

Tipología y Ciclo de Vida de los Datos

Morten Nyborg

Práctica 1

Introducción

Este informe presenta información adicional para comprender el contenido de los datos entregados para la Práctica 1 de la asignatura Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de datos de la UOC.

Esta práctica consiste en la aplicación de técnicas de web scraping en todas las páginas de H&M en Europa para poder conseguir todos los datos que hay en la sección de ofertas en los distintos países de Europa en los que H&M tiene esta sección.

El resultado del web scraping en forma de base de datos y el código que hemos usado para conseguir la información lo compartimos en Zenodo y GitHub.

Esta práctica ha sido elaborada de forma individual por Morten Nyborg.

1. Contexto.

Explicar en qué contexto se ha recolectado la información.

H&M es una marca de ropa originada en Suecia que hoy tiene 4664 tiendas en 77 mercados (<https://hmgroup.com/about-us/markets-and-expansion/market-overview/>, agosto 2022).



Facts and figures

Net sales SEK 198,967 billion in 2021.
Around ~4,664 stores in 77 markets and 57 online markets.
In the H&M group, 71% of the employees in positions of responsibility were women in 2021.
Tripled the share of recycled materials used in our garments from 5.8% to 17.9% towards our goal 30% recycled materials by 2025.
We have reduced plastic packaging by 27.8% compared to our baseline year 2018.

Las tiendas de la marca se encuentran en todos los continentes del mundo, aunque la mayoría de ellas está en Europa.

AFRICA

- Egypt | مصر
- Morocco | المغرب
- South Africa
- Tunisia | تونس

ASIA

- Cambodia
- China's Mainland (en) | 中国大陆
- Hong Kong SAR (en) | 香港特別行政區
- India
- Indonesia
- Japan | 日本
- Kazakhstan | Қазақстан
- Macao SAR (en) | 澳門特別行政區
- Malaysia
- Philippines
- Singapore
- South Korea | 대한민국
- Taiwan Region (en) | 台灣地區
- Thailand | ประเทศไทย
- Vietnam | Việt Nam

EUROPE

- Austria | Österreich
- Belarus | Беларусь
- Belgium | België | Belgique
- Bosnia & Herzegovina | Bosna i Hercegovina
- Bulgaria | България
- Croatia
- Cyprus
- Czech Republic | Česká republika
- Denmark | Danmark
- Estonia
- Finland (sv) | Suomi
- France
- Georgia | საქართველო
- Germany | Deutschland
- Greece | Ελλάδα
- Hungary | Magyarország
- Iceland | Ísland
- Ireland
- Israel | ישראל
- Italy | Italia

MIDDLE EAST

- Kosovo
- Latvia
- Lithuania
- Luxembourg
- Netherlands | Nederland
- North Macedonia
- Norway | Norge
- Poland | Polska
- Portugal
- Romania | România
- Russia | Россия
- Serbia
- Slovakia | Slovenská republika
- Slovenia
- Spain | España
- Sweden | Sverige
- Switzerland | Schweiz | Svizzera | Suisse
- Turkey | Türkiye
- Ukraine | Україна
- United Kingdom

NORTH AND SOUTH AMERICA

- Bahrain | البحرين
- Jordan | الأردن
- Kuwait | الكويت
- Lebanon | لبنان
- Oman | سلطنة عمان
- Qatar | قطر
- Saudi Arabia | السعودية
- United Arab Emirates | الإمارات

NORTH AND SOUTH AMERICA

- Canada (fr) | Canada (en)
- Chile
- Colombia
- Costa Rica
- Ecuador
- Guatemala
- Mexico | México
- Panama
- Peru | Perú
- United States
- Uruguay

OCEANIA

- Australia
- New Zealand

La marca dispone de productos para mujer, hombre, bebé, niños, hogar, deporte y las distintas tiendas tienen surtidos de productos diferentes. Para cada categoría de cliente (mujer, hombre, etc) las páginas contienen o no una sección de ofertas. La existencia o no de ofertas de distintos tipos entre las páginas de la marca en distintos países nos ha llamado la atención, por lo tanto, queremos crear un dataset que permita analizar y comparar las ofertas en distintas categorías en distintos países de Europa para ver si hay alguna diferencia entre ellos.



Para hacernos una idea de la estructura de la página miramos en webs de H&M en distintos países y comprobamos que no todos los países tienen ofertas, por ejemplo observamos que en Noruega la marca no tiene una sección de ofertas, mientras que en otros países hay múltiples ofertas.

El dataset que proporcionamos contiene todos los productos ofertados en los distintos países, junto a la categoría de producto (hombre, mujer...), la web detallada del producto y puede ser útil para analizar las distintas campañas que tiene la marca H&M en distintos países de Europa y intentar descubrir patrones en las ofertas que existen, ver si hay campañas similares en países cercanos cultural o geográficamente, ver si hay diferencias en los precios de los productos ofertados entre países con menores y mayores rentas.

Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

Los datos que proporcionamos vienen del sitio web <https://www.hm.com/entrance.ahtml> que nos dirige a los distintos mercados que tiene H&M en el mundo. El objetivo de nuestro proyecto es conseguir los datos descritos en Europa, por lo tanto nos limitamos solamente a los países que H&M ha agrupado bajo Europa..

Para conseguir los datos que queremos, en cada país buscamos la sección de ofertas para cada categoría de producto: sabemos que existen ofertas en hombre, mujer, niño, etc, y todas ellas las podemos identificar con la palabra “deals”, “offers”, “angebote” en la url.

The sidebar on the right side of the page contains three main sections:

- Novedades**
 - Ver todo
 - Prendas
 - Zapatos y Accesorios
 - Sport
- Lookbook de otoño**
 - Estilos casuales
 - Looks elegantes
 - Looks urbanos
- Ofertas**
 - Multipacks desde 7,99 €
 - Calcetines - 3 x 2

Al seguir el link de una de las promociones que ofrece H&M en España encontramos los calcetines 3x2. Aquí buscamos extraer el título del producto, su precio y un vínculo a la página del producto.

CALCETINES - 3 X 2

ORDENAR POR ▾ TALLA ▾ COLOR ▾ ESTAMPADO ▾ ESTILO ▾ TODOS LOS FILTROS 112 artículos [Modelo](#) [Producto](#)

3 por 2 3 por 2 3 por 2 3 por 2

Calcetines en punto jacquard 2,99 € ● ● ● +8 Nueva Colección	Calcetines en punto jacquard 2,99 € ● ● ● +8 Nueva Colección	Calcetines estampados 3,99 € ● ● ● +17 Nueva Colección	Calcetines en punto jacquard 2,99 € ● ● ● +8 Nueva Colección
---	---	---	---

All llegar al final de la lista, cargaremos más productos entrando en la url detrás de “MOSTRAR MÁS”.



Repetimos el proceso descrito en todos los países de Europa.

2. Título

Ofertas de H&M en Europa: tipo de oferta, nombre del producto y precio.

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 información sobre todos los productos de H&M que se encuentran en la sección de ofertas en todas las páginas web de la marca en países de Europa.

4. Representación gráfica.

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

Representamos gráficamente el dataset con un diagrama de elaboración propia que muestra los pasos que hemos seguido desde la identificación del punto de partida del ejercicio de scraping hasta la publicación de los resultados y el código usado para conseguirlos en Zenodo y Github.



5. Contenido.

Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

El dataset consta de los siguientes campos:

main page - la dirección URL de la oferta donde puedes encontrar el producto (cadena de caracteres) Ejemplo típico:

https://www2.hm.com/es_es/hombre/deals/calcetines-3x2.html

name - el nombre del producto (cadena de caracteres) Ejemplo típico:

Calcetines en punto jacquard

url - la dirección URL del producto (cadena de caracteres) Ejemplo típico:

/es_es/productpage.0701134069.html

price - el precio del producto en moneda local (cadena de caracteres) Ejemplo típico:

"2,99 €"

Tenemos 42910 registros en el dataset.

Hay 19 países con ofertas.

El periodo de tiempo de los datos 16/11/2022

Nota sobre los links: un link lo consideramos como un dato único en nuestro dataset, pero el usuario puede tomar en consideración la estructura de dichos enlaces para obtener más información. Los enlaces de main_page contienen datos del tipo de cliente (hombre, mujer, niño, bebé, hogar, etc), del país y el idioma (es_es), del nombre de la oferta (calcetines-3x2).

6. Propietario.

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

El propietario del conjunto de datos es H & M Hennes & Mauritz AB, así lo podemos comprobar en el pie de página de cualquiera de las páginas web de la marca. En su página H&M nos permite ver el número de productos que están bajo un tipo de oferta (parte superior derecha), pero para obtener el número de productos en oferta en todo el país para todos los tipos de clientes, tendríamos que navegar a cada categoría y contar el número de productos en oferta. Por ejemplo Hombre - Multipacs, Hombre - Calcetines 3x2, Mujeres - Multipacs, etc. Esto no nos sorprende porque una tienda online no tiene como objetivo principal proporcionar estadísticas sobre los productos que vende.

Nuestro dataset reune los datos del nombre de la oferta en la url, además contiene la descripción del producto, el precio y el código del producto en la url y su presentación en forma de csv facilitará al usuario manipular los datos de varias maneras.



Por otra parte, tenemos datos de productos de H&M y sus imágenes en varios juegos de datos en Kaggle, como por ejemplo <https://www.kaggle.com/datasets/odins0n/handm-dataset-128x128> o <https://www.kaggle.com/datasets/odins0n/hm256x256>.

Los datasets parecen ser específicos de un mercado y ninguno de estos dos juegos de datos contiene información sobre las ofertas de la marca.

Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Para asegurarnos que estamos consiguiendo los datos de forma legal hemos consultado el fichero robots.txt y ahí hemos visto que todas las restricciones para los scrapers están relacionadas con cuentas de usuario, cesta y proceso de compra.

En la parte de condiciones de uso tampoco hemos encontrado ninguna información que nos indique que no podemos usar los datos, en este apartado toda la información está relacionada con la privacidad del usuario y el uso de sus datos

(https://www2.hm.com/en_sg/customer-service/legal-and-privacy/privacy-link.html,

https://www2.hm.com/es_es/service-clients/legal-and-privacy/privacy-link.html,

<https://hmgroup.com/legal-notice/>). Al haber consultado las páginas mencionadas hemos llegado a la conclusión que podemos obtener los datos sin estar violando las normas de uso de H&M.

Para realizar el web scraping hemos usado Scrapy, que en su configuración permite introducir retardos entre solicitudes y también respetar las recomendaciones del fichero robots.txt y hemos hecho uso de ambas funcionalidades para asegurarnos poder continuar usando los datos como para

no saturar el servidor. Para conseguir los datos hemos usado un user agent con los datos de nuestro navegador y sistema operativo ("Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36").

Adicionalmente, en el libro de Web Scraping with Python de Richard Lawson encontramos que está permitido hacer web scraping y publicación de información que es un hecho, por ejemplo, ubicaciones de las tiendas y horarios. Así por ejemplo lo ha determinado el juez en el caso ofir.dk vs home.dk. Nosotros nos encontramos justamente en el caso de conseguir y publicar la información que la misma empresa ha compartido sobre sus tiendas.

7. Inspiración.

Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Podemos conseguir los datos de los productos en oferta navegando por cada una de las secciones de los distintos mercados de la marca en Europa, aunque este método de conseguir los datos requiere mucho tiempo incluso si se tratara de un solo país. Como hemos comentado en el apartado 6, existen algunos conjuntos de datos de productos de H&M con imágenes que contienen todos los productos de la marca en un mercado. Nuestro dataset se diferencia de ellos por centrarse solamente en ofertas de productos y abarcar varios mercados de Europa.

Disponiendo de estos datos en formato CSV los interesados podrán realizar estudios para comparar las campañas de H&M en distintos países de Europa, comprobar si los mercados cercanos tienen ofertas similares y también si la renta de las personas en un mercado influye en el precio de los productos ofertados.

8. Licencia.

Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección.

Licencia elegida:

Database released under Open Database License, individual contents under Database Contents License. Permite compartir la base de datos, usar la base de datos para producir otros trabajos (siempre que el resultado del trabajo también tenga dicha licencia)

ODbL: Es una licencia que tiene como objetivo dar cobertura a juegos de datos. Esta licencia regula el uso público de los datos y los trabajos resultantes del uso de los datos deberán tener la misma licencia.

DCL: cubre el contenido de los datos.

9. Código.

El código que hemos escrito para obtener los datos está publicado en el siguiente enlace:

<https://github.com/mnyborg77/tcvd-practica1>

Para conseguir el dataset de los productos en ofertas de H&M en los distintos países de Europa hemos usado Python y Scrapy. Todas las librerías utilizadas en la elaboración del código se encuentran en el fichero requirements.txt en el repositorio de este proyecto.

El objetivo del código que compartimos es seleccionar las tiendas de H&M en Europa, encontrar sus ofertas en la sección dedicada a esto y obtener información de todos los productos que se encuentren en estas secciones.

El resultado del scraping se encuentra en el dataset que publicamos en Zenodo y también compartimos en el repositorio de github.

El scraping se realiza en varias etapas:

- En primer lugar seleccionamos todos los mercados dentro de Europa.
- En cada uno de los mercados encontramos las ofertas que existen. Las ofertas las identificamos con palabras clave en la url. La dificultad de esto consiste en que las páginas no usan exactamente las mismas palabras clave. Mientras que la mayoría usa “deals” para identificar las ofertas, algunos países usan “offers”, y la tienda de Alemania usa una palabra en alemán, “angebote”. Es posible que no hayamos conseguido toda la información que buscamos. En relación con esto, el scraper usa operadores lógicos para encontrar urls que contengan o bien “deal” o “offer” or “angebote”, y la coincidencia no necesita ser exacta.
- En los url de cada oferta podemos ver una lista de productos y extraemos una serie de datos de cada uno de ellos, sin llegar a la página de producto. Seguimos el link de paginación para navegar a la siguiente página que contiene más productos en oferta.

La dificultad de ejercicio consiste en identificar etiquetas y clases en el código fuente múltiples veces, identificar patrones de construcción de enlaces por parte de H&M.

10. Dataset.

El dataset está disponible en Zenodo <https://doi.org/10.5281/zenodo.7338342> y en repositorio GitHub <https://github.com/mnyborg77/tcvd-practica1>.

11. Vídeo.

Esta memoria también viene acompañada de un vídeo que presenta el proyecto de web scraping, incluyendo el interés en realizar este proyecto, el proceso seguido para conseguir los datos y el código y dataset resultantes.

https://drive.google.com/file/d/1_GQrx4-E7b5KSanlKjCoeGiIRQx4fkfX/view?usp=sharing

12. Contribuciones.

Por motivos personales y laborales el estudiante Morten Nyborg ha solicitado realizar esta práctica en solitario.

Contribuciones	Firma
Investigación previa	Morten Nyborg
Redacción de las respuestas	Morten Nyborg
Desarrollo del código	Morten Nyborg
Participación en el vídeo	Morten Nyborg