



11-4-2022

Tipología y ciclo de vida de los datos

PRA1: Web Scrapping



Yésica Fernández Ramos y Carlos Ruiz Salvador

Contenido

Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.	2
Título. Definir un título que sea descriptivo para el dataset.	2
Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.	2
Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.	2
Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.	3
Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.	4
Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.	4
Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.	4
Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R	4
Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.	5

Contexto. Explicar en qué contexto se ha recolectado la información.
Explicar por qué el sitio web elegido proporciona dicha información.

El conjunto de datos elegido, cuyo origen viene dado por el sitio <https://www.alitacomics.com/>, el cual es una web de referencia dentro del mundo de los cómics y productos relacionados (figuras, merchandising, etc) en Galicia. El objetivo es el análisis de los productos activos en catálogo de una empresa de referencia para disponer de información sobre las publicaciones activas por las editoriales a nivel nacional.

Título. Definir un título que sea descriptivo para el dataset.

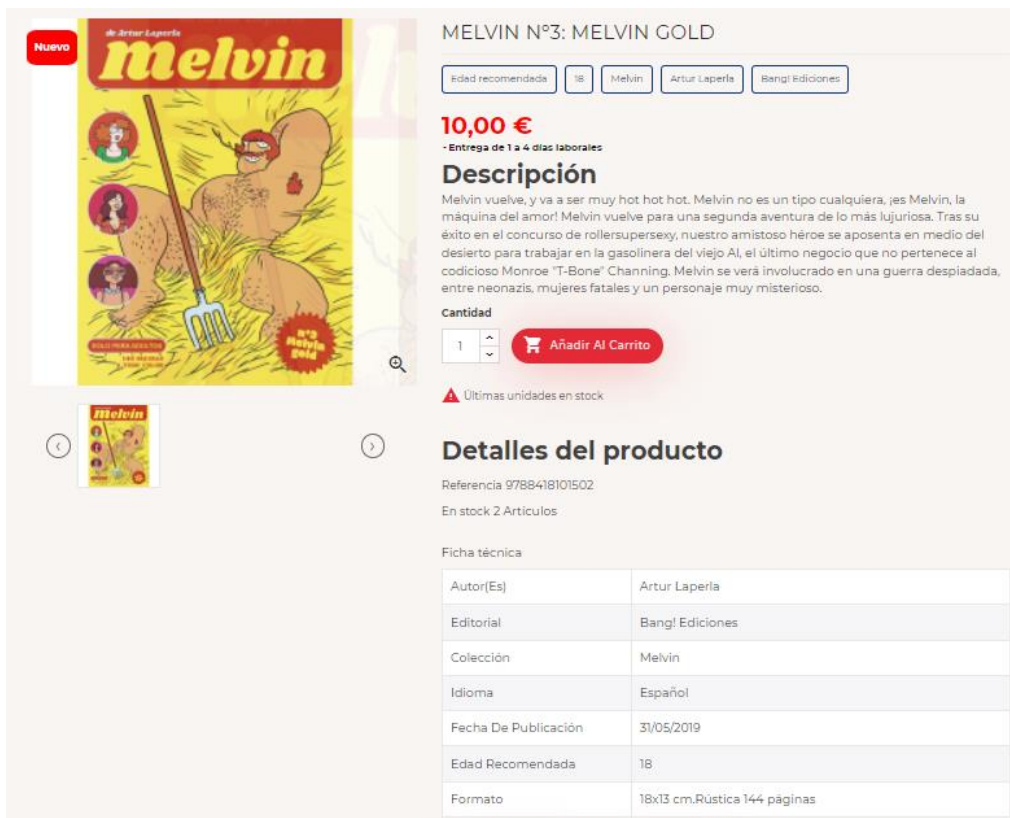
Catálogo de publicaciones del e-commerce de Alita Cómics.

Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

En este dataset se presenta el catálogo de publicaciones ofrecidas por la web de Alita Cómics a principios de abril de 2022. En el mismo se facilitan diversos campos de cada uno de los artículos para su análisis.

Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

Captura visual de los datos de un comic en la página web de alita.



The screenshot shows a product page for a comic book. On the left is the comic cover for 'Melvin Nº3: MELVIN GOLD', featuring a character in a red hat and a pitchfork. The right side of the page contains the following information:

- Product Title:** MELVIN Nº3: MELVIN GOLD
- Age Recommendation:** 18
- Author:** Artur Laperla
- Publisher:** Bang! Ediciones
- Price:** 10,00 €
- Delivery:** Entrega de 1 a 4 días laborales
- Description:** Melvin vuelve, y va a ser muy hot hot hot. Melvin no es un tipo cualquiera, ¡es Melvin, la máquina del amor! Melvin vuelve para una segunda aventura de lo más lujuriosa. Tras su éxito en el concurso de rollersupersexy, nuestro amistoso héroe se aposenta en medio del desierto para trabajar en la gasolinera del viejo Al, el último negocio que no pertenece al codicioso Monroe 'T-Bone' Channing. Melvin se verá involucrado en una guerra despiadada, entre neonazis, mujeres fatales y un personaje muy misterioso.
- Quantity:** 1 (with a dropdown arrow)
- Add to Cart:** Añadir Al Carrito
- Status:** Últimas unidades en stock
- Product Details:**
 - Referencia: 9788418101502
 - En stock: 2 Artículos
- Technical Sheet (Ficha técnica):**

Autor(Es)	Artur Laperla
Editorial	Bang! Ediciones
Colección	Melvin
Idioma	Español
Fecha De Publicación	31/05/2019
Edad Recomendada	18
Formato	18x13 cm. Rústica 144 páginas

- Comics
 - Reference
 - Price
 - date_add
 - name
 - description
 - category_name
 - link
 - quantity
 - features
 - Autor(es)
 - Editorial
 - Colección
 - Idioma
 - Fecha de publicación
 - Edad recomendada
 - Formato
 - ISBN

Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

En relación con el periodo de tiempo de recolección, estos datos recogidos muestran la “instantánea” de la mencionada web el pasado día 09 de abril de 2022.

Los campos que se han decidido obtener son:

- **'reference':** referencia del editor
- **'price':** muestra el precio del producto
- **'date_add':** fecha de publicación en el catálogo
- **'name':** nombre de la publicación
- **'description':** información adicional del artículo, si está disponible
- **'category_name':** clasificación del artículo, tipo de comic
- **'link':** URL en la cual el artículo ha sido publicado
- **'quantity':** indica la cantidad disponible
- **'Autor(es)':** indica el autor del cómic
- **'Editorial':** editorial de la publicación
- **'Colección':** indica a qué colección pertenece
- **'Idioma':** idioma de la publicación
- **'Fecha de publicación':** fecha en la cual el cómic fue publicado
- **'Edad recomendada':** edad mínima recomendada de lectura

- **‘Formato’**: formato de la publicación
- **‘ISBN’**: código ISBN del título

Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Como se ha indicado en diversos apartados anteriores, Alita Cómics (propiedad de Juan Carlos Sanmiguel Sesto) es quién ha alimentado el e-commerce del cual obtenemos el juego de datos.

Al ser información pública disponible en la web, el web scraping realizado cumple con la legalidad vigente. Los precios son públicos y la propiedad intelectual no se viola. A nivel ético: la obtención de la información recopilada no representa un perjuicio para el sitio web al no haberlo saturado mediante peticiones, al mismo tiempo que no se pretende presentar ni usar los datos como propios. También se ha tenido en cuenta el archivo <https://www.alitacomics.com/robots.txt> del sitio web para obtener el conjunto de datos.

Como ejemplo claro de este tipo de prácticas de análisis de competencia, tenemos el portal de Amazon que monitoriza constantemente la competencia para ajustar precios [1] y empresas que prestan sus servicios a tal efecto, como Nubiser. (<https://nubiser.com/servicio-scraping-a-medida/>).

[1] <https://www.retailgators.com/how-web-scraping-price-comparison-of-amazon-products.php>

Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Mediante este conjunto de datos, después de su análisis, se pueden definir y aplicar estrategias comerciales. Esas estrategias pueden ser de precio, de marketing, de mejora de la competitividad, de sincronización o de optimización del SEO, entre otras.

Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.

Released Under CC BY-SA 4.0 License. Por los siguientes motivos, tal y como se especifica en <https://creativecommons.org/licenses/by-sa/4.0/deed.es>:

- Se permite compartir el material, así como adaptarlo para cualquier uso, incluido el comercial.
- Atribución: Se debe hacer referencia al creador e indicar cualquier cambio.
- Compartir Igual: en caso de realizar cualquier contribución a posteriori, deberá distribuirse mediante la misma licencia.

Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R

<https://github.com/YesicaFdzt/Web-Scrapping>

Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

<https://doi.org/10.5281/zenodo.6449297>

Contribuciones	Firma
Investigación previa	YFR, CRS
Redacción de las respuestas	YFR, CRS
Desarrollo del código	YFR, CRS