Tipología y ciclo de vida de los datos. Práctica 1.

Autores: Gabriel Paladines y Jaime Pardo

Requisitos previos: instalar pip, builtwith, whois, requests, beautifulsoup4. Una vez instalado pip (https://www.liquidweb.com/kb/install-pip-windows/), el resto se pueden incluir mediante, por ejemplo, "pip install builtwith", y así sucesivamente.

Introducción: El objetivo de esta práctica es crear un dataset a partir de los datos contenidos en una página web.

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

Actualmente existe una pandemia que está afectando a muchos países en distinto grado. Uno de los problemas generalizados es la falta o escasez de determinados bienes de consumo, en particular comidas, bebidas y productos para el cuidado de la higiene. Además, es posible que se produzca un aumento de precios debido al cambio en las demandas.
El sitio web elegido es la página del centro comercial El Corte Inglés, que es uno de los centros de referencia en España para la adquisición de los productos que hemos nombrado.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Evolución de los precios y la disponibilidad de alimentos y otros productos para el hogar.

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 dataset contiene un listado de productos (básicamente de alimentación y cuidado de la higiene), para los cuales se puede ver el precio, la categoría y si tiene o no descuento. En el punto 5 desarrollaremos esta información.

El objetivo sería hacer una extracción de información cada 24h.

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

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Nombre,Marca,Categorias,Precio Original,Precio Final,Moneda,Estado,Descripcion,Fecha de captura														
2															
3	CARBONELL aceite de oliva suave 0,4º bidón 5 l,CARBONELL,"Alimentación General,Alimentación general,Aceites,Aceite de oliva",18.75,12.99,EUR,AVAILABLE,,2020-04-11														
4															
5	CARBONELL aceite de oliva suave 0,4º bidón 3 l,CARBONELL,"Alimentación General,Alimentación general,Aceites,Aceite de oliva",11.35,11.35,EUR,AVAILABLE,Lleva 3 y paga 2,2020-04-11														
6															
7	LA ESPAÑOLA aceite de oliva suave 0,4º botella 1 l,LA ESPAÑOLA,"Alimentación General,Alimentación general,Aceites,Aceite de oliva",3.99,3.79,EUR,AVAILABLE,Lleva 3 y paga 2,2020-04-11														
8															
9	LA ESPAÑOLA aceite de oliva intenso 1º botella 1 l,LA ESPAÑOLA,"Alimentación General,Alimentación general,Aceites,Aceite de oliva",3.99,3.79,EUR,AVAILABLE,Lleva 3 y paga 2,2020-04-11														
10															

El dataset incluye los siguientes campos:

Nombre: nombre completo del producto, suele incluir el formato
Marca: marca del producto
Categorías: grupo de hasta 4 categorías, de más general a más específica
Precio original: el precio original aparece debido a que hemos decidido filtrar para mostrar únicamente productos en oferta.
Precio final: precio de compra para el cliente
Moneda: Euro
Estado: disponibilidad (o no) del producto
Descripción: texto descriptivo de la oferta. Por ejemplo: “lleva 3 y paga 2”
Fecha de captura: fecha del momento de la extracción

El periodo de tiempo previsto es de 24h.

La recogida de datos se ha hecho, por simplicidad, únicamente con la primera página de la búsqueda y con los productos con descuento. Para ello se ha creado un script en Python, el cual mediante la técnica de web scraping extrae los productos directamente desde la siguiente página:

https://www.elcorteingles.es/ofertas-supermercado/

En dicha página identificamos las etiquetas y extraemos la información necesaria. Por ejemplo:


```
<div class="c12 grid-header">…</div>
<div class="c12 js-grid-container"> event
<div class=" grid c12 js-hidding-pages js-page-1 MAIN" data-page="1"> flex
<div class=" grid-item product_tile _supermarket dataholder js-product " data-scope="product" data-json="
{"id":"0110120902700044___","brand":"CARBONELL","category":[_nt":true,"status":"AVAILABLE","quantity":1,"currency":"EUR"}" data-
product-stock="true" data-product-current="0" data-product-catalog="supermarket" data-product-description="CARBONELL aceite de
oliva suave 0,4º bidón 5 l" data-product-id="0110120902700044___" data-product-sale_type="SELLING_TYPE_UNIT" data-product-
trigger="PLP"> flex
```

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

Agradecemos a la compañía El Corte Inglés por facilitar la captura de datos desde su página web.

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

Este conjunto de datos es interesante porque permite obtener un listado de productos con descuento, así como hacer una comparativa de precios para ver su evolución. Es especialmente relevante en situación de pandemia.

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**

El trabajo realizado en esta práctica tiene una finalidad educativa, en ningún caso comercial. Es por ello que la licencia elegida ha sido **CC BY-NC-SA 4.0** (“reconocimiento no comercial sin obra derivada”). Esta licencia permite compartir el

dataset. Sin embargo, al contrario que la CC BY-SA 4.0, no permite un uso comercial. Por el mismo motivo se ha descartado ofrecer una licencia de dominio público.

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

Se puede ver el código en el siguiente enlace:
https://github.com/gabrielpaladines/web-scraping

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

La publicación se ha llevado a cabo en https://zenodo.org/ y se puede encontrar en:
https://zenodo.org/record/3748734

El formado es ".csv" separado por comas (no punto y coma).

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

El DOI (Digital object identifier) proporcionado por Github/Zenodo es: