

## **Práctica 1. Web Scrapping**

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

Profesor: Diego Péres

Alumno: Juan Francisco Vallalta Rueda

## Contexto

Un cliente interesado en abrir una librería online de divulgación científica nos solicita los siguientes conjuntos de datos:

- Un dataset con el título, materia, ISBN, sinopsis e imagen 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 imagen, 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 Mrazbert

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
- Imagen

Los datos se extraen en una fecha y hora concreta 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

En preparación.

### Código

En preparación

```
import requests as rq  
from bs4 import BeautifulSoup
```

```
#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 física
```

```
url_inicial="https://www.amazon.es/s/ref=lp_902503031_nr_n_4?fst=as%3Aoff&rh=n%3A59  
9364031%2Cn%3A%21599365031%2Cn%3A902503031%2Cn%3A902508031&bbn=9025  
03031&ie=UTF8&qid=1571919270&rnid=902503031"
```

```
#Abrimos sesion
```

```
sesion=rq.Session()
```

```
sesion.post(url_base, headers=cabecera)
```

```
#Accedemos a las páginas
```

```
pagina = sesion.get(url_inicial, headers=cabecera)
```

```
soup = BeautifulSoup(pagina.content)
```

```
contador_pagina = 1
```

```
#Obtenemos número máximo de páginas
```

```
texto_nmax =soup.find(id="pagn")
```

```
contador_max = texto_nmax.find_all('span')
```

```
pagina_max = int(contador_max[7].get_text())
```

```
#range(pagina_max)
```

```
#Obtener las urls
```

```
url = []
```

```
#for contador in range
```

```
#Recorremos la página
```

```
resultado = soup.find(id="mainResults")
```

```
for tag_li in resultado.find_all('li'):
```

```
    asin = tag_li.get('data-asin')
```

```
    #Edicion digital
```

```
    if asin[0]=='B':
```

```
        #Busqueda digital
```

```
        print("Digital")
```

```
    else:
```

```
        url = url_base_dp+asin+"/"
```

```
        print(url)
```

```
        libro = sesion.get(url, headers=cabecera)
```

```
        soup_libro = BeautifulSoup(libro.content)
```

```
        reg_libro = []
```

```
        #Accedemos al titulo
```

```
        reg_libro.append(soup_libro.title.get_text())
```

```
#Accedemos al precio
precio_libro = soup_libro.find(id="buyNewSection")
reg_libro.append(precio_libro.get_text())
print(reg_libro)
```

```
#Accedemos url libro
#pagina = sesion.get(url3, headers=cabecera)
#soup = BeautifulSoup(pagina.content)
```

```
#Accedemos al titulo
#titulo = soup.title.get_text()
```

```
#Accedemos al precio
#precio_libro = soup.find(id="buyNewSection")
#precio = precio_libro.get_text()
```

```
#Accedemos al detalle del libro
#detalle_libro=soup.find(id="detail_bullets_id")
#detalle = detalle_libro.find_all('li')
#formato = detalle[0].get_text()
#editor = detalle[1].get_text()
#coleccion=detalle[2].get_text()
#idioma = detalle[3].get_text()
#isbn = detalle[5].get_text()
```

```
#Formateamos los datos
#print(titulo)
#print(formato)
#print(editor)
#print(coleccion)
#print(idioma)
#print(isbn)
#print(precio)
#Scraping libro
```

## Dataset

En preparación.

## Evaluación inicial

La URL que nos interesa es

[https://www.amazon.es/s?i=stripbooks&bbn=902503031&rh=n%3A599364031%2Cn%3A%21599365031%2Cn%3A902503031%2Cn%3A902508031%2Cp\\_6%3AA1AT7YVPFBWXBL%2Cp\\_n\\_availability%3A831278031&lo=image&pf\\_rd\\_i=902503031&pf\\_rd\\_m=A1AT7YVPFBWXBL&pf\\_rd\\_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf\\_rd\\_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf\\_rd\\_r=ETE9H1GNGT56MXNCVXZ2&pf\\_rd\\_r=ETE9H1GNGT56MXNCVXZ2&pf\\_rd\\_s=merchandised-search-1&pf\\_rd\\_t=101&ref=amb\\_link\\_5](https://www.amazon.es/s?i=stripbooks&bbn=902503031&rh=n%3A599364031%2Cn%3A%21599365031%2Cn%3A902503031%2Cn%3A902508031%2Cp_6%3AA1AT7YVPFBWXBL%2Cp_n_availability%3A831278031&lo=image&pf_rd_i=902503031&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf_rd_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf_rd_r=ETE9H1GNGT56MXNCVXZ2&pf_rd_r=ETE9H1GNGT56MXNCVXZ2&pf_rd_s=merchandised-search-1&pf_rd_t=101&ref=amb_link_5)

Esta URL nos muestra la relación de libros de la materia física disponibles en el catálogo de Amazon. Para cada libro existe una ficha de producto que contiene la información que nos interesa:

[https://www.amazon.es/F%C3%ADsica-para-ciencia-tecnolog%C3%ADa-Vol/dp/8429144293/ref=sr\\_1\\_1?m=A1AT7YVPFBWXBL&pf\\_rd\\_i=902503031&pf\\_rd\\_m=A1AT7YVPFBWXBL&pf\\_rd\\_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf\\_rd\\_r=ETE9H1GNGT56MXNCVXZ2&pf\\_rd\\_s=merchandised-search-1&pf\\_rd\\_t=101&qid=1571714645&refinements=p\\_6%3AA1AT7YVPFBWXBL%2Cp\\_n\\_availability%3A831278031&s=books&sr=1-1](https://www.amazon.es/F%C3%ADsica-para-ciencia-tecnolog%C3%ADa-Vol/dp/8429144293/ref=sr_1_1?m=A1AT7YVPFBWXBL&pf_rd_i=902503031&pf_rd_m=A1AT7YVPFBWXBL&pf_rd_p=112d02d7-2c00-40a4-a4ea-fee637dd6484&pf_rd_r=ETE9H1GNGT56MXNCVXZ2&pf_rd_s=merchandised-search-1&pf_rd_t=101&qid=1571714645&refinements=p_6%3AA1AT7YVPFBWXBL%2Cp_n_availability%3A831278031&s=books&sr=1-1)

Título

Autor

Precio

Sinopsis

Páginas

Editorial

Idioma

ISBN-13

Imagen

Otros libros que compraron los clientes

#### 1. Archivo robots.txt

Accedemos <https://www.amazon.es/robots.txt> y se obtiene el siguiente resultado:

```
User-agent: *
Disallow: /dp/product-availability/
Disallow: /dp/rate-this-item/
Disallow: /exec/obidos/account-access-login
Disallow: /exec/obidos/change-style
Disallow: /exec/obidos/dt/assoc/handle-buy-box
Disallow: /exec/obidos/flex-sign-in
Disallow: /exec/obidos/handle-buy-box
Disallow: /exec/obidos/refer-a-friend-login
```

```
Disallow: /exec/obidos/subst/associates/join
Disallow: /exec/obidos/subst/marketplace/sell-your-collection.html
Disallow: /exec/obidos/subst/marketplace/sell-your-stuff.html
Disallow: /exec/obidos/subst/partners/friends/access.html
Disallow: /exec/obidos/tg/cm/member/
Disallow: /gp/cart
Disallow: /gp/content-form
Disallow: /gp/customer-images
Disallow: /gp/customer-media/upload
Disallow: /gp/customer-reviews/common/du
Disallow: /gp/customer-reviews/write-a-review.html
Disallow: /gp/flex
Disallow: /gp/gfix
Disallow: /gp/history
Disallow: /gp/item-dispatch
Disallow: /gp/legacy-handle-buy-box.html
Disallow: /gp/reader
Disallow: /gp/registry/wishlist/*/reserve
Disallow: /gp/richpub/listmania/createpipeline
Disallow: /gp/music/clipserve
Disallow: /gp/recsradio
Disallow: /gp/sign-in
Disallow: /gp/slides/make-money
Disallow: /gp/structured-ratings/actions/get-experience.html
Disallow: /gp/twitter/
Disallow: /gp/vote
Disallow: /gp/voting/
Disallow: /gp/yourstore
Disallow: /ap/signin
Disallow: /gp/registry/search.html
Disallow: /gp/orc/rml/
Disallow: /gp/dmusic/mp3/player
Disallow: /gp/entity-alert/external
Disallow: /gp/customer-reviews/dynamic/sims-box
Disallow: /review/dynamic/sims-box
Disallow: /gp/redirect.html
Disallow: /gp/customer-media/actions/delete/
Disallow: /gp/customer-media/actions/edit-caption/
Disallow: /gp/dmusic/
Allow: /gp/dmusic/promotions/AmazonMusicUnlimited
Disallow: /gp/customer-media/product-gallery/B007HCCOD0
Disallow: /gp/help/customer/display.html?*nodeId=200534000
Disallow: /gp/feature.html?*docId=1000632623
Disallow: /gp/aag
Disallow: /gp/socialmedia/giveaways
Disallow: /gp/aw/so.html
Disallow: /gp/pdp/profile/
Disallow: /gp/product/product-availability
Disallow: /gp/offer-listing
Disallow: /dp/twister-update/
Disallow: /dp/e-mail-friend/
Disallow: /gp/registry/wishlist/
Disallow: /wishlist/
Allow: /wishlist/universal
Allow: /wishlist/vendor-button
Allow: /wishlist/get-button
```



```

Disallow: /gp/wishlist/
Allow: /gp/wishlist/universal
Allow: /gp/wishlist/vendor-button
Allow: /gp/wishlist/ipad-install
Disallow: /registry/wishlist/
Disallow: /local/ajax/
Disallow: /gp/rentallist
Disallow: /gp/video/dvd-rental/settings
Disallow: /gp/rl/settings
Disallow: /gp/video/settings
Disallow: /gp/video/watchlist
Disallow: /gp/video/library
Disallow: /gp/profile/
Disallow: /reviews/iframe
Disallow: /gp/ask-widget/askWidget*
Disallow: /ss/customer-reviews/lighthouse/
Disallow: /gp/aw/ol/
Disallow: /gp/promotion/
Disallow: /hz/leaderboard/top-reviewers/
Disallow: /hz/leaderboard/hall-of-fame/
Disallow: /review/top-reviewers/
Disallow: /review/top-reviewers
Disallow: /review/hall-of-fame
Disallow: /reviews/top-reviewers/
Disallow: /reviews/top-reviewers
Disallow: /reviews/hall-of-fame

User-agent: EtaoSpider
Disallow: /

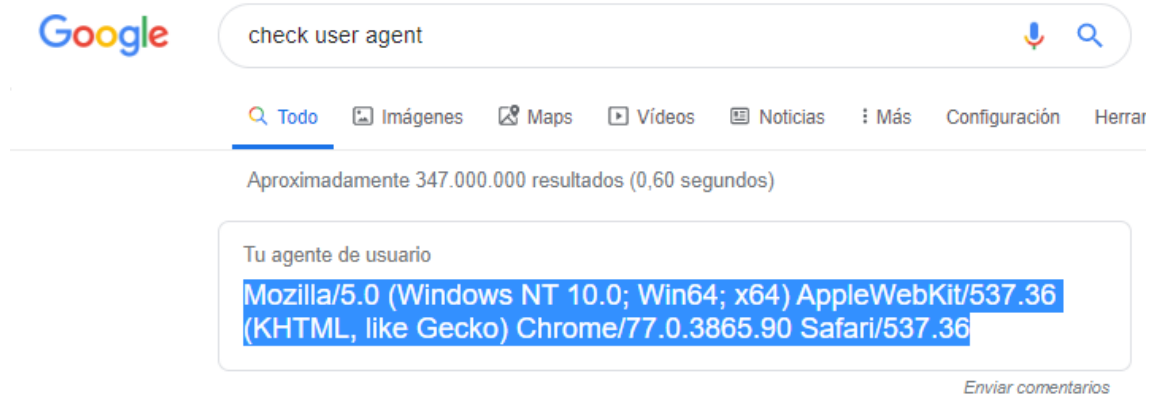
# Sitemap files
Sitemap:
https://www.amazon.es/sitemaps.2307ea63773dfee.SitemapIndex_0.xml.gz
Sitemap:
https://www.amazon.es/sitemaps.d449d7f825f081e.SitemapIndex_0.xml.gz
Sitemap:
https://www.amazon.es/sitemaps.e49bfbf08ac5517.SitemapIndex_0.xml.gz
Sitemap:
https://www.amazon.es/sitemaps.95918f5af3f77b0.SitemapIndex_0.xml.gz
Sitemap:
https://www.amazon.es/sitemaps.02b411f7e4baecc.SitemapIndex_0.xml.gz

```

El archivo robots.txt nos muestra la exclusión de directorios para los robots y restricción completa para el robot EtaoSpider.

2. Mapa del sitio web.
3. Tamaño del sitio.
- 4.

Al intentar acceder a la página de un producto se observa que Amazon detecta que es script de software y bloquea el acceso. Para ello procedemos a modificar el *user agent* la cabecera HTTP cambiándolo por el user agent que utiliza nuestro navegador:



Google

check user agent

Todo Imágenes Maps Vídeos Noticias Más Configuración Herrar

Aproximadamente 347.000.000 resultados (0,60 segundos)

Tu agente de usuario

Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/77.0.3865.90 Safari/537.36

Enviar comentarios