

Dataset: Llibres de ciència-ficció a “Casa del Libro”

Gabriel Izquierdo i Mireia Olivella

Abril 2020

Context

El conjunt de dades generat com a part d'aquesta pràctica es correspon als llibres de la secció de ciència-ficció de la web *Casa del Libro*. Entre ells ens podem trobar tot tipus de llibres: de tapa tova, de tapa dura, de butxaca i eBooks.

Representació gràfica



Figura 1: Prestatge amb llibres

Descripció

Com s'ha comentat, el contingut del conjunt de dades es correspon a un llistat de llibres procedents de la secció de ciència-ficció de la web *Casa del Libro*. L'objectiu ha sigut generar un llistat de tots aquest llibres on, per cada llibre, es crea un registre per cada un dels formats en el que es comercialitza (tapa dura, tapa tova, ...).

Contingut

Per cada llibre de la secció de ciència-ficció s'ha recollit la següent informació:

- **Title:** títol del llibre.
- **Authors:** autors del llibre.
- **Rate:** puntuació rebuda pels lectors.
- **Availability:** indica si el llibre està disponible a la web.

Si el llibre està disponible s'afegiran els atributs següents:

- **bookType:** tipus de llibre segons el format: tapa dura, tapa tova, de butxaca i eBooks.
- **Price:** preu del llibre.

Si el llibre no està disponible a la web, el tipus de llibre i el preu tindran valor “—”.

El temps de les dades és desconegut, però s'interpreta que la informació és recent, ja que la botiga està en actiu i els productes que tenen a la venda han de ser els que hi ha ara mateix en el mercat.

La web de *Casa del Libro* (<https://www.casadellibro.com/>) no s'inicia d'una càrrega, sinó que es van llençant scripts en JavaScript per tal d'emplenar la informació per parts. Aquest tipus de càrregues es dur a terme per tal de mostrar informació d'una manera àgil i no tenir temps d'espera molt elevats, ja que no es mostraria res fins que tota la informació estigués carregada.

Per tal d'obtenir les dades d'una web amb aquest tipus de càrrega s'ha procedit a utilitzar Selenium, ja que proporciona les eines necessàries per dur a terme amb èxit aquesta tasca.

Agraïments

Les dades han estat recollides de la pròpia web *Casa del Libro* (secció *Ciencia Ficción*): <https://www.casadellibro.com/libros/literatura/narrativa-en-bolsillo/ciencia-ficcion-en-bolsillo/121005001/p>

Hem utilitzat el llenguatge de programació Python i tècniques de *web Scraping* utilitzant Selenium per tal de poder extreure la informació.

Inspiració

Aquest conjunt de dades es pot utilitzar en molts àmbits, però sobretot està enfocat al camp de la mineria de dades. El conjunt de dades només està enfocat a la secció de ciència ficció de la web, però com totes les seccions tenen el mateix format es pot aplicar el mateix script canviant la url a la dessitjada.

Això pot servir per obtenir llistats amb tota la informació de cada llibre de la web. Amb aquesta informació es poden crear sistemes recomanadors on, segons els gustos d'una persona aquest li pugui recomanar llibres procedents de la web *Casa del Libro*.

Llicència



La llicència escollida pel conjunt de dades ha sigut CC BY-SA 4.0 License. Els motius de l'elecció són els següents:

- S'ha de mantenir el nom del creador del conjunt de dades, i especificar els canvis fets respecte a la versió original.
- Es permet un ús comercial. Això permetria que una empresa utilitzés les dades generades per crear projectes que reconeguessin l'autor original.
- Les contribucions es distribuiran segons els paràmetres que el propi autor hagi plantejat.

Codi

El codi per poder obtenir el dataset es pot trobar trobar accedint al següent enllaç de GitHub: <https://github.com/rascundampelcuf/Web-scraping>

Dataset

El dataset pot ser consultat a la següent adreça de Zenodo: <https://zenodo.org/record/3749018#.XpMd8pntZPY>

Recursos

1. Subirats Maté, L., Calvo González, M. (2019). Web scraping. Oberta UOC Publishing, SL.
2. Selenium-python.readthedocs.io. 2020. Selenium With Python — Selenium Python Bindings 2 Documentation. [online] Available at: <https://selenium-python.readthedocs.io> [Accessed 9 April 2020].