Práctica 1 - Tipología y Ciclo de Vida de los Datos

Asignatura: M2.851 / Semestre: 2022-2023 / Fecha: 22-11-2022

URL del sitio web elegido: https://www.amazon.es/gp/bestsellers/books/

Autores

- Araceli Najar García- anajar@uoc.edu (AN)
- Bianca Nathalie Palacios Pinargotti- bplaciosp@uoc.edu (BP)

Resolución de los apartados

1. Contexto.

Amazon es una corporación de comercio electrónico que ofrece cientos de productos distintos, uno de ellos son libros en formato físico o electrónico. Su página en España muestra los más vendidos cada día, sin importar su categoría. Por tanto, este conjunto de datos aporta información de los libros más vendidos en Amazon España y permite conocer las opciones de lectura actualmente en una cultura donde la lectura está perdiendo su valor. Los datos han sido recogidos mediante su página web https://www.amazon.es/gp/bestsellers/books/.

2. Título del dataset

Libros más vendidos en Amazon España el 9 de Noviembre del 2022.

3. Descripción del dataset.

Dataset con los datos obtenidos de Amazon España el 9 de Noviembre del 2022 y que muestran el ranking, título, autor, precio en euros, puntuación dada por los lectores, número de reviews dejados en la página y la imagen de portada de los libros más vendidos en España.

4. Representación gráfica.

5. Contenido.

El dataset creado contiene los siguientes campos:

- **Ranking** es la posición en que se encuentra el libro entre los más vendidos del día.
- **Título** del libro.
- **Autor** (o autores) del libro.
- **Precio en euros** del producto en la fecha actual.
- **Puntuación** otorgada al producto. Para la valoración, Amazon calcula las calificaciones del libro mediante modelos de aprendizaje automático en vez de hacer una media simple. Se toman en cuenta varios factores como la edad de las personas que dan las opiniones, los votos de los clientes en un rango de 1 a 5 estrellas o si la opinión es de compra verificada.

Fuente: https://www.amazon.com/gp/help/customer/display.html?nodeId=GQUXAMY73JFRVJHE

- **Número de comentarios (reviews)** u opiniones otorgado a cada libro. Los clientes valoran las opiniones objetivas sobre los libros lo que permite al resto de clientes leer los comentarios de otros lectores sobre sus libros y tomar decisiones de compra.
- **Imagen de portada** de los libros. Se adjunta un enlace a la imagen.

La recolección de los datos fue realizada el 9 de Noviembre del 2022.

6. Propietario.

El propietario del conjunto de datos es **Amazon.com Inc**, una corporación estadounidense de comercio electrónico. En concreto, los datos obtenidos mediante web scraping se han tomado de la vertiente española de esta web y, dentro de esta, del apartado de libros más vendidos (*bestsellers*) en el día de extracción de los datos.

No hemos encontrado análisis anteriores que se basen en estos datos ni en esta información concreta de los libros más vendidos por amazon, a excepción de algunas revistas o webs de estrategias SEO (por ejemplo, los artículos : https://www.elle.com/es/living/ocio-cultura/g22618134/libros-mas-vendidos-amazon/ y https://www.agenciadeseo.es/libros-mas-vendidos-en-amazon/), en las que podemos encontrar artículos que listan y comentan cuáles son los libros más vendidos por Amazon España en ese momento. Sin embargo, este parece ser un análisis manual que no va más allá de registrar el título, sinopsis, ranking y precio del top 20 o top 15 de libros. Sin embargo, estos artículos no contienen información sobre las mismas variables que se han escogido en el conjunto de datos obtenido en esta práctica.

Por otro lado, a la hora de la obtención del dataset se han seguido una serie de pautas para actuar de acuerdo a los **principios legales y éticos** en el contexto del proyecto. Se han verificado las condiciones de uso, no se ha parseado el HTML manualmente (sino con el uso de la librero *BeautifulSoup*) y de manera inicial se ha consultado el archivo *robots.tx*t para tener en cuenta las restricciones de la web de Amazon España a la hora de ser rastreada. Además, se ha rastreado y obtenido exclusivamente información pública, a la que cualquier usuario puede acceder sin necesidad de iniciar sesión o registrarse en el sitio web, y tampoco se ha incluido ningún dato privado o restringido. Del mismo modo, se ha seguido el principio de no causar daño, es decir, se ha evitado sobrecargar el servidor web con un número excesivo de peticiones y la finalidad del proyecto es utilizar la información obtenida de manera justa y con un mero objetivo académico e informativo.

7. Inspiración.

El conjunto de datos obtenidos en este proyecto puede resultar de interés para diferentes objetivos. Por ejemplo, una finalidad sería analizar el tipo de libros que más se venden en España y evaluar si los autores locales tienen más éxito entre los usuarios españoles o si bien predominan las obras extranjeras (esto, a su vez, podría ser una información interesante para comparar con la disponibilidad de libros de autores locales y extranjeros en otras fuentes de comercio como pueden ser otras webs dedicadas a la venta de libros o pequeñas librerías, ya que es posible que en Amazon primen los *bestsellers* extranjeros si este tipo de literatura es más difícil de encontrar en otras tiendas).

También se podría responder con este dataset la pregunta de cuánto dinero se suele gastar la mayoría de usuarios de Amazon en consumir libros y literatura, y si el precio de los libros puede llegar a ser determinante a la hora de venderse más o menos (o si, en cambio, libros con un precio superior a la media se encuentra en posiciones muy altas del ranking de más vendidos y, por tanto, los compradores no se ven frenados por el un precio excesivo cuando se trata de un libro muy esperado o que suscita mucho interés por el público).

Además, variables como la puntuación del libro puede ser de utilidad para evaluar si realmente los libros más vendidos son bien recibidos por los usuarios y bien puntuados por estos, o si por el contrario presentan calificaciones muy bajas. Esto plantearía un posible debate y análisis posterior de si los libros más vendidos son aquellos con una estrategia de marketing más potente o aquellos con una calidad literaria superior, o incluso si mucho de los *bestsellers* se convierten en tal por la publicidad y el 'boca a boca' aunque finalmente no acaben generando críticas positivas entre el público. Unir a esta información la proporcionada por la variable que indica el número de críticas de cada libro nos puede dar una perspectiva más fundamentada del éxito y la valoración general de los libros, ya que habría que tener en cuenta no solo la puntuación del libro sino también las veces que ha sido puntuado (por ejemplo, dos libros que presenten una puntuación de 4 pero uno tenga 2 valoraciones y otro 100 son situaciones muy diferentes).

Finalmente, otra de las variables registradas en este dataset puede ser de interés y ser el punto de partida de un análisis más allá de los datos numéricos o categóricos: la variable que registra las imágenes de las distintas portadas de los libros más vendidos. Es útil para realizar un estudio sobre la influencia del estilo (tipografía, colores, diseño, etc.) de las portadas para ser más atractivas a los consumidores. Es decir, la presentación de la portada podría ser un factor que promueva la venta de un libro.

En definitiva, el conjunto de datos obtenido en la práctica de *web scraping* otorga listados con información de los libros más vendidos y con la opción de obtenerlos en diferentes fechas. Por tanto, puede ser interesante para evaluar y describir las tendencias en el mundo literario español y analizar cómo cambian aspectos como el precio en diferentes épocas del año. También nos permitiría responder preguntas que van más allá del análisis descriptivo, como las comentadas anteriormente.

8. Licencia.

La licencia que se ha seleccionado para el dataset resultante ha sido una **Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0)**. Esta licencia se ha escogido ya que se trata de un conjunto de datos que se pueden compartir y adaptar de manera libre. La presente práctica emplea información publicada en la web de Amazon disponible para cualquier usuario con acceso a dicha web y cuyos valores e información no se encuentran limitados en absoluto. Por ejemplo, en la web de más vendidos de Amazon se puede consultar el precio o valoraciones de todos los libros del catálogo de venta al ser información abierta al público. Es decir, la información contenida en este conjunto de datos puede ser copiada y redistribuida en cualquier medio o formato y, a su vez, se permite su transformación y adaptación para cualquier propósito, sea este comercial o no.

9. Código.

El código se ha realizado utilizando el lenguaje de programación Python. En la carpeta source se encuentran:

- El fichero **scraper.py** con el código comentado que contiene las funciones usadas para obtener la información de los libros más vendidos.
- El fichero **main.py** con el ejecutable del scraper.

Dentro de los aspectos más relevantes del código cabe destacar que la función *bestsellers_scraper()* se encarga de realizar todo el proceso de scraping y obtención de las variables que se han escogido para formar parte del dataset resultante, que se obtiene en formato CSV. Además se presenta la opción de indicar el número de páginas del listado completo de libros más vendidos en Amazon España de las que se desea extraer la información de los ejemplares. En cada página aparecen un total de 50 libros, así, según el valor que se le dé al argumento de la función, se obtendrá un total de 50**num_páginas* libros. El valor por defecto de este argumento se ha especificado en 1 (es decir, se recopilan los top 50 libros más vendidos).

```
def bestsellers_scraper(self, num_paginas=1):
    #-------------------------------
    # Esta función realiza el proceso de scraping completo
    # y devuelve el dataset final en formato .csv
    # Su argumento de entrada es el número de páginas
    # que queremos incluir a la hora de recopilar los bestsellers.
    # 1 página = 50 libros bestsellers.
    #-------------------------------
    libros_total = []
    for pagina in range(1, num_paginas+1):
        # ruta de la web utilizada para este trabajo
        url = 'https://www.amazon.es/gp/bestsellers/books/'

        # se modifica el url para recoger información de otras páginas
        if pagina > 1:
            pagina_url =  f'https://www.amazon.es/gp/bestsellers/books/ref=zg_bs_pg_{pagina}?ie=UTF8&pg={pagina}'

        soup = self.__contenido_pagina(url)
        sleep(randint(1,5))

        # para cada libro, se extrae su información de la página web
        libro_info = soup.find_all('div', id='gridItemRoot')
        libros_total += libro_info
    libro_dict = self.__datos_libros_full(libros_total) # se unifica la info extraída del conjunto de libros
    bestseller_df = pd.DataFrame(libro_dict) # df con la info de todos los libros
    bestseller_df.to_csv(f'amazon_{num_paginas*50}bestsellers.csv', index = True) # se guarda el df en un fichero .csv

def __contenido_pagina(self, url):
    #-------------------------------
    # Esta función utiliza la librería requests para acceder al contenido de
    # la web, mostrando un aviso si se produce un error durante el proceso.
    # Posteriormente se parsea el contenido de la web con BeautifulSoup
    #-------------------------------
    page = requests.get(url)
    if page.status_code != 200:
        raise Exception('Error al acceder al contenido de la web')
    soup = BeautifulSoup(page.content, 'html.parser')
    return soup

def __datos_libros_full(self, libro_info):
    #-------------------------------
    # Esta función recibe toda la información de todos los libros de la página
    # Es decir, el contenido de la web donde se encuentra la información que las anteriores funciones
    # irán encontrando. Inicia un diccionario vacío sobre el que se incluirá la información de cada libro
    # Utiliza la función datos_libro() para obtener las variables del dataset para cada uno de los libros
    #-------------------------------
    libro_dict = {
        'ranking' : [],
        'titulo' : [],
        'autor' : [],
        'precio_euros' : [],
        'puntuacion' : [],
        'num_reviews' : [],
        'portada' : []
    }
    sleep(randint(1,5))

    for item in libro_info:
        datos_libros_full = self.__datos_libro(item)
        libro_dict['ranking'].append(datos_libros_full['ranking'])
        libro_dict['titulo'].append(datos_libros_full['titulo'])
        libro_dict['autor'].append(datos_libros_full['autor'])
        libro_dict['precio_euros'].append(datos_libros_full['precio_euros'])
        libro_dict['puntuacion'].append(datos_libros_full['puntuacion'])
        libro_dict['num_reviews'].append(datos_libros_full['num_reviews'])
        libro_dict['portada'].append(datos_libros_full['portada'])

    return libro_dict
```

El proceso de scraping se inicia obteniendo el contenido de la web con la librería *requests* y parseando el resultado obtenido con la librería *BeautifulSoup*. En caso de que no sea posible obtener el contenido de la url indicada, se para el proceso y se devuelve un mensaje de error.

Tras esto se obtienen todas las etiquetas que engloban la información de los ejemplares más vendidos por cada página y se almacenan en una lista.

Sobre cada libro de esta lista se aplica la función *datos_libro*(), la que a su vez ejecuta el resto de funciones definidas en el scraper para ir obteniendo cada una de las variables que compondrán el conjunto de datos final.

```python
def __datos_libro(self, libro):
    #-----------------------------
    # Esta función aplica las funciones definidas con anterioridad
    # para obtener el título, autor, puntuación, nº reseñas y precio
    # Además busca en el contenido de la página para obtener el ranking y la portada
    # Todo esto lo hace para un único libro
    #-----------------------------
    ranking = libro.find('span', class_="zg-bdg-text").text.strip('#')
    portada = libro.find('img')['src']
    titulo, autor = self.__titulo_autor(libro)
    puntuacion, num_reviews = self.__puntuacion_numreviews(libro)
    precio = self.__precio_libro(libro)
    sleep(randint(1,5))
    return {
        'ranking' : ranking,
        'titulo' : titulo,
        'autor' : autor,
        'precio_euros' : precio,
        'puntuacion' : puntuacion,
        'num_reviews' : num_reviews,
        'portada' : portada
    }
```

```python
def __titulo_autor(self, libro):
    #-----------------------------
    # Esta función busca y devuelve el título del libro y el nombre del autor
    # dentro del resultado obtenido con BeautifulSoup
    # Si no se encuentra registrado se devuelve el valor 'None'
    #-----------------------------
    titulo_autor = libro.find_all('div', class_="_cDEzb_p13n-sc-css-line-clamp-1_1Fn1y")
    if len(titulo_autor) < 2:
        titulo_doc = libro.find_all('div', class_="_cDEzb_p13n-sc-css-line-clamp-2_EWgCb")
        if len(titulo_doc)<1:
            titulo = 'None'
        else:
            titulo = titulo_doc[0].text
        autor_doc = libro.find_all('div', class_="_cDEzb_p13n-sc-css-line-clamp-1_1Fn1y")
        if len(autor_doc)<1:
            autor = 'None'
        else:
            autor = autor_doc[0].text
    else:
        titulo = titulo_autor[0].text
        autor = titulo_autor[1].text
    return titulo, autor
```

```python
def __puntuacion_numreviews(self, libro):
    #-----------------------------
    # Esta función busca y devuelve la puntuación media del libro y el nº de reseñas
    # Si no se encuentran registrados se devuelve el valor 'None'
    #-----------------------------
    puntuacion_numreviews = libro.find('div', class_="a-icon-row")
    if puntuacion_numreviews == None:
        puntuacion = 'None'
        num_reviews = 'None'
    else:
        puntuacion = puntuacion_numreviews.find('i').text.split(' ')[0]
        reviews = puntuacion_numreviews.find('span', class_="a-size-small").text
        num_reviews = ''
        for i in reviews.split(','):
            num_reviews += i
    return puntuacion, num_reviews
```

```python
def __precio_libro(self, libro):
    #-----------------------------
    # Esta función busca y devuelve el precio del libro
    # Si no se encuentra registrado se devuelve el valor 'None'
    # Se elimina el símbolo del € para solo mostrar la cifra numérica
    #-----------------------------
    precio_doc = libro.find('span', class_="p13n-sc-price")
    if precio_doc == None:
        precio = 'None'
    else:
        precio = precio_doc.text.strip('€')
    return precio
```

10. Dataset.

El dataset obtenido tras realizar el scraping en la web de libros más vendidos en Amazon España (https://www.amazon.es/gp/bestsellers/books/) incluye la siguiente información de los 100* libros más vendidos en el día 09/11/2022:

- Variable **'ranking'**: posición de cada libro dentro del listado de más vendidos (numérica integer).
- Variable **'titulo'**: título del libro (cadena de caracteres).
- Variable **'autor'**: autor o autores del libro (cadena de caracteres).
- Variable **'precio_euros'**: precio del libro en euros (numérica float).
- Variable **'puntuacion'**: puntuación del libro según las estrellas que los compradores han otorgado a dicho libro (numérica float, rango 0-5).
- Variable **'num_reviews'**: número de comentarios en los que se ha comentado o reseñado el libro (numérica integer).
- Variable **'portada'**: enlace a la imagen de portada del libro (cadena de caracteres).

El dataset se encuentra publicado en formato CSV en Zenodo. El enlace del DOI de este dataset es el siguiente: https://doi.org/10.5281/zenodo.7338204

*En este caso se ha decidido tomar los 100 libros más vendidos para nuestro conjunto de datos definitivo, pero según se ha comentado en apartados anteriores a la hora de hacer el web scraping es posible seleccionar la cantidad de libros del listado total de bestsellers de los que se quiere obtener la información (con un aumento o decremento de 50 unidades, es decir, los 50 libros más vendidos, los 100 libros más vendidos, los 150 libros más vendidos, etc.).

11. Vídeo.

Se adjunta el enlace del video almacenado en Google Drive UOC:

https://drive.google.com/drive/folders/1ChyLGRmZ7othKB4nalXHcv5jhb63vGzW?usp=share_link

Contribuciones	Firma
Investigación previa	AN, BP
Redacción de las respuestas	AN, BP
Desarrollo del código	AN, BP
Participación en el vídeo	AN, BP