Tipologia i Cicle de Vida de les Dades

PRAC 1: WEB SCRAPING

09/10/2020

Sergio Costa
Lorena Casanova

Índex

1. Context

El joc de dades que presentem s'ha recol·lectat en el marc de la Pràctica 1 de l'assignatura **Tipologia i Cicle de Vida de les Dades** del **Màster Universitari de Ciència de Dades** de la **Universitat Oberta de Catalunya (UOC)** durant el primer semestre del Curs 20/21.

L'objectiu de la pràctica és la construcció d'un joc de dades a partir de la informació present en una pàgina web a través de *web scraping*.

Per a desenvolupar el projecte s'ha escollit la pàgina web *goodreads* com a pàgina objectiu. *Goodreads* és una web social de catalogació, referència i recomanació de llibres propietat d'Amazon i amb una extensa comunitat d'usuaris que han catalogat més de 10.000.000 llibres [1]. La quantitat i diversitat d'informació present a la seva extensa base de dades fa que sigui un bon candidat per al desenvolupament de la pràctica.

L'objectiu del projecte és, doncs, la creació d'una base de dades de llibres amb informació sobre els llibres (títol, autor, pàgines, editorial, isbn, etc.), sobre el seu contingut (gènere, protagonistes, localització, etc.), i informació provinent de la interacció social dels usuaris (puntuació, número de vots, percentatge de *likes*, etc.). A més, s'aprofitarà per recollir informació de preus dels llibres des del portal IberLibro, un mercat online de llibres propietat de AbeBooks el qual, al mateix temps, és una filial d'Amazon i pioner en el mercat online de venta de llibres [2]. S'utilitzarà la informació del ISBN extreta anteriorment de la pàgina de GoodReads per buscar el llibre en el portal IberLibro i recuperar el preu d'aquest.

Malgrat no s'ha pogut incloure la informació en aquest data set, per tal de tenir una comparativa de preus amb altres formats de llibre, s'ha desenvolupat el codi necessari extreure el preu dels llibres en format electrònic de l'aplicació Kindle d'Amazon, un lector de llibres electrònics portàtil que permet comprar, emmagatzemar i llegir llibres digitalitzats. En aquest cas la cerca es fa a través del títol i l'autor de cada llibre, informació extreta de GoodReads.

Per tal de concretar el desenvolupament del projecte s'ha decidit fer *web scraping* dels 52478 llibres de la llista *Bests Books Ever*, la més extensa del web GoodReads i que conté, segons GoodReads, els millors llibres de la història.

Tot i haver concretat el projecte en aquesta llista el codi s'ha desenvolupat de tal manera que permet a l'usuari introduir la URL de qualsevol llista de GoodReads i fer-ne l'scraping dels llibres que conté.

2. Definir un títol pel joc de dades

GoodReads Best Books Ever dataset

3. Descripció del joc de dades

El joc de dades conté 25 camps d'informació i la imatge de la portada dels 52478 llibres de la llista *Best Books Ever* de GoodReads. Val a dir que atenent a possibles conflictes amb els drets d'autor de les portades dels llibres s'ha decidit no publicar-les com a part del joc de dades. Alternativament, ates el valor afegit que considerem que aquestes aporten, s'ha afegit un camp amb la URL de la portada del llibre i s'ha desenvolupat el codi necessari per a que l'usuari pugui descarregar-les directament des de la font. Al repositori GitHub (https://github.com/scostap/goodreads_bbe_dataset) es troba la documentació necessària i un exemple de la descàrrega d'imatges.

A més de les dades obtingudes de GoodReads s'ha afegit el preu dels llibres del la botiga IberLibro. Una segona versió del joc de dades inclourà el preu dels llibres electrònics del web de Kindle d'Amazon.

4. Representació gràfica

[3]

5. Contingut

El joc de dades conté 25 atributs sobre 52478 de llibres descarregats entre el primer i el 3 de novembre de 2020 mitjançant un *web scraper* basat en *Selenium* i desenvolupat en *python*.

Els 25 camps del joc de dades són:

- **bookId:** Identificador del llibre a GoodReads.com
- **title:** Títol del llibre.
- **series:** Nom de la saga a la quan pertany el llibre.
- **author:** Autor/s del llibre.
- **rating:** Puntuació global del llibre a GoodReads.com.
- **description:** Descripció del llibre.
- **Language:** Llengua en la que s'ha publicat
- **isbn:** Referència ISBN de 13 xifres.
- **genres:** Gèneres als quals pertany el llibre, per ordre de votació.
- **characters:** Nom dels personatges principals.
- **bookFormat:** Tipus d'encuadernació.
- **edition:** Tipus d'edició (p.ex: *Anniversary Edition*).
- **pages:** Número de pàgines.
- **publisher:** Editorial.
- **publishDate:** Data de publicació.
- **firstPublishDate:** Data de publicació de la primera edició.
- **awards:** Llistat de premis.
- **numRatings:** Nombre total de valoracions a GoodReads.com.
- **ratingsByStars:** Nombre de valoracions per puntuació (5-1 estrelles).
- **likedPercent:** Percentatge de lectors als quals ha agradat el llibre, camp derivat i calculat seguint el càlcul de GoodReads (% de valoracions superiors a dos estrelles).
- **setting:** Localització de la història.

- **coverImg:** URL de la imatge de portada.
- **bbeScore:** Puntuació del llibre a la llista *Best Books Ever*
- **bbeVotes:** Nombre de vots a la llista *Best Books Ever*
- **price:** Preu del llibre a IberLibro.

6. Agraïments

Les dades han estat recollides de la pàgina web GoodReads.com per la publicació i accés a les dades dels llibres i a IberLibro.com per la informació sobre preus dels llibres.

7. Inspiració

El principal interès del joc de dades creat en aquesta pràctica és el de disposar de la informació sobre llibres present a la base de dades de GoodReads.com en un format coherent i estructurat que permeti el seu anàlisi i l'extracció efectiva de coneixement.

Les varietat d'atributs recollits permeten de fer multitud d'anàlisi, entre les quals podríem destacar:

- Identificació dels atributs que tenen més influència en la valoració dels llibres, en el nombre de vots, l'score o el preu.
- Anàlisi d'agrupament i clústering: Analitzar si els atributs recollits permeten d'agrupar els llibres en diferents grups de característiques similars i identificar i definir les característiques de cada grup.
- Anàlisis segregats per anys de publicació, per localització de la història, etc.
- Autors més prolífics o amb llibres millor valorats.
- Regressions sobre variables com la valoració dels llibres, el percentatge de gent que li agrada el llibre.
- A més considerem d'especial rellevància també l'anàlisi de les imatges de portada dels llibres, que juntament amb les variables del joc de dades permetria l'entrenament d'algorismes per al reconeixement de text a les portades (per exemple: títol, noms d'autor, editorial, etc.).
- També es podria estudiar si existeix relació entre el disseny de la portada dels llibres i el gènere al qual pertanyen, per exemple en llibres d'autoajuda.

- A més el joc de dades també permetria el desenvolupament d'una aplicació que retornés tota la informació sobre un llibre a partir del reconeixement de la imatge de portada.

- En última instància el joc de dades també permetria de desenvolupar un sistema de recomanació de llibres a partir d'una entrada de llibres llegits.

8. Llicència

La llicència que s'ha escollit per al joc de dades és la de Attribution-NonComercial 4.0 International (CC BY-NC 4.0).

https://creativecommons.org/licenses/by-nc/4.0/

Aquesta permet copiar i redistribuir el material en qualsevol mitjà i format sempre i quan es reconeixi l'autoria de manera apropiada i s'indiqui si s'ha fet alguna modificació però limita el seu ús a finalitats comercials.

Considerem que les clàusules d'aquesta llicència són les més idònies per a la publicació d'un joc de dades desenvolupat en el marc d'un treball acadèmic atès que permeten la lliure distribució de la informació i en promouen la redistribució i modificació sempre que no sigui amb una finalitat lucrativa i, per tant, d'ús lliure en l'àmbit acadèmic.

9. Codi

El codi font del programa desenvolupan el Python per a l'extracció de les dades necessàries per a la creació del job de dades es troba al repositori GitHub (https://github.com/scostap/goodreads_bbe_dataset) junt a la documentació i exemples d'ús. El codi desenvolupat por ser usat per a generar un joc de dades a partir de qualsevol llista de llibres de GoodReads.

10. Dataset

El joc de dades pot ser accedit a través del següent DOI:
http://doi.org/10.5281/zenodo.4265096

I pot ser referenciat de la següent manera:

Lorena Casanova Lozano, & Sergio Costa Planells. (2020). Best Books Ever Dataset (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4265096

11. Anàlisi de resultats

S'ha fet un anàlisi global del dataset obtingut de la informació extreta de la pàgina GoodReads.

Atributs	Completitud %
bookId	100
title	100
series	45
author	100
rating	100
description	97
language	93
isbn	92
genres	91
characters	26
bookFormat	97
edition	9
pages	96
publisher	93
publishDate	98
firstPublishDate	59
awards	20
numRatings	100
ratingsByStars	97
likedPercent	99
setting	22
coverImg	99
bbeScore	100
bbeVotes	100
price	73

12. Bibliografia i Webgrafia

[1] "Goodreads - Viquipèdia, l'enciclopèdia lliure." [Online]. Available: https://ca.wikipedia.org/wiki/Goodreads. [Accessed: 30-Oct-2020].
[2] "IberLibro.com: Información de la Empresa." [Online]. Available: https://www.iberlibro.com/docs/CompanyInformation/. [Accessed: 30-Oct-2020].
[3] Available: https://soybibliotecario.blogspot.com/2017/12/sitios-web-y-bibliotecas-libros-gratis.html
[4] H. Brody, "The Ultimate Guide To Web Scraping," 2013.
[5] B. S. Mózo, *Python Automation Cookbook*, vol. 53, no. 9. 2017.
[6] A. . Fallis, *Data Visualization with Python and Javascript*, vol. 53, no. 9. 2013.
[7] M. Heydt, *Python Web Scrapping Cookbook: Over 90 Proven Recipes to Get Your Scraping with Python, Microservices, Docker and AWS.* 2018.
[8] S. vanden Broucke and B. Baesens, *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. 2018.
[9] V. Nair, *Getting Started with Beautiful Soup*. 2014.

13. Taula de contribucions

Contribucions	**Signa**
Recerca Prèvia	Sergio Costa, Lorena Casanova
Redacció de les respostes	Sergio Costa, Lorena Casanova
Desenvolupament codi	Sergio Costa, Lorena Casanova