Tipología y ciclo de vida de los datos
WEB SCRAPING

Esteban Salazar - alessalazar
Juan Carlos Morales - jcmorales840801

April 13, 2019

Contents

1 Contexto

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

En un contexto competitivo, los portales se han convertido en la centralización de mercados para los clientes, es decir, plataformas como Amazon, Mercado Libre, Linio, entre otros han logrado que las grandes compañías, así como las personas de negocios pequeños e inclusive el mercadeo de productos de segunda mano converjan en un solo lugar, generando así un espectro amplio para el cliente final o consumidor. La página que se escogió para esta práctica es Mercadlo Libre, la cual está dedicada a las compras entre usuarios inscritos a su servicio de compras, ventas y pagos por Internet con enfoque en Latinoamérica, el objetivo de realizar el web scraping en esta página es hacer un análisis de mercado donde podamos converger los tres conceptos indicados anteriormente, el objetivo es analizar y comparar precios, configuraciones, observaciones de los clientes, inclusive visualización rápida de las imágenes de los productos que se están ofertando; asumiendo que este trabajo podría servir como base para un proyecto real, recolectar información como esta podría darnos la opción de brindar un servicio el cual le evite al cliente o a la casa matriz en este caso, LG o HP, una búsqueda exhaustiva de los productos que buscan o producen respectivamente, es decir, actualmente si se ingresa a dichos portales y se realizan búsquedas de portátiles de estas marcas, pueden aparecer más de 400 tipos diferentes de opciones, lo cual es bastante grande, cabe decir que estos portales cuenta con filtros que ayuden a reducir los resultados, pero aun así se debe ingresar página por página para saber sus características o calificaciones de los clientes que han utilizado estos proveedores anteriormente, basados en esta complejidad para visualizar la información de manera ágil, tener estos datos disponibles de una manera más sencilla, ayudaría con el ahorro de tiempo para no ingresar a cada link que oferte, mejorando la experiencia del usuario de la compra o el usuario de la investigación de manera más agradable.

2 Definir un título para el dataset.

Elegir un título que sea descriptivo

Shopping de precios y características.

Listado detallado en opciones de compra para portátiles LG y HP en portal de ventas por internet, indicando sus características, precio, observaciones e imágenes.

3 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).

El objetivo principal de este conjunto de datos esta segmentado en tres partes,

la primera es indicar las características del producto, tales como información, marcas, precio y características, las segunda es indicar las características del proveedor, como sus calificaciones, estado y opiniones, por último se agrega a este dataset el nombre y las imágenes que hacen parte de cada una de las páginas de las cual se espera poder extraer información.

Descripción del Data set:

El data set cuenta con 12 variables, de las cuales 10 variables son de tipo carácter y 2 variables numéricas, las variables y el tipo son:

- Título - Character
- Modelo - Character
- Marca - Character
- Tipo de pantalla - Character
- Precio - Integer
- Condición - Character
- Envío - Character
- Opiniones - Integer
- Vendedor - Character
- Tipo de vendedor - Character
- Nombre de la imagen - Imagen
- url - Character

Limitaciones del Data set:

Las variables de tipo carácter, presentan una novedad al bajar la información, cuando el texto tiene signos de puntuación o acentuación, se generan caracteres no convencionales, que dificultaran el procesamiento posterior al raspado, por lo que antes de comenzar con el análisis de la información se debe realizar una depuración de la base de datos extraída.
Por otra parte, cabe mencionar que no siempre la información de cada atributo extraído cuenta con las características estándares que se requieren para el análisis de la información, en cada campo se espera ver una serie de características para cada producto, donde después de realizar la depuración pertinente, el usuario final pueda realizar comparación de precios y características entre los productos de las diferentes marcas o fabricantes, para lo cual es indispensable que variables como el precio, sean lo más exactas posible

4 Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente

Figure 1: Esquema del Data Set

5 Contenido del Data Set

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

La información del raspado se obtiene como una fotografía en el momento de la consulta a la web, es decir, nos da el estatus del momento y puede que si se realiza el raspado en un momento de tiempo diferente, la información tenga cambios.

El conjunto de datos o dataset, está conformado por las siguientes columnas que contienen la siguiente información:

- Título: Hace alusión al título indicando por el vendedor del producto indicado en la plataforma
  - Ejemplo: HP 240 G6 con Win 10 Pro
- Modelo: Indica la versión del equipo, con las características que este tiene.
  - Ejemplo: 240 5 generación
- Marca: Indica la fábrica de quien produjo el portátil.
  - Ejemplo: HP (Hewlett Packard)

- Tipo Pantalla: Indica la característica o tecnología que contiene la pantalla.
  - Ejemplo: LED - LCD
- Esta columna confirma el precio del producto en el portal.
  - Ejemplo: $ 2'800.000 COP
- Condición: Con este campo se le hace saber al usuario cual ha sido la oferta del producto en el mercado.
  - Ejemplo: 350 vendidos desde 2019
- Envió: Indica si al comprar el producto él envió tiene algún coste o si este es gratis.
- Opiniones: Porcentaje a la calificación del proveedor del producto, este comienza desde 0 como muy malo hasta 5 como excelente.
- Vendedor: Indica la url de información del proveedor, en la cual se podrán encontrar datos mas detallados respecto a sus operaciones y calificaciones.
  - Ejemplo: https://perfil.mercadolibre.com.mx/GRUPODECME?brandId=1700
- Tipo Vendedor: El portal presenta categorías de proveedores, los cuales están basados en las calificaciones y comentarios hechos por sus compradores, ayudan al usuario final a entender la capacidad y estatus de quien ofrece los servicios.
  - Ejemplo: Mercado líder platinum
- Nombre Imagen: En este dataset también se están descargando las imágenes que se tienen de los productos en los encabezados de la página, por tal razón se está ingresando en esta columna el nombre de la imagen indicada al inicio, este mismo nombre eta siendo guardada en las carpetas de destino.
  - Ejemplo: Notebook HP 5 generación
- Url Imagen: indica las Url de las imágenes donde esta alojadas en la página y se guardan para su fácil acceso

6 Agradecimientos.

Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Se agradece el portal de Mercado Libre, por permitir el acceso a sus datos para realizar un Web Scraping de su página o raspado por sus términos en español, este material se utilizara con un único fin el cual es el académico, con el objetivo de entender la metodología de web scraping, funcionalidades, ventajas en el mercado etc. dicho material acá obtenido no se utilizara para ningún fin comercial y solo se gestionado para esta práctica académica.

Casos de éxito web scraping:

- La empresa Peryco.com es una compañía que se dedica al e-commerce de perfumería y cosmética, ellos utilizan la técnica para obtener información de los precios de los productos objetivo, y de acuerdo al análisis realizado con los precios, se encargan de canalizar a los usuarios con las mejores opciones de compra. [1]
- Un caso de éxito es el de la web TRIVAGO.COM es un comparador de precios en todo lo referente a hoteles en diferentes páginas web, lo que hace es tomar la mejor opción en cuanto a precio y direccionar al shopper al sitio web oficial del hotel, en este modelo de negocio se cobra por cada cliente direccionado al sitio oficial. [2]

Problemática alrededor de la información web

Este tipo de problemáticas es muy usual, dado que los consumidores finales en la actualidad están en busca de herramientas tecnológicas que les permitan optimizar el tiempo y el dinero, es por eso que existe un auge en cuanto al uso de nuevas tecnologías para el análisis y la explotación de la información, si empre enfocado a mejorar la experiencia del consumidor final, cosa que lo llevara a que el mismo sea fidelizado.

Valor agregado del scraper planteado:

Este trabajo permitirá al shopper tomar la mejor decisión en cuanto a la adquisición de aparatos tecnológicos en portales web, lo ideal es que con la información compilada se genere una sugerencia en cuanto al mejor precio encontrado para el equipo con las características solicitadas.

7 Inspiración.

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El interés de trabajar con un conjunto de datos obteniendo la información de un portal proviene por entender cómo se puede visualizar de una mejor manera para un cliente final que pueda tener varias dudas respecto al producto que está buscando una visualización de una manera más sencilla y fácil de leer, sin la necesidad de ingresar a muchas páginas para buscar el producto que realmente necesita el usuario final o comprador en este caso.
A su vez nos parece interesante si analizamos también el caso como una casa matriz, la cual sabemos que se encarga de la manufactura de estos productos, pero una vez producidos pasan a las manos de miles de comercializadoras, paginas

[1] http://www.odiseadigital.org/peryco-comparadores-webscraping-y-marketing-de-afiliacion-por-coterobarros
[2] https://www.trivago.com.co/

como esta podría ayudarles a ellos, como están siendo vendidos sus productos, que proveedores tienen mejores resultados y con qué productos en específico logran esta información.

8 Licencia.

Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above).
- Unknown License

De acuerdo con lo estipulado en el código legal creative commons1 y de acuerdo a las características del presente trabajo, se define para el mismo la licencia CC BY-NC-SA 4.0 License, la cual otorga los siguientes puntos a tener en cuenta a la hora de compartir y/o modificar el presente contenido, cito textualmente el contenido [3]

- **Atribución:** La persona que utilice el contenido del presente trabajo, debe dar crédito de manera correcta, referenciar el enlace en el cual se encuentra el contenido, y también debe indicar si se le realizan cambios o adiciones al contenido original.
- **No comercializar:** El presente contenido no se puede utilizar con fines lucrativos y/o comerciales
- **Compartir de la misma forma:** En este punto del código, se hace referencia a que si por algún motivo se llegase a distribuir la información del contenido revisado, ese debe distribuirse bajo la misma licencia original. El presente contenido no se puede utilizar con fines lucrativos y/o comerciales

9 Código.

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

Se ubica en el Github creado: *https://github.com/jcmorales840801/web-scrapping-prac-1*

[3] https://creativecommons.org/licenses/by-nc-sa/2.5/co/

10 Dataset.

Presentar el dataset en formato CSV

Se ubica en el Github creado: *https://github.com/jcmorales840801/web-scrapping-prac-1*

11 Integrantes.

Contribuciones	Firma
Investigación previa	ES - JCM
Redacción de las respuestas	ES - JCM
Desarrollo código	ES - JCM

12 Referencias.

References

[1] Subirats, L., Calvo, M. (2019). Web Scraping.

[2] Richard Lawson, (2015).Web Scraping with Python