Práctica 1 Tipología y ciclo de vida de los datos

Nombre de los integrantes del grupo: Juan Lara Chups y Oriol Mössinger Sanahuja

1. Contexto

Para hacer este trabajo de web scraping debíamos elegir una página web “real” que se le pudiera hacer web scraping para obtener información. Para nuestro caso de estudio hemos elegido Norma Comics.

Norma Comics es una tienda física de cómics americanos, mangas, animes, merchandising, productos japoneses, etc. muy famosa en Barcelona situada en el Passeig de Sant Joan.

Debido al avance de las tecnologías de la información en los últimos años, Norma Comics abrió su propia tienda online para que aficionados de toda España pudieran comprar sus productos y ampliar su cuota de mercado más allá de Barcelona y sus alrededores.

El contexto en el cual hemos recolectado la información es centrarnos en la oferta de cómics americanos para realizar un web scraping de todos los cómics que tienen a la venta en su sección de “Marvel Cómics”. Dicha sección se va ampliando periódicamente ya que la oferta de cómics va en constante aumento.

Finalmente, la página web que hemos usado para hacer web scraping es:
https://www.normacomics.com/

2. Título

El título que proponemos es “Oferta de cómics de Marvel en Norma Comics”.

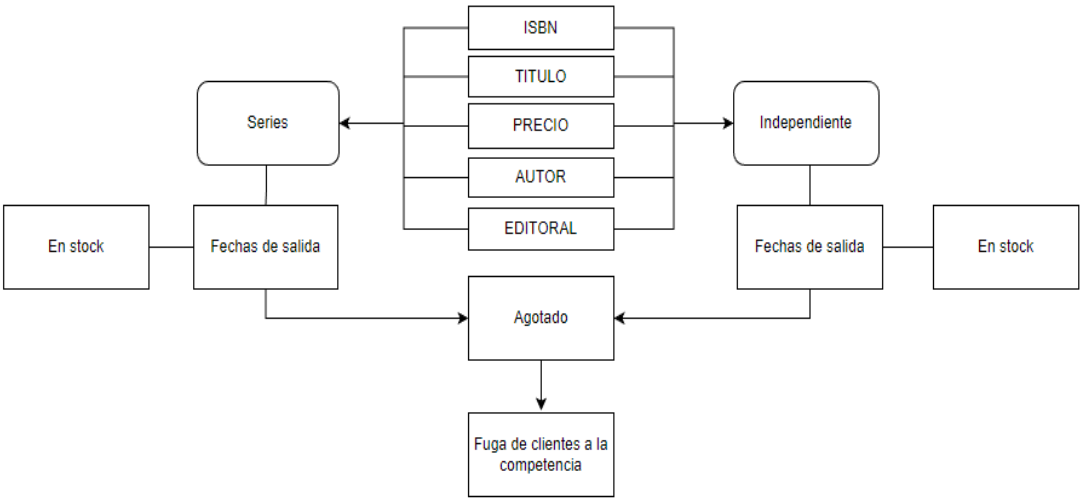
3. Descripción del dataset

El dataset que hemos obtenido a través de web scraping contiene las siguientes columnas: Nombre, Precio, Editorial, ISBN, Fecha de lanzamiento, Formato, Número de páginas y Disponibilidad.

En el dataset se recogen todos los títulos que Norma Comics oferta en su página web filtrados por cómics americanos. El periodo de tiempo de los datos se encuentran comprendidos por años, meses y días desde el año 2009 hasta la actualidad. Los datos no han realizado la fase de preproceso o limpieza, por lo que pueden existir inconsistencias en ellos. El formato del dataset es un fichero CSV.

En este dataset hemos querido compilar en un fichero csv los diferentes elementos que se consideran más importantes en la venta de un cómic y que están a disposición del comprador en la página web de https://www.normacomics.com/.

4. Representación gráfica



En el gráfico anterior podemos observar la estructura visual de los datos considerando las variables stock y fecha que contienen el dataset. El tipo de cómics se pueden dividir en dos grandes grupos, series e independientes.

Los cómics denominados series son aquellos que siguen una cronología siguiendo la historia principal a través de numerosos cómics. En nuestro dataset se identifican con el número entre paréntesis que aparece al final del título del comics.

Los cómics independientes son aquellos que cuentan historias paralelas en el universo Marvel pero que tienen su inicio y final en el mismo cómic.

Cuando se menciona la fuga de clientes se debe a que el hecho de que no haya stocks disponibles podría acarrear una pérdida de clientes o su compra a la competencia.

5. Contenido

Los campos incluidos en el dataset y que hemos mencionado anteriormente en el punto 3 son los siguientes:

- Nombre: La columna "Nombre" del dataset contiene el nombre de cada uno de los cómics de Marvel que se ofrecen en la tienda online "Norma Comics". Cada entrada de la columna es una cadena de caracteres (string) que representa el título del cómic.

- Precio: La columna "Precio" en el conjunto de datos contiene el precio de venta al público de cada cómic de Marvel que se ha extraído de la página web de Norma Comics. El precio está en formato de texto y se encuentra en euros (€). Finalmente, se convierte la columna de precios a formato numérico para que se puedan realizar operaciones matemáticas con ella.

- Autor: La columna "Autor" del conjunto de datos contiene información sobre el autor o autores del cómic. Es probable que en algunos casos el campo esté vacío o contenga una cadena como "Varios autores" o "Equipo creativo".

- Editorial: La columna "Editorial" contiene el nombre de la editorial que publicó cada cómic de Marvel.

- ISBN: La columna "ISBN" contiene los números de identificación internacional de libro (ISBN) de cada cómic en la lista. El ISBN es un identificador único de libros y publicaciones seriadas utilizado en todo el mundo para identificar de manera unívoca cada título o edición de un libro o publicación seriada. El ISBN consta de 13 dígitos (anteriormente de 10 dígitos) y puede utilizarse para buscar información detallada sobre un libro específico en una base de datos de bibliotecas o en línea. En el conjunto de datos, esta columna se utiliza para identificar de manera única cada cómic en la lista. Por tanto, en un futuro se podría usar como clave primaria para realizar trabajos con SQL.

- Fecha de lanzamiento: La columna "Fecha de lanzamiento" contiene la fecha en la que se puso a la venta cada cómic en el sitio web de Norma Comics. Esta información se obtiene haciendo web scraping de la página de detalles de cada cómic. La fecha de lanzamiento se representa en formato de fecha y se guarda en formato de cadena de caracteres en el dataframe. Para facilitar su manipulación posterior, se convierte en objeto de fecha utilizando la biblioteca datetime. Si la fecha no se puede analizar, se asigna un valor vacío a la celda correspondiente en el dataframe.

- Formato: La columna "Formato" del conjunto de datos contiene información sobre el formato de los cómics de Marvel disponibles en la tienda en línea Norma Comics. Esta columna indica el tipo de formato de cada cómic, que puede ser tapa dura, tapa blanda, grapas, entre otros. La información de esta columna puede ser útil para los compradores que prefieren un tipo específico de formato o que buscan colecciones completas de una serie de cómics.

- Número de páginas: La columna "Número de páginas" indica la cantidad de páginas que tiene cada cómic en la base de datos. Esta información es útil para los lectores que desean saber cuánto tiempo les tomará leer un cómic o para aquellos que prefieren cómics más cortos o más largos en función de su tiempo disponible para la lectura.

- Disponibilidad: La columna "Disponibilidad" indica si el cómic está actualmente disponible para la compra en la tienda o no. Puede tomar 2 valores:
    - En stock: indica que el cómic está disponible para su compra.
    - Agotado: indica que el cómic ha sido vendido y ya no está en stock en ese momento.

6. Propietario

El propietario intelectual del dataset es la empresa Norma Editorial, propietaria de la tienda online Norma Comics. Al ser una empresa privada que ofrece un servicio de venta de cómics al por menor, debemos comprobar sus condiciones de uso en el siguiente enlace:

https://www.normacomics.com/condiciones-de-uso

Una vez leídas las condiciones de uso, nos encontramos con el siguiente párrafo:

PROPIEDAD INTELECTUAL
Los derechos de propiedad intelectual e industrial derivados de todos los textos, imágenes, así como de los medios y formas de presentación y montaje de su sitio web pertenecen, por sí o como cesionaria, a NORMA EDITORIAL, S.A.. Serán, por consiguiente, obras protegidas como propiedad intelectual por el ordenamiento jurídico español, siéndoles aplicables tanto la normativa española y comunitaria en este campo, como los tratados internacionales relativos a la materia y suscritos por España.

Todos los derechos reservados. En cumplimiento de la Ley de la Propiedad Intelectual se prohíbe expresamente la reproducción, distribución, comunicación pública y utilización, de la totalidad o parte de los contenidos de su sitio web sin el consentimiento expreso de NORMA EDITORIAL, S.A.

Llamamos al teléfono de contacto de Norma Comics, 93 244 81 25, fuimos redirigidos al departamento de Ecommerce y después de hablar con el responsable del departamento mandamos el siguiente mail. Se adjunta captura de pantalla en la siguiente página de esta memoria.

Oriol Mössinger Sanahuja <omossinger@uoc.edu>

Consulta legal Web Scraping Norma Comics

1 message

Oriol Mössinger Sanahuja <omossinger@uoc.edu> Wed, Apr 19, 2023 at 1:07 PM
To: info@normacomics.com
Cc: Juan Lara Chups <jlarachu@uoc.edu>

Buenas tardes Jose Antonio,

Me llamo Oriol Mössinger Sanahuja (DNI: 39410385T) y mi compañero Juan Lara Chups (DNI: 31024872B) y yo estamos estamos realizando el máster en Ciencia de Datos en la UOC.

En la asignatura de Tipología y ciclo de vida de los datos nos piden que hagamos un trabajo sobre web scraping. Adjunto el enunciado para que ustedes verifiquen qué nos solicita el profesor.

Solicitamos que nos den permiso para poder realizar web scraping en su página web.

Queremos captura los siguientes elementos de los cómics americanos:

- Nombre
- Precio
- Autores
- Editorial
- ISBN
- Fecha de venta
- Formato
- Número de páginas
- Si está en stock o agotado

Muchas gracias,

Cordialmente,

Oriol.

El responsable del departamento de Ecommerce nos llamó de vuelta para consultarnos qué tipo de consulta le haríamos ya que nunca había escuchado de la técnica de captación de datos web scraping. Nosotros le explicamos en qué consistía y que solamente era una consulta académica y no recurrente en el tiempo.

Nos dio el visto bueno por teléfono con la condición de que la consulta que hiciéramos con web scraping no saturaran el servidor de la tienda online. Le solicitamos que nos confirmara por mail dicho permiso para realizar web scraping en la tienda online, nos dijo que nos contestaría más tarde al mail y a fecha de entrega de este trabajo aún no nos contestó a este mail a pesar de tener su aceptación verbal.

A pesar de no tener la confirmación escrita y solamente tener su confirmación verbal, continuamos el trabajo y de forma ética nos comprometemos a usar este código de web scraping para fines estrictamente académicos. También nos comprometemos a no usar ningún bot que pudiese sobrecargar el servidor de Norma Comics. Igualmente, en sus condiciones de uso, en ningún momento cita explícitamente la prohibición del uso de web scraping en todo el portal de normacomics.com .

Investigando por Github no hemos encontrado ningún análisis de mercado sobre alguna tienda de cómics, por tanto hemos buscado otros análisis de mercado y hemos encontrado el siguiente:
https://github.com/sophiachann/WebScrapingProject-EcommercesAnalysis

Dicho proyecto de análisis de negocio compara diferentes KPIs entre el ecommerce hongkonés HKTVmall y el ecommerce estadounidense Amazon.

Básicamente en este análisis de los dos mercados se plantean las siguientes cuestiones:

- Identificación de nuevas oportunidades de negocio
- Entendimiento de las necesidades del cliente
- Determinación de los factores de crecimiento

7. Inspiración

Este dataset puede ser interesante desde el punto de vista de marketing para realizar estudios de diversas características:

- Estudio de mercado: si es bien sabido el universo Marvel se ha dado más a conocer gracias al universo cinematográfico, por lo que resulta de gran interés observar los cómics de las películas asociadas y poder realizar ofertas de aquellos cómics que coincidan con los superhéroes aparecidos en la gran pantalla.
- Del mismo modo, podemos identificar las editoriales y los autores que más cómics tienen publicados en la página o, a su vez, identificar editoriales con menos potencial o con menos cómics a la venta.
- Riesgo de fuga de clientes potenciales o identificación de producto estrella: con el dataset y el código realizado para captar los datos podemos identificar los cómics que tienen stock o no. Esto puede ser interesante ya que el hecho de que esté sin stock podría deberse a que el cómic en cuestión es un producto estrella.
- De la misma manera, se pueden realizar hipótesis utilizando la fecha de lanzamiento del cómic para la toma de decisión de renovar el stock, retirar el producto o introducir una opción de reserva para maximizar los beneficios.
- Precio medio del cómic: podemos obtener los precios medios de los cómics filtrando por superhéroes y editorial y ordenar los cómics de mayor a menor precio.

En resumidas palabras, las preguntas que se pueden contestar con este dataset serían las siguientes:

- ¿Cuánto cuesta en promedio un cómic?
- ¿Cuántas páginas tiene en promedio un cómic?
- ¿Cuántos cómics salen al mercado cada mes?
- ¿Qué editoriales publican más cómics?
- ¿Qué autores son los que tienen más publicaciones?
- ¿Qué y cuántos cómics pueden ser conocidos debido al salto del UM a la gran pantalla?
- ¿Cuántos cómics se encuentran sin stock? ¿Cuáles son estos cómics?
- ¿Qué características comparten los cómics que están agotados?

Vinculado al punto 6 de este trabajo. Podemos analizar la elasticidad de la demanda de cómics, el precio medio de los cómics que están agotados (y por tanto podemos suponer que el consumidor le gustan), plantear ofertas para aumentar el consumo de los clientes como 3x2 en cómics o descuentos si compras una serie de cómics entera.

También, aunque nosotros en este trabajo no dispongamos de dicha información, el propietario de los datos, Norma Comics, podría cruzar el dataset que estamos trabajando con la base de datos de sus clientes para hacer recomendaciones personalizadas.

8. Licencia

En el caso de las licencias a elegir para nuestro conjunto de datos podemos elegir entre dos de ellas: CC BY-NC-SA 4.0 y CC-BY-SA 4.0.

Las dos licencias cumplen con el criterio de Creative Commons, por lo que el dataset queda a disposición del público. Se permite a otros distribuir, remezclar, ajustar y construir sobre el material siempre y cuando se le otorgue el crédito adecuado al creador original.

Las diferencias principales que existen entre las dos licencias recaen en su uso comercial y la derivación posterior de trabajos basados en nuestro conjunto de datos.

Con la licencia CC BY-NC-SA 4.0 el uso del conjunto de datos queda prohibido su uso comercial, o con otras palabras, para fines lucrativos. En cambio, con la licencia CC-BY-SA 4.0, queda totalmente libre su uso para fines comerciales.

Cuando hablamos de la derivación de trabajos nos referimos a que en el caso de aplicar la licencia CC BY-NC-SA 4.0 se le obliga la persona que va a realizar el trabajo sobre nuestro a dataset el hecho de derivar el resultado con la misma licencia. En cambio, con la licencia CC-BY-SA 4.0 solo requiere que los trabajos derivados se distribuyan bajo la misma licencia o una licencia similar.

Siguiendo las pautas éticas de nuestro trabajo, hemos decidido seleccionar la licencia CC BY-NC-SA para restringir los usos del dataset a fines estrictamente académicos y defender al dueño, Norma Comics, de posibles repercusiones que podrían generar trabajos derivados.

9. Código

El código ha sido realizado y manipulado en lenguaje python con la implantación de la librería selenium. El intérprete utilizado ha sido google colab y pycharm. El código fuente se encuentra en la carpeta de github tal y como se especifica en el enunciado.

10. Dataset

A continuación, se presenta una parte de los datos de nuestro dataset resultante. Al final del apartado aparece el DOI en zenodo y el enlace directo al dataset.

VENENO 14 (62)	Al Ewing, Bryan Hitch	Panini Comics	9770005528007000	3,14 €	05/04/2023	Grapa	24	En stock
SALVAJES VENGAD	Carlos Magno, David Pepose	Panini Comics	9770005576008000	3,14 €	05/04/2023	Grapa	24	En stock
LOS CUATRO FANT	Ibán Coello, Ryan North	Panini Comics	9770005543000000	3,14 €	05/04/2023	Grapa	24	En stock
MILES MORALES: S	Cody Ziglar, Federico Vicentini	Panini Comics	9770005550008000	3,61 €	05/04/2023	Grapa	32	En stock
MOTORISTA FANTA	Benjamin Percy, Cory Smith	Panini Comics	9770005594002000	5,23 €	05/04/2023	Grapa	48	En stock
EL INVENCIBLE IRO	Gerry Duggan, Juan Frigeri	Panini Comics	9770005443003001	4,37 €	05/04/2023	Grapa	40	En stock

El DOI al dataset en zenodo: 10.5281/zenodo.7856522

Enlace directo al dataset: https://zenodo.org/record/7856522#.ZEUC2XZByAs

A continuación se muestra la participación de los integrantes del grupo:

CONTRIBUCIONES	**FIRMA**
Investigación previa	OMS, JLC
Redacción de las respuestas	OMS, JLC
Desarrollo del código	OMS, JLC
Participación en el video	OMS, JLC

```
# Importamos las diferentes librerías que deberemos usar para
# realizar el ejercicio de web scraping:

from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from datetime import datetime
import pandas as pd
from selenium.webdriver.chrome.service import Service
from bs4 import BeautifulSoup
import locale

# Iniciamos el controlador de Chrome para hacer web scraping:
s = Service('/usr/local/bin/chromedriver')
driver = webdriver.Chrome(service=s)

# Definimos la URL base y los parámetros de búsqueda
base_url = 'https://www.normacomics.com'
params = '/comics/comic-americano/marvel-comics.html?p='


# Creamos una lista de URLs completas de cada página de resultados
urls = [base_url + params + str(page) + '&product_list_limit=72' for page in
range(1, 23)]

# Hacemos las peticiones HTTP y extraemos el contenido HTML
soups = []
for url in urls:
   driver.get(url)
   WebDriverWait(driver,
10).until(EC.presence_of_element_located((By.CSS_SELECTOR,
"li.item.product.product-item")))
   soups.append(BeautifulSoup(driver.page_source, 'html.parser'))

# Extraemos los enlaces de cada cómic
comic_links = []
```

```
for soup in soups:
    comics = soup.find_all('li', class_='item product product-item')
    for comic in comics:
        link = comic.find('a', class_='product-item-link')['href']
        comic_links.append(link)

# Hacemos las peticiones HTTP de cada enlace y extraemos su contenido HTML
soups2 = []
for link in comic_links:
    driver.get(link)
    WebDriverWait(driver,
10).until(EC.presence_of_element_located((By.CSS_SELECTOR, "td.col.data")))
    soups2.append(BeautifulSoup(driver.page_source, 'html.parser'))

# Extraemos las características de cada cómic en la página correspondiente
comic_names = [soup2.find('span', class_='base').text.strip() for soup2 in
soups2]
comic_prices = [soup2.find('span', class_='price').text.strip() for soup2 in
soups2]
comic_authors = [soup2.find('td', class_='col data').text.strip() for soup2 in
soups2]
comic_editorials = [soup2.find('td', attrs={'data-th':
'Editorial'}).text.strip() for soup2 in soups2]
comic_isbns = [soup2.find('td', attrs={'data-th': 'ISBN'}).text.strip() for
soup2 in soups2]

# Para no tener problemas con la configuración regional española:
locale.setlocale(locale.LC_TIME, 'es_ES')

# Para la fecha de lanzamiento haremos uso de datetime
# ya que en la página web sale en un formato "incómodo":
date_without_format = [soup2.find('td', attrs={'data-th': 'Fecha de
venta'}).text.strip() for soup2 in soups2]

# Convertimos las fechas en formato de string a datetime:
release_dates = []
for date in date_without_format:
    try:
```

```
            release_date = datetime.strptime(date, '%d %b %Y')
        except ValueError:
            release_date = None
        release_dates.append(release_date)

# Formateamos las fechas y les asignamos un nombre:
comic_releases = [release_date.strftime('%d/%m/%Y') if release_date else None
for release_date in release_dates]
comic_formats = [soup2.find('td', attrs={'data-th': 'Formato'}).text.strip()
if soup2.find('td', attrs={'data-th': 'Formato'}) else '' for soup2 in soups2]
comic_pages = [soup2.find('td', attrs={'data-th': 'Num páginas'}).text.strip()
if soup2.find('td', attrs={'data-th': 'Num páginas'}) else '' for soup2 in
soups2]
comic_stocks = ['En stock' if soup2.find('span',
class_='label-availability').find_next_sibling('span').text.strip() == 'En
stock' else 'Agotado' for soup2 in soups2]

# Creamos un dataframe con las características de cada cómic
df = pd.DataFrame({
'Nombre': comic_names,
'Autor': comic_authors,
'Editorial': comic_editorials,
'ISBN': comic_isbns,
'Precio': comic_prices,
'Fecha de lanzamiento': comic_releases,
'Formato': comic_formats,
'Páginas': comic_pages,
'Disponibilidad': comic_stocks
})

#Guardamos el dataframe como un archivo CSV
df.to_csv('comics_marvel.csv', index=False)

driver.quit()
```