Integrantes Equipo: Lisardo Gayán Tremps

José Luis Melo Santos

Curso: Master Ciencia de Datos

Asignatura : Tipología y ciclo de vida de los datos

Nombre Profresor : Diego Pérez Trenard

Actividad : Práctica 1

Fecha Entrega : 04/2019

PRACTICA 1

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

   Proyecto en Python realizado para la asignatura de Tipología y ciclo de vida de los datos en el cual se extraen los precios de todas las categorías de Media Markt, siendo este el portal de referencia para el mercado de consumo electrónico y que se puede tomar como base para la comparación de precios.

2. **Definir un título para el dataset.**

   Precios diarios Media Markt España.

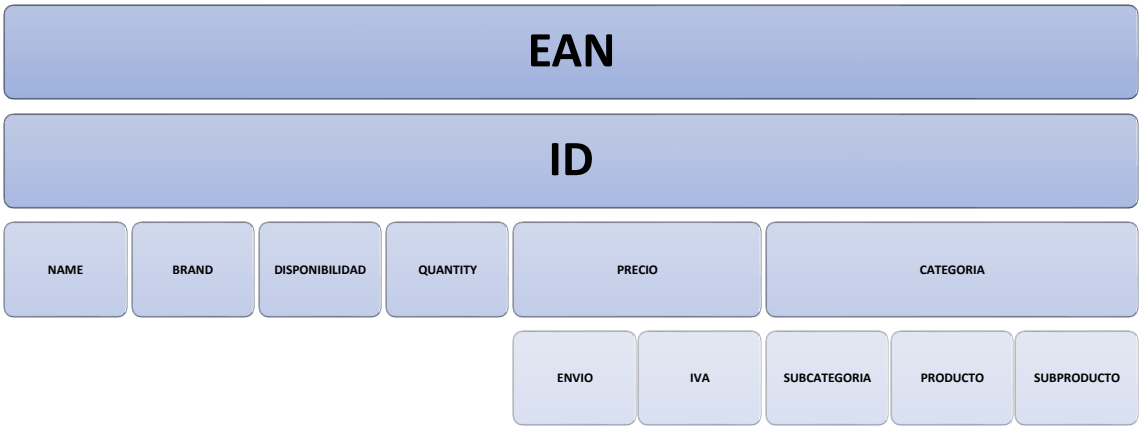
3. **Descripción del dataset.**

   Se genera un dataset o conjunto de datos en el fichero “productinfo_yyyymmdd.csv” (yyyymmdd, representa el día en que se extrajo la información), donde se detalla la información de todos y cada uno de los artículos listados en la web con los siguientes campos.

   - Ean
   - ID
   - Categoría
   - Subcategoría
   - Producto
   - Subproducto
   - Brand
   - Name
   - Precio
   - Quantity
   - Date
   - Iva Aplicado
   - Stock Status
   - Coste envío

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente**

Los campos y su relación se detallan a continuación.



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

Los campos incluidos son los antes mencionados y el título de cada uno de ellos describe en claridad su uso. Dado que los precios en la web se actualizan diariamente se ha decidido realizar una extracción inicial del sitio web español y alemán para comprobar el buen funcionamiento y escalabilidad del código, una vez confirmado, se ha creado un dataset por día durante 3 días para la realización de esta práctica. El scraper adjunto en el proyecto incluye la documentación de su funcionamiento.

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

La información de precios y marcas pertenece al grupo Media Markt España y se puede consultar en su página web en su formato original. La publicación de esta información solo es de carácter informativo y su objetivo es comprobar el funcionamiento del scraper.

Se analizaron las siguientes fuentes:

Titulo Articulo: Brinkhuis/Mediamarkt

Sitio Web: GitHub

URL: https://github.com/Brinkhuis/Mediamarkt

Titulo Articulo: Python web scraping with Beauifulsoup to extract variables

Sitio Web: Stack Overflow

URL: https://stackoverflow.com/questions/47186825/python-web-scraping-with-beauifulsoup-to-extract-variables

Titulo Articulo: Python BeautifulSoup scrape tables

Sitio Web: Stack Overflow

URL: https://stackoverflow.com/questions/18966368/python-beautifulsoup-scrape-tables

Titulo Articulo: Common User Agent List «Networking How To's

Sitio Web: Networkinghowtos.com

URL: http://www.networkinghowtos.com/howto/common-user-agent-list/

Titulo Articulo: Python's Requests Library (Guide) – Real Python

Sitio Web: Realpython.com

URL: https://realpython.com/python-requests/

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

Dado que Media Markt es la web de referencia para el mercado de consumo electrónico, es muy viable realizar un estudio del rango y variación de sus precios, ya sea para tener un parámetro de comparación para crear un sitio web o realizar un estudio del comportamiento de las marcas y precios.

En el código se incluye la creación del gráfico de la distribución de precios de una categoría para mostrar brevemente el posible análisis que se podría realizar a partir de los datos recabados. Además de que este proyecto es escalable a las webs de los diferentes países donde Media Markt tiene presencia.

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

Publicado bajo la licencia **CC BY-SA 4.0**

https://creativecommons.org/licenses/by-sa/4.0/

Esta licencia permite

Compartir - copiar o redistribuir el material en cualquier medio o formato.

Adaptar – combinar, transforma y construir sobre el material para cualquier propósito, incluso comercialmente.

Bajo los siguientes términos.

Atribución – Se debe dar crédito apropiado, proveer link a la licencia e indicar si se hicieron cambios. Esto se puede hacer de cualquier manera razonable, pero no en cualquier forma que sugiera que el licenciador te respaldo o su uso.

ShareAlike – Si se remezcla, transformas o construye sobre el mismo material, se debe distribuir las contribuciones bajo las misma licencia que la original.

El motivo para haber elegido esta licencia, es que la información que se obtiene con el proyecto, si bien es pública, el hecho de que cambie con el tiempo hace que resulte interesante conservarla, para realizar análisis futuros.

En cuanto al código se ha mejorado parte del código de un trabajo anterior que no tenía licencia definida. Hemos optado por hacer referencia a este trabajo y elegir una licencia que permita su uso, para que de igual forma nuestro trabajo sea mejorado en el futuro.

9. Código.

El fichero con el código se encuentra en Mediamarkt_Web_Scraping.ipynb.

También proporcionamos el mismo código en formato .py, para poder ejecutar en programas de Schedule como por ejemplo el task scheduler de Windows. En cualquier caso, es necesario crear las carpetas /data y /plots.

10. Dataset.

El fichero con la extracción de los datos se encuentra en la carpeta Data, fichero productinfo_"yyyymmdd".csv. Se genera un fichero por día.

Contribuciones	Firma
Investigación previa	L.G., J.L.M.
Redacción de las respuestas	L.G., J.L.M.
Desarrollo código	L.G., J.L.M.