PRÁCTICA 1: Web scraping

Tipología y ciclo de vida de los datos

Máster en Ciencia de Datos

Daniel Laureano Cerviño Cortínez
Juan Kevin Trujillo Rodríguez

Universitat Oberta de Catalunya

Tabla de contenido

Preguntas de la práctica 1 1

Contribuciones al trabajo 4

Bibliografía 4

Preguntas de la práctica 1

Esta práctica consiste en la creación de un dataset a partir de los datos contenidos en una web. Con ella cumplimos los siguientes puntos:

1. Contexto

Decidimos utilizar la página web (Toscrape.com, n.d.) para realizar scrape porque los datos que contienen no están sujetos a ninguna licencia y no tiene ninguna licencia que nos impida utilizarlos. Esta web se trata de una página de prueba para ejecutar este tipo de técnicas. Además, hemos seleccionado esta debido a su naturaleza similar a un e-commerce. Cabe destacar que las páginas web centradas en el comercio electrónico son muy utilizadas para conocer los distintos precios de los productos. En este caso, podríamos utilizar este proyecto, para implementarlo en cualquier página web similar (siempre teniendo en cuenta las diferencias existentes).

2. Título del dataset

Books to Scrape

3. Descripción del dataset

De la web (Toscrape.com, n.d.) se han extraído todos los datos de los 1.000 libros disponibles. De estos se han extraído los datos relativos expuestos en la página, tales como, el título, el precio, el código universal del libro, etc. La página web disponible de un conjunto de 50 categorías entre las cuales se distribuyen todos los libros.

4. Representación gráfica

Title	ProductType	Category	Rating	Product Description	UPC	Price Excl. Tax	Price Incl.Tax	Tax	Avaliability	Number of Reviews	image
It's Only the H	Books	Travel	Two	"Wherever you go, w	a2212	£45.17	£45.17	£0.00	In stock (19 available)	0	http://books.toscrape.c
Full Moon ove	Books	Travel	Four	Acclaimed travel writ	ce604	£49.43	£49.43	£0.00	In stock (15 available)	0	http://books.toscrape.c
See America:	Books	Travel	Three	To coincide with the 2	f9705c	£48.87	£48.87	£0.00	In stock (14 available)	0	http://books.toscrape.c
Vagabonding:	Books	Travel	Two	With a new foreword	18092	£36.94	£36.94	£0.00	In stock (8 available)	0	http://books.toscrape.c
Under the Tus	Books	Travel	Three	A CLASSIC FROM THE	a9435(	£37.33	£37.33	£0.00	In stock (7 available)	0	http://books.toscrape.c
A Summer In I	Books	Travel	Two	On her thirtieth birth	cc193(	£44.34	£44.34	£0.00	In stock (7 available)	0	http://books.toscrape.c
The Great Rai	Books	Travel	One	First published more t	48736c	£30.54	£30.54	£0.00	In stock (6 available)	0	http://books.toscrape.c
A Year in Prov	Books	Travel	Four	National BestsellerIn	9e609	£56.88	£56.88	£0.00	In stock (6 available)	0	http://books.toscrape.c
The Road to L	Books	Travel	One	The hilarious and lovi	366a2	£23.21	£23.21	£0.00	In stock (3 available)	0	http://books.toscrape.c
Neither Here	Books	Travel	Three	Bill Bryson's first trav	747cf7	£38.95	£38.95	£0.00	In stock (3 available)	0	http://books.toscrape.c
1,000 Places t	Books	Travel	Five	Around the World, co	228ba5	£26.08	£26.08	£0.00	In stock (1 available)	0	http://books.toscrape.c
Sharp Objects	Books	Mystery	Four	WICKED above her hi	e00eb4	£47.82	£47.82	£0.00	In stock (20 available)	0	http://books.toscrape.c
In a Dark, Dar	Books	Mystery	One	In a dark, dark wood I	19ed2	£19.63	£19.63	£0.00	In stock (18 available)	0	http://books.toscrape.c
The Past Neve	Books	Mystery	Four	A simple task, Attorne	5ee945	£56.50	£56.50	£0.00	In stock (16 available)	0	http://books.toscrape.c
A Murder in Ti	Books	Mystery	One	Beautiful and brilliant	f733e8	£16.64	£16.64	£0.00	In stock (16 available)	0	http://books.toscrape.c
The Murder o	Books	Mystery	Four	In the village of King's	c7b51	£44.10	£44.10	£0.00	In stock (15 available)	0	http://books.toscrape.c
The Last Mile	Books	Mystery	Two	In his #1 New York Ti	3bc89	£54.21	£54.21	£0.00	In stock (14 available)	0	http://books.toscrape.c
That Darkness	Books	Mystery	One	As a forensic investig	0c7b9c	£13.92	£13.92	£0.00	In stock (14 available)	0	http://books.toscrape.c
Tastes Like Fe	Books	Mystery	One	Sarah Hilary won the	2d1e3	£10.69	£10.69	£0.00	In stock (14 available)	0	http://books.toscrape.c
A Time of Tor	Books	Mystery	Five	Jerome Burnel was or	4416c4	£48.35	£48.35	£0.00	In stock (14 available)	0	http://books.toscrape.c
A Study in Sca	Books	Mystery	Two	In the debut of literat	63ee5	£16.73	£16.73	£0.00	In stock (14 available)	0	http://books.toscrape.c
Poisonous (M	Books	Mystery	Three	Teen-aged Internet b	abdd91	£26.80	£26.80	£0.00	In stock (12 available)	0	http://books.toscrape.c
Murder at the	Books	Mystery	Four	Murder at the 42nd St	e7fe3b	£54.36	£54.36	£0.00	In stock (12 available)	0	http://books.toscrape.c
Most Wanted	Books	Mystery	Three	Lisa Scottoline deliver	c039f5	£35.28	£35.28	£0.00	In stock (12 available)	0	http://books.toscrape.c

5. Contenido

Los datos de los libros que fueron introducidos en un fichero CSV el 6 de noviembre de 2020, los podemos desglosar en los siguientes atributos:

- ***Title***: Se trata de un campo de tipo texto que hace referencia al nombre del producto.
- ***Image***: Este atributo está compuesto por los enlaces a las imágenes.
- ***Rating***: Consisten en las valoraciones realizadas por los usuarios en la web, este campo de tipo texto está compuesto por los valores entre “One”, “Two”, “Three”, “Four” y “Five”.
- ***Description***: Atributo de tipo texto que se basa en una descripción de cada uno de los productos.
- ***Category***: Campo de tipo texto que hace referencia a la categoría de cada producto.
- ***UPC***: (*Universal Product Code*). Se trata del número identificativo de cada producto, en formato texto que está compuesto tanto por números como por letras, por ejemplo, “be5cc846f45496fb”.
- ***Product type***: Hace referencia al tipo de producto expuesto en la web. En nuestro caso se trataría de un campo textual compuesto por la palabra “Books”.
- ***Price excl. tax***: Atributo basado en el precio sin impuestos, de tipo texto porque también se extrajo el tipo de divisa de la página web, por ejemplo, “£13.12”.
- ***Price incl. tax***: Atributo basado en el precio incluyendo impuestos, de tipo texto porque también se extrajo el tipo de divisa de la página web, por ejemplo, “£13.12”.
- ***Tax***: Tipo de impuesto asociado al producto, de tipo texto que, ya que incluye el tipo de divisa junto con la cantidad, por ejemplo “£0.00”.
- ***Availability***: Es un atributo de tipo texto donde se encuentra el número de artículos disponibles para ese producto. Un ejemplo de este atributo es el siguiente: “In stock (19 available)”.
- ***Number reviews***: Atributo compuesto por el número de críticas realizadas por los consumidores de este producto. Se basaría en un campo de tipo numérico y entero.

Antes de incluir los datos en el CSV fueron entrecomillados para convertirlos a un formato de texto (*string*) para un mejor manejo y porque los datos están separados por comas “,” así evitamos que en caso de los textos tuvieran comas, no cambien la estructura del dataset, añadiendo más columnas de las que realmente tiene, que son 12.

Es importante destacar que, para la realización de esta técnica, la página web seleccionada no posee un fichero robots.txt ni una API para extraer los datos de la web. Esto es debido a que esta se realizó con fines didácticos para la práctica de técnicas de web scraping.

No tenemos un periodo de tiempo definido porque se ha utilizado una página web con libros de prueba para practicar web scraping.

6. Agradecimientos

Como agradecimiento principal sería para el propietario de la página web (Toscrape.com, n.d.) y así mismo el de los datos, pero, desconocemos su identidad ya que tras realizar una búsqueda en la web y en internet no hemos tenido accesible dicha información relativa a la propia de la web.

Existen otros análisis previos que realizar el web scraping de esta página en concreto, podemos destacar los siguientes:

- (Hong Khai, 2019)
- (Oheix, 2018)

Por otra parte, a los autores de los siguientes recursos que hemos utilizado:

- (Subirats Maté & Calvo González, 2019)
- (Lawson, 2015)

7. Inspiración

Nuestra principal inspiración y objetivo era encontrar una web con datos sin una licencia copyright, que nos permitiera extraer los datos que necesitáramos y poder trabajar con los mismos. Los datos extraídos no tienen información relevante debido a que se trata de una página web para practicar web scraping con datos de carácter ficticio, pero se aporta un ejemplo de como utilizar esta técnica con el lenguaje de programación Python.

8. Licencia

Hemos optado por una licencia ***Released Under CC0: Public Domain License***, dado que los datos son ficticios y queremos apoyar el open source. Debido a que nos gustaría que todo el mundo tuviera acceso a este tipo de técnicas con un ejemplo en concreto y que funcione como herramienta didáctica para cualquier persona que quiera practicar web scraping. Además, éticamente consideramos que establecer cualquier otro tipo de licencia a este proyecto no sería adecuado debido a que el web scraping se le realizó a una página web generada para practicar estas técnicas.

9. Código

El código fuente está disponible en nuestro repositorio de GitHub (Cerviño Cortínez & Trujillo Rodríguez, 2020a) .

10. Dataset

El dataset fue registrado en Zenodo (European Organization for Nuclear Research, n.d.) donde obtuvimos el DOI: ***10.5281/zenodo.4263215***, y que se puede consultar en la página web del DOI (Cerviño Cortínez & Trujillo Rodríguez, 2020b).

Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	DLCC, JKTR
Redacción de las respuestas	DLCC, JKTR
Desarrollo código	DLCC, JKTR

Bibliografía

Cerviño Cortínez, D. L., & Trujillo Rodríguez, J. K. (2020a). *Web Scraping of Books to Scrape*. https://github.com/Danielcc97/WebScraping

Cerviño Cortínez, D. L., & Trujillo Rodríguez, J. K. (2020b). *Books to Scrape*. https://doi.org/10.5281/ZENODO.4263215

European Organization for Nuclear Research. (n.d.). *Zenodo*. Retrieved 8 November 2020, from https://zenodo.org/

Hong Khai, T. (2019, December 18). *Extract Transform Load (ETL) for Books to Scrape*. https://medium.com/analytics-vidhya/extract-transform-load-etl-for-books-to-scrape-b0ff5f83095d

Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd.

Oheix, J. (2018, December 11). *An introduction to web scraping with Python*. https://towardsdatascience.com/an-introduction-to-web-scraping-with-python-a2601e8619e5

Subirats Maté, L., & Calvo González, M. (2019). *Web scraping*. Editorial UOC.

Toscrape.com. (n.d.). *Books to Scrape*. Retrieved 7 November 2020, from https://books.toscrape.com/