

## **Práctica 1. Web Scrapping**

Asignatura Tipología y ciclo de vida de los  
datos-Aula 2

Profesor: Diego Pérez

Alumno: Juan Francisco Vallalta Rueda

## Contexto

Un cliente interesado en abrir una librería online de divulgación científica nos solicita el siguiente conjunto de datos:

- Un dataset con el título, materia, ISBN, sinopsis, formato y precio de los libros de divulgación científica en física disponibles en el mercado español para cargarlos como productos en su tienda online.

## Título de dataset

El título del dataset es: Libros de física.

## Descripción del dataset

El dataset está formado por la relación de libros de física que se encuentran disponibles para la venta en la librería online de Amazon.es.

## Representación gráfica del dataset


El dataset se representa visualmente por los resultados de la búsqueda en la web de Amazon.es de todos los libros de la materia física. Se muestra para cada libro, su título, autor, formato y precio.

1-16 de más de 40.000 resultados para Libros : Ciencias, tecnología y medicina : Física

Más vendido



**Física para la ciencia y la tecnología, Vol. 1: Mecánica, oscilaciones y ondas, termodinámica, 6ª Edición** 1 enero 2010  
de Paul Allen Tipler y Gene Mosca

Tapa blanda  
**EUR 37,00**   
Recíbelo Mañana, oct 31  
Envío GRATIS disponible  
Sólo hay 1 en stock. Cómpralo cuanto antes.  
Más opciones de compra  
**EUR 29,00** usado y nuevo (19 ofertas)

★★★★★ ~ 41



**FISICA Y QUIMICA 1 BACHILLERATO - 9788448191542** 31 marzo 2015  
de Rodríguez y McGraw - Hill

Tapa blanda  
**EUR 36,05**   
Recíbelo Mañana, oct 31  
Envío GRATIS disponible

★★★★★ ~ 7



**LA+SB Física y Química 3 ESO** 15 junio 2015  
de Pozos Magariños, Antonio y Yuste Muñoz, Migue Angel

Tapa blanda  
**EUR 33,20**   
Recíbelo Mañana, oct 31  
Envío GRATIS disponible  
Más opciones de compra  
**EUR 14,11** usado y nuevo (11 ofertas)

★★★★★ ~ 8

Más vendido



**Las paradojas cuánticas: Schrödinger y la mecánica ondulatoria (NATGEO CIENCIAS)** 25 noviembre 2017  
de David Blanco Laserna

Versión Kindle  
**EUR 0,00**   
**EUR 4,74** por la compra

Tapa dura  
**EUR 14,25**   
Envío GRATIS disponible



**Breves respuestas a las grandes preguntas (Fuera de Colección)** 30 octubre 2018  
de Stephen Hawking y David Jos Mraabent

Tapa dura  
**EUR 17,00**   
Recíbelo Mañana, oct 31  
Envío GRATIS disponible  
Más opciones de compra  
**EUR 14,00** usado y nuevo (18 ofertas)

★★★★★ ~ 55

Versión Kindle  
**EUR 12,34**

Otros formatos: Encuadernación desconocida

## Contenido:

El dataset se obtiene a través de la web de Amazon buscando todos los libros de la materia física. Se muestran según la siguiente URL (30/10/2019 06:04):

[https://www.amazon.es/s/ref=lp\\_902503031\\_nr\\_n\\_4?fst=as%3Aoff&rh=n%3A599364031%2Cn%3A21599365031%2Cn%3A902503031%2Cn%3A902508031&bbn=902503031&ie=UTF8&qid=1572411363&rnid=902503031](https://www.amazon.es/s/ref=lp_902503031_nr_n_4?fst=as%3Aoff&rh=n%3A599364031%2Cn%3A21599365031%2Cn%3A902503031%2Cn%3A902508031&bbn=902503031&ie=UTF8&qid=1572411363&rnid=902503031)

El dataset contiene los siguientes campos:

- Título
- Autor
- Precio
- Formato
- Sinopsis
- Páginas
- Editorial
- Idioma
- ISBN-13

Los datos se extraen en una fecha y hora concreta (11/11/2019 22:18) y corresponden a la situación del catálogo de Amazon.es en ese instante del tiempo.

### **Agradecimientos**

El propietario de los datos es Amazon.es, librería online líder en el mercado nacional. Se caracteriza por disponer en stock de todos los libros disponibles para la venta.

### **Inspiración**

Este conjunto de datos es interesante para cualquier librería online pues le permite disponer de un catálogo de todos los libros en venta en España para una determinada materia, evitando la laboriosa carga inicial de productos en la web.

### **Licencia**

El dataset se distribuye bajo licencia CC BY-NC-SA 4.0. Esta licencia le permite al usuario del dataset a copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y cite la obra de la forma especificada por el autor o el licenciante para fines no comerciales. Las obras derivadas se deben compartir bajo una licencia idéntica.

### Código

```
import requests as rq
import string
from bs4 import BeautifulSoup
import pandas as pd
import time

#Cabecera navegador
cabecera = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/77.0.3865.90 Safari/537.36"}

#Url sitio scraping
url_base = "https://www.amazon.es"
url_base_dp = url_base + "/dp/"

#Página inicial con el resultado de la búsqueda de libros de fisica
url_inicial="https://www.amazon.es/s/ref=lp_902503031_nr_n_4?fst=as%3Aoff&rh=n%3A599364031%2Cn%3A%21599365031%2Cn%3A902503031%2Cn%3A902508031&bbn=902503031&ie=UTF8&qid=1571919270&rnid=902503031"

#Abrimos sesion
sesion=rq.Session()
sesion.post(url_base, headers=cabecera)

#Accedemos a las página inicial
pagina = sesion.get(url_inicial, headers=cabecera)
soup = BeautifulSoup(pagina.content)
num_pagina = 1

#Obtenemos número máximo de páginas de resultados
texto_nmax =soup.find(id="pagn")
contador_max = texto_nmax.find_all('span')
pagina_max = int(contador_max[7].get_text())

#Creamos dataframe libros_df
libros_df = pd.DataFrame(columns=('titulo', 'precio', 'sinopsis', 'formato', 'editor','coleccion', 'idioma', 'isbn'))

#Función que extrae los atributos del libro que nos interesan
def datos_libro(arg):
    #Retardamos peticiones
    t0 = time.time()
```

```

libro = sesion.get(arg, headers=cabecera)
response_delay = time.time() - t0
time.sleep(10 * response_delay)
#Accedemos al contenido de la página
soup_libro = BeautifulSoup(libro.content)
reg_libro = []
#Accedemos al titulo
reg_libro.append(soup_libro.title.get_text().strip())
#Accedemos al precio
precio_libro = soup_libro.find(id="buyNewSection")
if precio_libro == None:
    reg_libro.append("")
else:
    reg_libro.append(precio_libro.get_text().strip("\n"))
#Accedemos a la sinopsis
sinopsis=soup_libro.find(id="bookDescription_feature_div")
if sinopsis == None:
    reg_libro.append("")
else:
    reg_libro.append(sinopsis.div)
#Accedemos detalle libro
detalle_libro=soup_libro.find(id="detail_bullets_id")
if detalle_libro == None:
    reg_libro.append(["", "", "", "", ""])
else:
    detalle = detalle_libro.find_all('li')
    formato = detalle[0].get_text().strip()
    reg_libro.append(formato)
    editor = detalle[1].get_text().strip()
    reg_libro.append(editor)
    coleccion=detalle[2].get_text().strip()
    reg_libro.append(coleccion)
    idioma = detalle[3].get_text().strip()
    reg_libro.append(idioma)
    isbn = detalle[5].get_text().strip()
    reg_libro.append(isbn)
return reg_libro

```

```

#Recorremos las paginas resultados
for paginas in range(pagina_max):

```

```

    if num_pagina == 1:
        resultado = soup.find(id="mainResults")

```

```

for tag_li in resultado.find_all('li'):
    asin = tag_li.get('data-asin')
    #Edicion física
    if asin[0] != 'B':
        url = url_base_dp+asin+"/"
        libro_amazon = datos_libro(url)
        libros_df.loc[len(libros_df)]=libro_amazon

url_pag_sig = soup.find(id="pagnNextLink")
url_pag_sig = url_base + url_pag_sig.get("href")
num_pagina = num_pagina + 1
else:
    pagina = sesion.get(url_pag_sig, headers=cabecera)
    soup = BeautifulSoup(pagina.content)
    tag_div = soup.find_all('div')
    for item in tag_div:
        if item.has_attr('data-asin'):
            asin = item.get('data-asin')
            #Edicion física
            if asin[0] != 'B':
                url = url_base_dp+asin+"/"
                libro_amazon = datos_libro(url)
                libros_df.loc[len(libros_df)]=libro_amazon
if num_pagina < pagina_max:
    partes_url = url_pag_sig.split("&")
    url_0 = partes_url[0]
    pagina = 'page='+ str(num_pagina+1)
    url_2 = partes_url[2]
    ref = 'ref=lp_902508031_pg_' + str(num_pagina)
    url_pag_sig = url_0 + '&' + pagina + '&' + url_2 + '&' + ref
    num_pagina = num_pagina + 1

```

## Dataset

librosFisicaAmazon.csv

Contribuciones	Firma
Investigación previa	<b>Juan Francisco Vallalta Rueda</b>
Redacción de las propuestas	<b>Juan Francisco Vallalta Rueda</b>
Desarrollo código	<b>Juan Francisco Vallalta Rueda</b>