Crear un dataset a partir de los datos contenidos en una web

1. Títol del dataset.

Listado productos

2. Subtítol del dataset.

Listado de productos y precios de la tienda online de Mediamarkt

3. Imatge.

Haciendo un poco juego con la publicidad de Media Markt, he usado a modo irónico esta imagen[1], como imagen representativa del dataset, ya que lo que pretendemos con este dataset es ver si la empresa nos toma el pelo con las ofertas en días especiales.

[1] Origen de la foto: https://m.forococoches.com/foro/showthread.php?t=5279959&page=5

4. Context.

El conjunto de datos representa un listado con todos los productos disponibles en la tienda online de Mediamarkt organizados por categoría. Se genera un fichero por cada día que se ejecuta el programa (el nombre del .csv incluye la fecha) y se obtiene el listado de productos actualizado (más de 23000), incluyendo el precio y si está en oferta o no.

Este dataset podría servir indistintamente para cualquier tienda online que quisiéramos analizar, pero habría que adaptar el scraping para que recupere la información de manera correcta

5. Contingut.

Los campos que incluye el dataset son:

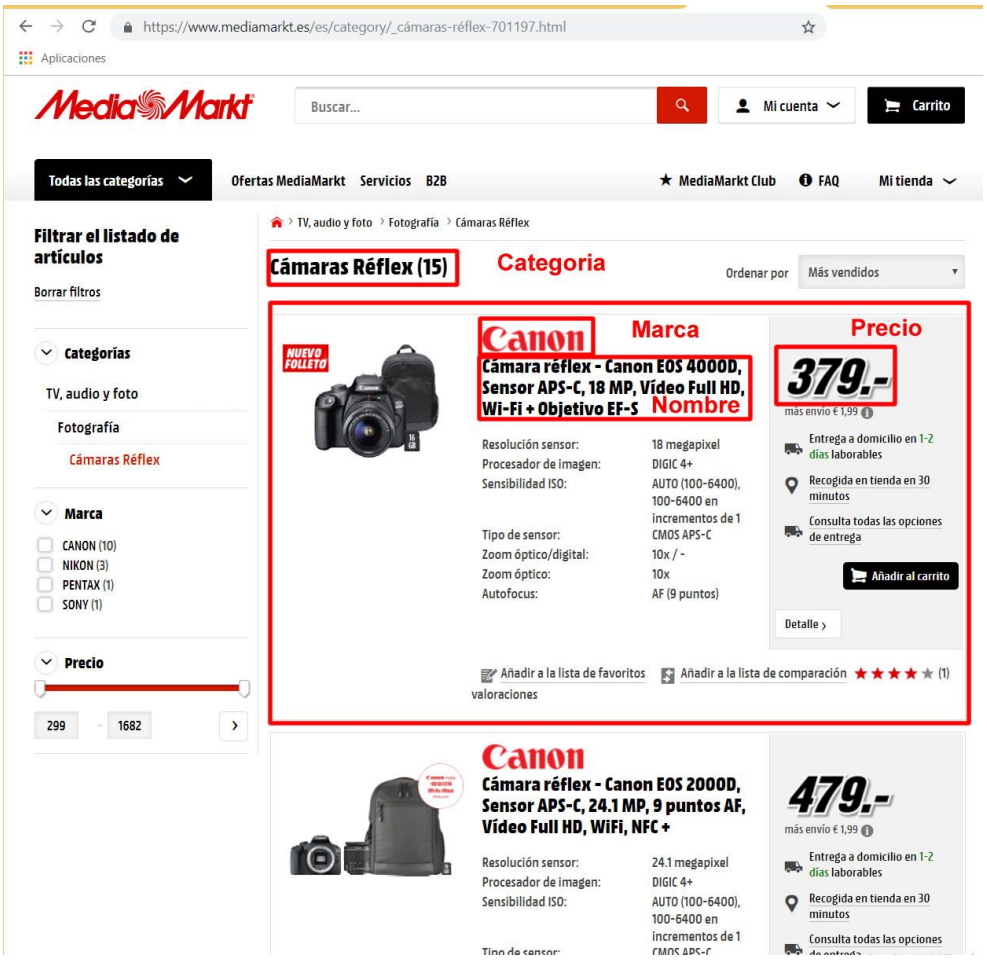
categoría	categoría a la que pertenece el producto
identificador	identificador del producto en la tienda. Cada producto diferente tiene su propio identificador que es único y que nos ayuda a identificar unívocamente el producto
nombre	Nombre completo del producto
marca	Fabricante del producto
Precio actual	PVP de venta actual del producto
precio anterior	En caso de estar en oferta, este campo contiene el precio anterior, para que el cliente sepa qué descuento se le ha aplicado

Tras revisar el robots.txt, en el cual no se restringe el acceso a los datos, obtenemos la ubicación del sitemap y a través del sitemap accedemos a las diferentes subsecciones. Para realizar el scraping hemos usado la lista de productos[2] que contiene un listado con más de 500 categorías disponibles.

Para obtener todos los productos iteramos por todas las listas y recuperamos los productos listados. No hace falta entrar en el detalle de cada producto, porque la información que necesitamos está en la propia lista.
En algunas categorías, existe paginación, así que recorremos todas las páginas de todas las categorías.

[2] https://www.mediamarkt.es/sitemap/sitemap-productlist.xml

Para recuperar la información de cada producto, nos fijamos en la siguiente figura que contiene un ejemplo de listado. En la figura se aprecian claramente de qué partes de la web obtenemos cada campo del dataset, a exception del identificador del producto, que no se ve a simple vista, pero que al inspeccionar el código se aprecia claramente



Por cada ejecución del programa se genera un fichero .csv que contiene un listado de todos los productos de la tienda online, con su precio y si tiene descuento o no.
El nombre del fichero tiene el siguiente formato: 181111_listaproductos.csv, donde los primero caracteres representan la fecha de ejecución YYMMDD_listadeproductos.csv
Para poder hacer un buen análisis se debería ejecutar el programa todos los días,sobre todo en los días de promociones especiales (black friday, cyber monday, días sin iva, nochebuena, reyes...) y comparar la evolución de los precios por productos, categorías o marcas.
La ejecución del programa lleva un tiempo, entre 30min y 1hora, ya que ejecuta más de 1000 llamadas a urls distintas y viene a costar del orden de 2-5s procesar cada petición.

6. Agraïments.

Los datos han sido obtenidos de la página web de mediamarkt, cuyo propietario es Media Saturn Multichannel S.A.U
La OCU[3] ha realizado varios estudios alertando de los supuesto descuentos de las principales tiendas online que nos han servido de inspiración
https://www.ocu.org/consumo-familia/compras-online/noticias/precios-black-friday

Según los estudios de la OCU:

> Desde este punto de vista, para los productos de nuestro estudio, los resultados de este Black Friday son muy diferentes:
> - Casi la mitad de los productos (48%) tenían en este Black Friday el mismo precio que el mínimo visto durante el mes anterior.
> - Un 41% tenían un precio superior a ese mínimo del periodo.
> - Sólo un 11% cumplirían los requisitos para poder indicar un precio inferior al mínimo alcanzado en el mes anterior.
>
> Es decir, apenas uno de cada diez productos se puede ver más barato estos días que en el mes precedente, mientras que cuatro de cada diez están más caros.

7. Inspiració.

Con el listado de productos y precios se pueden hacer muchas cosas, en primer lugar se podrían sacar estadísticas del número de productos por categoria o qué productos tienen mayores descuentos o cual es la tendencia de los precios.

Pero mi intención a la hora de seleccionar los datos es que hay muchas quejas por parte de los clientes de la tienda relacionados con que la empresa sube los precios antes de determinados días especiales, como son los días sin iva, black friday, cyber Monday o navidades. [1, 2, 3, 4 ,5 ,6]

El objetivo es lanzar el programa y obtener los precios de todos los productos todos los días para luego poder analizar los datos y ver si la empresa infla los precios de los productos antes de dichos eventos. Está claro que algún cliente que sigue un producto lo puede apreciaR y puede poner una queja, pero estaría bien saber, en caso que haga esa mala práctica, saber en cuantos productos lo hacen, qué cantidad de producto quitan de la web, a qué categorías pertenecen y en cuanto incrementa el precio y si al final hay descuento real en comparación con el precio medio anual de los productos.

[3] Oficina del Consumidor

[1] https://www.elconfidencial.com/tecnologia/2015-12-01/la-ocu-acusa-a-media-markt-de-fraude-por-manipular-los-precios-del-black-friday-y-el-cyber-monday_1111047/
[2] http://catalania.blogspot.com/2013/02/mediamarkt-la-estafa-del-dia-sin-iva.html
[3] https://www.adslzone.net/article10619-las-trampas-de-mediamarkt-en-el-dia-sin-iva.html
[4] https://www.ofertitas.es/toda-la-verdad-sobre-el-dia-sin-iva/4628/
[5] https://www.meneame.net/m/tecnolog%C3%ADa/mediamarkt-retira-productos-sube-precios-antes-dia-sin-iva
[6] https://www.meneame.net/story/dia-sin-iva-mediamarkt-no-ellos-no-son-tontos

8. Llicència.

9. Codi:

El código completo del scraping se encuentra en el fichero mediamarket_prueba_2.py

10. Dataset:

En github se puede encontrar el dataset “181111_listaproductos.csv” correspondiente al día 11 de Noviembre de 2018 y contiene más de 23000 registros

	A	B	C	D	E	F
1	Afeitadoras faciales	1385320	Afeitadora - Philips S7522/50, Anillos SkinGlide, Especial pieles sensibles, seco y húmedo, 50 min	PHILIPS	172	SINDESCUENTO
2	Afeitadoras faciales	1187436	Afeitadora - Philips AT750/26, Recargable, Cabezales redondeados, Completamente lavable	PHILIPS	57,99	68,99
3	Afeitadoras faciales	1418923	PANASONIC ES-CV51-S803	PANASONIC	192	SINDESCUENTO
4	Afeitadoras faciales	1373773	Afeitadora - Braun Series 5 5195CC, Wet&Dry, Estación de limpieza y carga, Cabezal 8 direcciones,	BRAUN	254	SINDESCUENTO
5	Afeitadoras faciales	1418922	PANASONIC ES-LV6N-S803	PANASONIC	244	SINDESCUENTO
6	Afeitadoras faciales	1418921	PANASONIC ES-LV6Q-S803	PANASONIC	304	328
7	Afeitadoras faciales	1418557	Afeitadora - River 03615-1016 Travel Shave, Recargable, Sin cables, Con bolsa para viajes y		19,99	SINDESCUENTO
8	Afeitadoras faciales	1402906	Pack - Afeitadora Braun Serie 7 7893S + Cepillo Oral B pro 500, Wet & Dry, 5 modos, Flexible,	BRAUN	269	SINDESCUENTO
9	Afeitadoras faciales	1418920	PANASONIC Afeitadora - Panasonic ES-ST 3 N, 3 cuchillas, LED, Seco/Humedo, Negro	PANASONIC	79	SINDESCUENTO
10	Afeitadoras faciales	1376067	Afeitadora - Remington PR1340GP, Cabezal comfort pivot, Con cable + Cortador facial	REMINGTON	49,9	SINDESCUENTO
11	Afeitadoras multifunción	1381617	Afeitadora - Philips Multigroom Series 3000 MG3730/15, Multifunción, 8 accesorios, 60 minutos	PHILIPS	23,2	SINDESCUENTO
12	Afeitadoras multifunción	1406321	Afeitadora corporal - Philips BG3011/15, Apta para la ducha, 1 peine-guía, 50 minutos autonomía	PHILIPS	29	SINDESCUENTO
13	Afeitadoras multifunción	1381614	Barbero - Cortapelos - Afeitadora multifunción Philips MG7715/15, 13 en 1, seco y húmedo,	PHILIPS	54,99	SINDESCUENTO
14	Afeitadoras multifunción	1381615	Barbero - Cortapelos - Afeitadora Multifunción - Philips MG7710/15, 12 accesorios, Autonomía 120	PHILIPS	48,99	SINDESCUENTO
15	Afeitadoras multifunción	1432974	BG7020/15	PHILIPS	59	SINDESCUENTO
16	Afeitadoras multifunción	1381616	Afeitadora - Philips Multigroom Series MG5730, Multifunción, 11 accesorios, 80 minutos de	PHILIPS	48,99	SINDESCUENTO
17	Afeitadoras multifunción	1402855	Barbero - Cortapelos - Afeitadora multifunción - Braun MGK3045, Set de afeitado 7 en 1	BRAUN	40,99	SINDESCUENTO
18	Afeitadoras multifunción	1172267	Barbero - Cortapelos --Afeitadora Multifunción Babyliss E 823 E KIT MULTI TRIMMER-6 RECHARG		23,99	SINDESCUENTO
19	Afeitadoras multifunción	1427942	TN8960 MULTISTYLER 9 EN 1 BASIC	ROWENTA	29,99	SINDESCUENTO
20	Afeitadoras multifunción	1159932	Afeitadora corporal - Philips TT2039/32 Recargable, 5 medidas de corte, Completamente lavable	PHILIPS	29,9	SINDESCUENTO
21	Afeitadoras multifunción	1172490	Barbero - Cortapelos - Afeitadora multifunción - Babyliss E 824 E KIT MULTI TRIMMER-8 RECHARG		33,99	39,99
22	Afeitadoras multifunción	1167616	Afeitadora corporal - Philips BG2025/15, Seco y húmedo, Recargable, Completamente lavable	PHILIPS	29	SINDESCUENTO
23	Afeitadoras multifunción	1433055	Recortador - Philips Multigroom Series 5000 MG5730/18, Tecnología DualCut, Resistente al agua, Negro	PHILIPS	48,99	SINDESCUENTO
24	Afeitadoras multifunción	1430907	MGK3980 BLK/BLK	BRAUN	69	SINDESCUENTO
25	Afeitadoras multifunción	1414040	Afeitadora corporal - Philips BG5020/15, Apta para la ducha, 3 peines/guía, Recargable	PHILIPS	48,99	SINDESCUENTO
26	Afeitadoras multifunción	1117615	Afeitadora corporal - Philips Bodygroom TT2040/32 Recargable, Funcionamiento bajo la ducha,	PHILIPS	56,99	59,99
27	Cortapelos	1418668	Cortapelos - Braun HC5090 + Gillette, 17 ajustes de longitud, Base cargadora, Plata	BRAUN	62,99	SINDESCUENTO
28	Cortapelos	1226121	Cortapelos - Philips HC5440/ 80 Cuchillas de acero inoxidable, 24 posiciones de longitud, uso sin	PHILIPS	33,99	SINDESCUENTO
29	Cortapelos	1279483	Cortapelos - Philips HC5438/ 80 Cuchillas de acero inoxidable, 24 posiciones de longitud, Uso sin	PHILIPS	27,99	29,99
30	Cortapelos	1215629	Cortapelos - Tristar TR-2587, Para nariz y orejas, Con cubierta, Funciona con pilas	TRISTAR	2,99	SINDESCUENTO
31	Cortapelos	1308076	Cortapelos - Wahl Nasal Trimmer 5642-135, Nariz, Orejas, Cabezal lavable, Inox	WAHL	5,99	6,99

11. Conclusión

La práctica ha servido para aprender a usar scraping para recuperar información de una web, pero para completar el estudio habría que analizar la información.

Para poder analizar los datos, se debería ejecutar el programa diariamente y cargar los datos en una herramienta tipo Elasticseach y visualizar las tendencias por medio de Kibana. De este modo podríamos de manera dinámica crear las gráficas que necesitemos:

- Ver tendencia de un productos
- Ver tendencia de todos los productos de una categoría
- Ver tendencia de todos los productos
- Ver tendencias agrupando los descuentos por categorías.

Otra opción sería añadir datos de otras tiendas para analizar el comportamiento conjunto.