PRA 1

Tipología y ciclo de vida de los datos

Jose Lainer

Lego Product Dataset

Contexto

La información recolectada proviene de la página web de Lego, que es una empresa que se dedica a la fabricación y venta de juguetes de construcción. El sitio web proporciona información detallada sobre los productos que ofrece, incluyendo su código, tema, nombre, precio en dólares, puntuación según las revisiones de los clientes, número de piezas y puntos VIP que se pueden obtener al comprar el producto. La información se recolectó mediante técnicas de web scraping con el objetivo de crear un dataset interesante y potencialmente útil para un proyecto analítico.

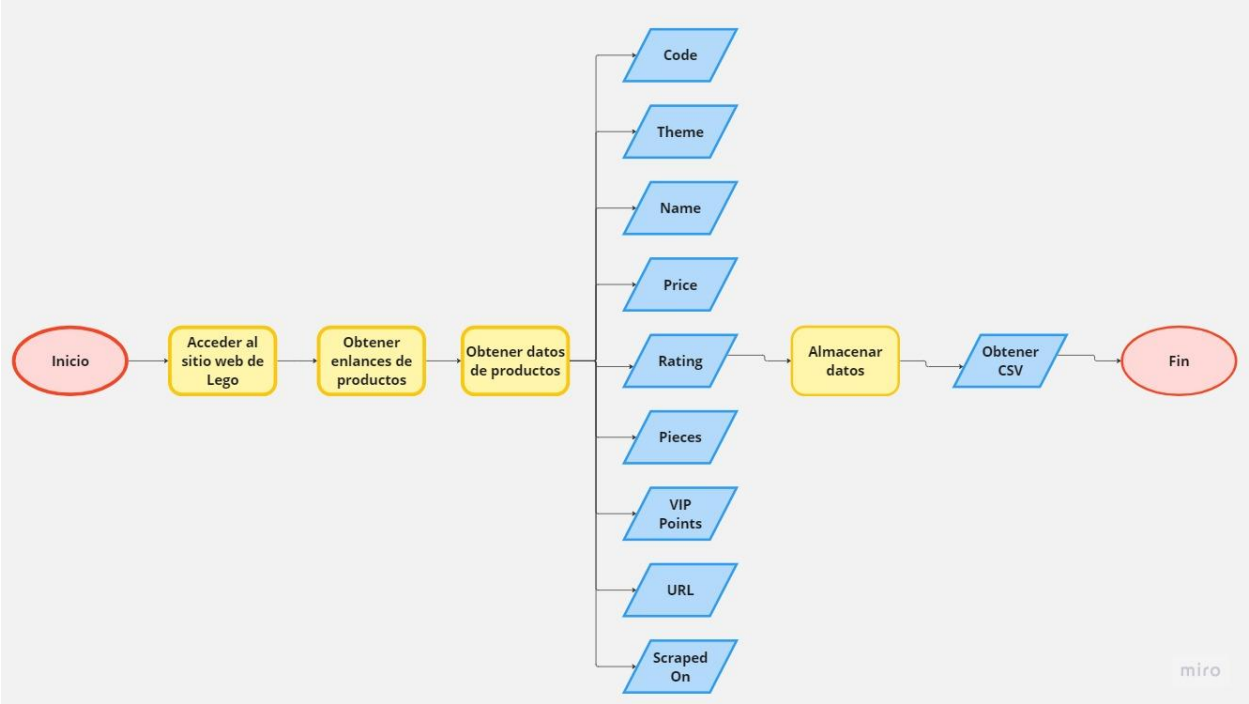
La elección de la página web de Lego se debe a que la empresa cuenta con una amplia gama de productos, y cada uno de ellos pertenece a un tema específico. Además, el sitio web proporciona una gran cantidad de información detallada sobre cada producto, lo que lo convierte en un sitio adecuado para extraer datos relevantes. La información recolectada puede ser utilizada para realizar análisis comparativos de productos por tema, identificar patrones de compra de los clientes, detectar tendencias en los precios y puntuaciones, y muchas otras aplicaciones potenciales en el ámbito del análisis de datos.

La dirección del sitio web es: https://www.lego.com/en-us/themes

Descripción del dataset

El conjunto de datos extraído del sitio web de Lego contiene información de cada uno de los productos disponibles en Abril del 2023, incluyendo su código, tema, nombre, precio en dólares, puntuación según las revisiones de los clientes, número de piezas, puntos VIP y el enlace al producto. Con un total de 759 registros, este conjunto de datos proporciona una amplia muestra de la variedad de productos que ofrece Lego, lo que lo convierte en una fuente de información valiosa para proyectos en diferentes áreas, desde análisis de mercado hasta estudios de preferencias de los clientes. Además, los datos han sido recolectados mediante técnicas de web scraping eficientes y adecuadas, garantizando la calidad y precisión de la información obtenida.

Representación gráfica



Contenido

Del sitio web de Lego, se obtienen los siguientes datos de cada producto:

- **Code**: El código de identificación único asignado a cada producto de Lego.
- **Theme**: El tema al que pertenece el producto, por ejemplo: Architecture, Star Wars, Harry Potter, Marvel, entre otros.
- **Name**: El nombre del producto Lego.
- **Price**: El precio del producto Lego en dólares estadounidenses.
- **Rating**: La puntuación promedio del producto según las reviews de los clientes.
- **Pieces**: El número de piezas incluidas en el producto Lego.
- **VIP Points**: Los puntos VIP que se otorgan al comprar el producto Lego y que se pueden gastar en descuentos para futuras compras, sets y artículos exclusivos.
- **URL**: El enlace URL que lleva a la página web del producto Lego.
- **Scraped On**: La fecha en la que se obtuvieron los datos del sitio web de Lego.

Los datos fueron extraídos del sitio web de Lego en una sola ocasión, y el período de tiempo específico al que pertenecen los datos es el 23 de abril de 2023. Entonces, es posible que algunos de los datos hayan sido actualizados en la página de Lego desde que se realizó la extracción inicial de datos.

Propietario

Dado que este conjunto de datos fue recolectado como parte de una actividad práctica, no existe un propietario específico para este conjunto de datos. Sin embargo, podemos considerar que la empresa Lego es la propietaria original de la información que se ha extraído de su página web.

En cuanto al análisis previo de este conjunto de datos, debido a su naturaleza académica y práctica, no se han encontrado análisis previos de este conjunto de datos específico. Sin embargo, podemos observar que el análisis de datos en el ámbito de los productos de Lego podría ser un tema de interés para analistas de datos.

En cuanto a la ética y legalidad en la recolección de datos, se siguieron los siguientes pasos para actuar de acuerdo con los principios éticos y legales:

- Se eligió un sitio web público y accesible para la extracción de datos.
- Se verificó el archivo robots.txt del sitio web de Lego para asegurarse de que se cumplían sus directrices en cuanto a la extracción de datos, y se siguió lo recomendado en dicho archivo para garantizar una extracción responsable y ética de los datos.
- Se reconoció a Lego como el propietario original de los datos recolectados y se utilizaron los datos exclusivamente para fines académicos y prácticos.

Inspiración

Este conjunto de datos extraído de la página web de Lego puede ser interesante para diversos fines, como por ejemplo:

- Análisis de precios y rentabilidad: La recolección de datos sobre los precios de los productos de Lego y los puntos VIP que se obtienen al comprarlos permiten analizar la rentabilidad de cada producto y determinar qué productos son más rentables para la empresa.
- Análisis de valoración de los clientes: La recolección de datos sobre la puntuación que los clientes otorgan a los productos de Lego permite analizar la valoración y percepción del público sobre cada producto.

Además, se podrían responder preguntas como:

- ¿Cuáles son los temas de productos de Lego más populares?
- ¿Cuál es el producto de Lego más caro y cuál es el más rentable?
- ¿Cuál es el tema de producto de Lego mejor valorado por los clientes?
- ¿Cómo se relacionan el precio y la puntuación del producto en la valoración de los clientes?

Licencia

Se eligió la licencia Creative Commons Attribution 4.0 International (CC-BY 4.0) para el conjunto de datos resultante porque permite el uso y la distribución de los datos de forma

amplia y libre, siempre y cuando se cite en todo momento al propietario de la recopilación de datos original. Esta licencia permite la creación de obras derivadas, así como su reutilización, y es compatible con la mayoría de plataformas y herramientas de análisis de datos. Además, esta licencia cumple los principios de apertura de datos y permite que otros investigadores utilicen y amplíen la recopilación de datos de forma transparente y ética.

Código

Mediante el comando `pip3 freeze > requirements.txt` se obtienen las librerías y versiones utilizadas en este trabajo:

- `async-generator==1.10`
- `attrs==23.1.0`
- `certifi==2022.12.7`
- `cffi==1.15.1`
- `exceptiongroup==1.1.1`
- `h11==0.14.0`
- `idna==3.4`
- `numpy==1.24.3`
- `outcome==1.2.0`
- `pandas==2.0.0`
- `pycparser==2.21`
- `PySocks==1.7.1`
- `python-dateutil==2.8.2`
- `pytz==2023.3`
- `selenium==4.9.0`
- `six==1.16.0`
- `sniffio==1.3.0`
- `sortedcontainers==2.4.0`
- `trio==0.22.0`
- `trio-websocket==0.10.2`
- `tzdata==2023.3`
- `urllib3==1.26.15`
- `wsproto==1.2.0`

El código utilizado para la recolección de datos se ha implementado en Python haciendo uso de la herramienta de scraping Selenium. El proceso de recolección de datos es:

1. Se accede al sitio y se aceptan las cookies.
2. Se obtienen los enlaces a cada tema de Lego.
3. Para cada tema, se obtienen los enlaces a cada uno de los productos.
4. Para cada producto, se obtienen los datos de interés.
5. Los datos se almacenan en una lista y luego se guardan en un archivo CSV.

La mayor dificultad del sitio web de Lego es que hay diferentes productos para cada tema, por lo que hay múltiples páginas en algunos temas. Entonces, para obtener los enlaces de todos los productos hay que acceder a todas las páginas o dar clic en el botón “Show All” que mostrará todos los productos (pero se deberá hacer scroll hacia abajo y esperar a que se carguen todos los productos).

Para resolver esto, primeramente se espera a que los elementos estén visibles y se agregan los links de los productos a una lista. Si existe un botón de "next" en la página, se hace clic en él y se repite el proceso para la siguiente página. Si no existe un botón de "next", se detiene la iteración y se devuelven los links recolectados. Una vez se tienen los links, se procede a obtener los datos de interés en cada producto.

Se adjunta el enlace del código en GitHub: https://github.com/jose-lainer/lego-crawler

Dataset

Se djunta el enlace del DOI del dataset: https://doi.org/10.5281/zenodo.7860808

Video

Se adjunta el enlace del video en Google Drive:
https://drive.google.com/drive/folders/1p59e_PU-Cj8_3LuHPcTpaCQfdvDqUxi-?usp=share_link