# Solución PRA 1 – Tipología y ciclo de vida de los datos

Arturo Hernández Sánchez

Laia Cebey Ripoll

*Máster Universitario en Ciencia de Datos*

## Tabla de contenidos



### 1. Contexto

Para esta práctica, hemos elegido la web https://books.toscrape.com/ , la cual se trata de una web ficticia con un catálogo de libros, perfecta para practicar *Web Scraping*, ya que no es necesario pedir permisos y demás, al tratarse de una web pensada para practicar y aprender con las distintas técnicas de *Web Scraping*.

En esta web, tenemos en la pantalla inicial, una cuadrícula con un catálogo de libros, con la posibilidad de filtrar por género. Al hacer clic en cada libro, nos aparecen datos como precio, el *rating* del libro, el stock disponible, o un código que identifica de manera unívoca cada libro. Nuestro objetivo es extraer estas variables numéricas, para ver si existe una cierta correlación entre ellas y extraer conclusiones relevantes para el negocio.

Esta idea, podría ser extrapolable a cualquier web de compras, ya que son variables que suelen aparecer en cualquier tienda que venda productos físicos.

### 2. Título

El título del dataset propuesto es:

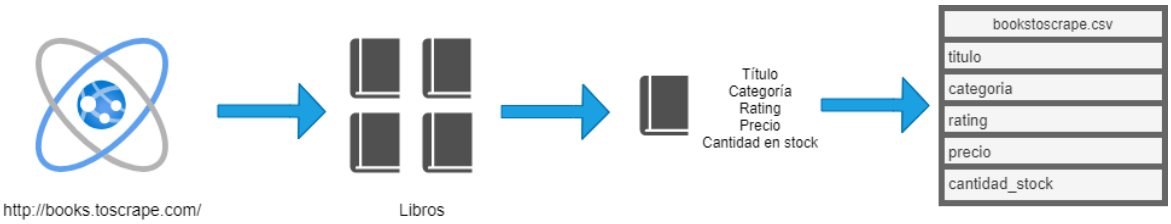
"Datos relevantes del catálogo de los libros presentes en la web *Books to Scrape*".

### 3. Descripción del *dataset*

Como se ha indicado anteriormente, este *dataset* constará con 5 variables que describen ciertos datos de cada libro de la web *Books to Scrape*. Dos de estas variables serán cualitativas, y servirán para identificar cada uno de los libros, mientras que las otras tres, serán variables cuantitativas, las cuales serán el objeto de análisis para estudiar su dependencia.

## 4. Representación gráfica

En la página web seleccionada, encontramos un listado con información de 1000 libros (divididos en 50 páginas que muestran 20 libros). Si seleccionamos cada libro, se obtiene más información sobre él, concretamente el título, la categoría a la que pertenece, el rating, el precio y la cantidad en stock. Se guarda esta información de cada libro en un csv llamado bookstoscrape.




## 5. Contenido

El *dataset* cuenta con un total de 1000 observaciones que hacen referencia a cada uno de libros presentes en el catálogo de la web *Books to Scrape*, y 5 variables, las cuales son:

- *titulo* – Título del libro.
- *categoría* – Categoría de género a la que pertenece el libro.
- *rating* – Calificación del libro otorgada por los lectores del mismo.
- *precio* - Precio del libro.
- *cantidad_stock* – Número de ejemplares del libro disponibles en el almacén en el momento de la extracción *del dataset*.

Estos datos se han recogido del sitio web *Books to Scrape* mediante *Python*, haciendo uso de las librerías *requests y BeautifulSoup*.

Para obtener los datos, se ha iterado sobre las 50 páginas de la página web en las que hay 20 libros. Cada página tiene la estructura 'http://books.toscrape.com/catalogue/page-{}.html' donde '{}' es el número de página, es decir un número de 1 a 51. Una vez dentro de cada página, se han realizado llamadas a cada una de las urls correspondientes a cada libro para extraer sus datos y almacenarlos en cada una de las variables del *dataset*.

Además de este primer enfoque, se ha querido utilizar alguna librería más avanzada, como es el caso de selenium para obtener el mismo dataset. Se ha seguido el mismo enfoque, pero en este caso sin especificar cada url, simplemente utilizando el atributo "next" del código html. Además, se han añadido pausas entre las peticiones a la página web para simular el caso de una página web que limite las peticiones que se pueden hacer en un periodo de tiempo.

## 6. Agradecimientos

No se conoce el o los autores de esta página web. En el código fuente se puede ver el nombre de Óscar, así que se supone que la página la ha creado alguien llamado Óscar.

```
<!-- /container-fluid -->
<footer class="footer container-fluid">…</footer>
<!-- jQuery -->
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.9.1/jquery.min.j
s"></script>
<script>…</script>
<script src="static/oscar/js/jquery/jquery-1.9.1.min.js"></script>
<!-- Twitter Bootstrap -->
<script type="text/javascript" src="static/oscar/js/bootstrap3/bootstrap.mi
n.js"></script>
<!-- Oscar -->
<script src="static/oscar/js/oscar/ui.js" type="text/javascript" charset="u
tf-8"></script>
<script src="static/oscar/js/bootstrap-datetimepicker/bootstrap-datetimepic
ker.js" type="text/javascript" charset="utf-8"></script>
<script src="static/oscar/js/bootstrap-datetimepicker/locales/bootstrap-dat
etimepicker.all.js" type="text/javascript" charset="utf-8"></script>
<script type="text/javascript">
            $(function() {


    oscar.init();

    oscar.search.init();

        });
```

Respecto al análisis legal de la página, al ser una web con un catálogo de libros ficticio precisamente desarrollada para que personas que están aprendiendo *Web Scraping* puedan practicar, tal como se ve en este mensaje al abrir la web:

> **Warning!** This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.

El uso académico sin fines lucrativos, como es el caso de esta práctica, es totalmente legítimo.

## 7. Inspiración

Como se ha indicado en apartado anteriores, una de las preguntas que puede responder el análisis de este *dataset*, es el estudio de la correlación de las variables cuantitativas que describen estos libros.

Por otro lado, también sería posible utilizar algún método de análisis no supervisado para separar los libros en distintas categorías por precio o *rating* por ejemplo.

En caso de existir cierta correlación de variables, podríamos ser capaces de predecir por ejemplo el éxito que tendrá el libro, atendiendo a las predicciones obtenidas para la variable *rating*.

## 8. Licencia

Elegimos la licencia *Released Under CC BY-NC-SA 4.0 License*. Se trata de una licencia *Creative Commmons*. Elegimos esta por los siguientes motivos:

- *BY* - El beneficiario de la licencia podrá hacer uso para copiar, distribuir y mostrar el trabajo, siempre y cuando se cite al autor del mismo. De esta manera, se podrá sacar el mayor partido al trabajo realizado en esta práctica sin perder el reconocimiento del autor original.
- *NC* – El beneficiario podrá hacer el uso de la práctica que se ha indicado en el apartado anterior, siempre y cuando no se realice con fines comerciales o lucrativos. Al ser una práctica basada en una web de uso libre con fines académicos y para democratizar el acceso a material académico, consideramos preferible evitar que alguien se lucre por ello.
- *SA* – Todas las obras que deriven de esta práctica se podrán distribuir siempre y cuando se mantenga esta misma licencia. De esta manera, nos aseguramos que los dos puntos anteriores se mantengan en las obras derivadas.
  *(Fuente: https://es.wikipedia.org/wiki/Licencias_Creative_Commons)*

## 9. Código

El código de este proyecto se puede encontrar en el siguiente repositorio de github. Se pueden encontrar dos archivos .py con dos formas distintas de obtener el código. El primero PAC1.py con el código del primer enfoque (librerías requests y beautifulSoup) y el segundo PAC1_selenium.py con el código del segundo enfoque (librería selenium).

## 10. *Dataset*

El dataset publicado en formato CSV en Zenodo se puede encontrar aquí.

| Contribuciones | Firma |
|---|---|
| Investigación previa | *Arturo Hernández Sánchez, Laia Cebey Ripoll* |
| Redacción de respuestas | *Arturo Hernández Sánchez, Laia Cebey Ripoll* |
| Desarrollo código | *Arturo Hernández Sánchez, Laia Cebey Ripoll* |