adventures: Mad Love USA, Harley Quinn núm. 14 USA (extracto portada aniversario)".
- **product_publishing_date:** String. Mes i any de publicació del producte. Exemple: "Junio de 2018".
- **product_writer:** String. Escriptors/guionistes del còmic, separats per ",". Exemple: "Bruce Timm, Paul Dini".

- **product_artist:** String. Dibuixants del còmic, separats per ",". Exemple: "Bruce Timm".
- **product_edition:** String. Dades de l'edició del producte: format, dimensions, número de pàgines, data quan es fa disponible, etc. Exemple: "Cartoné (Deluxe), 144 págs. A color. Disponible el 15/05/2018".
- **product_isbn:** String. ISBN o número de referència del producte. Exemple: "978-84-17441-69-2".

Per definir el període temporal del nostre dataset, agafarem com a referència la columna "product_publishing_date". Observant aquests valors, podem dir que el període de temps de les nostres dades és de novembre de 2010 a maig de 2022. Per tant, tenim casi 12 anys de publicacions de fitxes de productes.

AGRAÏMENTS

Voldria agrair a l'editorial ECC Cómics i a l'empresa El Catálogo del Cómic, empresa propietària de l'editorial, per les dades publicades a la seva pàgina web.

El dataset no podrà ser publicat degut al que especifiquen a l'apartat "Derechos de Propiedad Intelectual e Industrial", del seu "Aviso legal" [1].

> Derechos de Propiedad Intelectual e Industrial
>
> ECC es la única y exclusiva titular de los derechos de propiedad intelectual e industrial sobre las marcas, imágenes, textos, diseños, animaciones, programación y diseño de las Webs o cualquier otro contenido o elementos de las mismas, o, en su caso, dispone de los permisos o licencias necesarias para su utilización. El Usuario reconoce y acepta que el acceso y/o descarga de cualquier contenido y/o elemento que se encuentre a su disposición a través de cualquiera de las Webs es para su uso personal e intransferible. Cualquier acto de reproducción, distribución, comunicación pública, puesta a disposición, o transformación, así como cualquier otra forma de explotación de todo o parte de dichos contenidos o elementos, realizado bajo cualquier forma o mediante cualquier medio, requerirá el consentimiento previo y por escrito de ECC, o en su caso, de su titular.
>
> Asimismo, todos los signos distintivos que aparecen en las Webs o son titularidad de ECC o cuenta con las correspondientes licencias y/o autorizaciones, y se encuentran debidamente registrados, quedando prohibida su reproducción o distribución bajo ningún medio, sin la debida, previa y expresa autorización de su titular.
>
> El acceso y navegación por las Webs en ningún caso se entenderá como una renuncia, transmisión, licencia o cesión total ni parcial de los derechos antes indicados por parte de ECC o, en su caso, del titular de los derechos al que correspondan.

Per actuar èticament, hem volgut només descarregar dades públiques que tothom pugui trobar al seu web. Totes les dades han sigut recopilades de llocs públics i visibles de la seva pàgina web, accessible sense necessitat de registre, identificació o permís previ.

També, per evitar danyar els seus servidors, hem posat una pausa d'un segon després de cada request. Així evitem sobresaturar el seu servidor i denegar-los el servei.

Per últim, no hem accedit a cap informació que no estigui permesa segons el seu robots.txt.

Els anàlisis que hem estudiat que poden estar relacionats (productes d'alguna pàgina web), es centren en agafar preus, descripcions i reviews/rating. És el cas de l'anàlisi referenciat a la bibliografia [2]. En el nostre cas, no tenim informació de reviews/ratings, però tenim molta més informació de format, paginació, autors, etc.

INSPIRACIÓ

Aquest dataset és interessant perquè conté bastant informació dels productes i els autors que d'ECC Cómics.

En general, des d'un punt de vista comercial, podríem voler tenir informació principalment dels preus. Aquesta seria la situació d'estar fent aquest anàlisi per comparar-lo amb una botiga teva similar, perquè estàs pensant obrir una editorial, etc.

També, des d'un punt de vista més centrat en el desenvolupament d'aplicacions, podríem voler tenir un dataset de totes les obres publicades per autor per mostrar a la nostra pàgina web o app. Aquest dataset també ens seria útil en aquest cas, ja que podem extreure les obres de tots els autors que han publicat.

En comparació amb l'anàlisi que comentàvem al punt anterior, encara que perdem informació important com les reviews/ratings, estem també guanyant molta informació que fins i tot, ens podria permetre crear una app o web amb aquesta informació.

LLICÈNCIA

Com hem explicat abans, a l'apartat d'Agraïments, no podem publicar el dataset resultant de l'execució del web scraper degut a les indicacions que trobem a l'apartat "Derechos de Propiedad Intelectual e Industrial", del seu "Aviso legal".

El dataset amb dades simulades serà publicat amb llicència "CC0: Public Domain License".

Hem escollit aquesta llicència degut a que volem que sigui un dataset d'accés completament públic i que tothom pugui treballar amb ell, distribuir-ho, millorar-ho i aportar els extres que cregui convenients.

A més, en ser un dataset amb dades simulades, no tenint cap restricció prèvia de copyright que ens obligui a publicar-ho amb una certa llicència.

CODI

El codi ha sigut publicat a un repositori públic de GitHub al meu perfil "jcuevam" associat al meu correu de la UOC:

jcuevam/UOC.TCVD.WebScraping: Repositori pel projecte de Tipología i Cicle de Vida de les Dades, Pràctica 1: Web Scraping. (github.com)

DATASET

Com hem comentat, hem hagut de publicar a Zenodo un dataset simulat. Es pot trobar aquí:

ECC Cómics Products | Zenodo

VÍDEO

Hem realitzat un vídeo explicatiu del nostre projecte per la pràctica 1. El podem trobar al Google Drive amb l'enllaç que s'ha facilitat amb el lliurament de la pràctica.

REFERÈNCIES

1. Aviso legal [en línia] [consulta: 01/04/2022]. Disponible a: https://www.ecccomics.com/comic/legal.aspx

2. Web Scraping in Python – How to Scrape an eCommerce Website Using Beautiful Soup and Pandas [en línia] [consulta: 03/04/2022]. Disponible a:

https://www.freecodecamp.org/news/scraping-ecommerce-website-with-python/

CONTRIBUCIONS

CONTRIBUCIONS	DISGNATURA
Investigació prèvia	J C M
Redacció de les respostes	J C M
Desenvolupament del codi	J C M