WEB SCRAPING

Práctica 1

Alberto Sarabia y Esther Gonzalez
Máster Ciencia de Datos

Tabla de contenido

1. Contexto

Hoy en día, la moda es un área muy atractiva para todo tipo de audiencia. Además, está disponible tanto para los más jóvenes como los más veteranos. En concreto, estudiaremos la moda actual en nuestro país.

La base de datos generada durante esta práctica estará alimentada por dos páginas de cadenas de moda distintas: Zara y Mango. Ambas están diseñadas para atraer a todos los públicos y que compren sus productos.

Las dos páginas, cada una acorde a la estratégica de marketing que siga, proporcionan toda la información relacionada con los productos que pone a la venta. Precios, materias de confección o maneras de cuidar el producto, entre otros, son algunos de los datos disponibles para recoger y crear el dataset.

2. Definición del título

Un título realmente descriptivo será el de Mango&ZaraFacts , Mango y Zara ya que son las compañías que estamos estudiando y Facts por que un hecho de que lo que nos están vendiendo y produciendo está alojado en su página web con todas las cosas que no vemos como la contaminación.

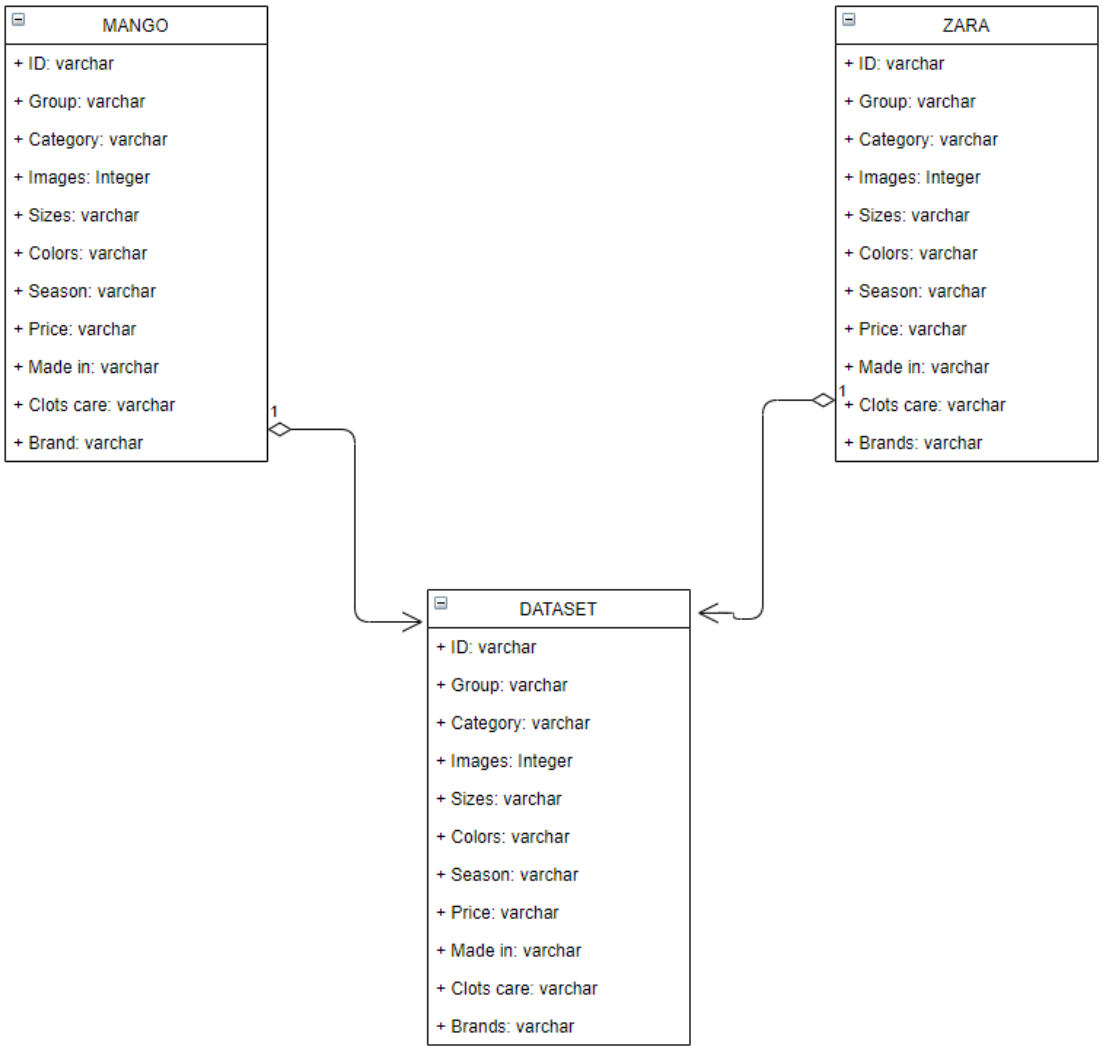
3. Descripción del dataset

Los datos que tenemos disponibles en la base de datos diseñada será un conjunto relacionado con la oferta de moda proporcionada a la población a nivel nacional. Se podría diferenciar los datos en un par de subconjuntos:

- **Información de la prenda**: características disponibles para la compra de la prenda: tallas, colores, precios, etc.
- **Cuidado y material de la prenda**: conjunto de información que detallará los cuidados requeridos y los materiales de confección, así como el lugar de fabricación de las prendas: lavado, secado, blanqueantes, etc.

4. Representación gráfica

La representación gráfica que identifica el dataset visualmente sería:



5. Contenido del dataset

Como se ha presentado anteriormente, en el tercer apartado, los datos recopilados se corresponden con las opciones de venta y los cuidados del producto disponible. En concreto, el conjunto de datos disponibles en el dataset será:

- **Referencia** (id): cadena de caracteres alfanumérico que utiliza cada página para identificar el producto.
- **Grupo** (group): conjunto del que forma parte el producto. Las posibilidades son: Hombre, Mujer, Niño o Niña.
- **Categoría** (category): tipo de prenda a la que pertenece. Las posibilidades son: Camisa, Camiseta, Pantalones, Vaqueros.
- **Imágenes** (images): cantidad de imágenes proporcionadas de la prenda.
- **Tallas** (size): Rango de tallas disponible. Algunos ejemplos serían: XS-XL, S-M-L, 34-46, 6 años (116 cm)-13-14 años (164 cm).
- **Variedad de colores** (colour): número de colores distintos disponibles del producto.

- **Temporada** (season): la moda se distribuye por distintas temporadas acorde al momento del año. Se podrán definir dos temporadas: primavera - verano u otoño - invierno.
- **Precio** (price): coste del producto
- **Fabricado** (made_in): Información del país dónde se ha creado el producto. Algunos ejemplos son: Portugal, Turquía o China.
- **Cuidado de la ropa** (clothes_care): el campo contiene información sobre los cuidados que debe tener el producto. La información se encontrará bajo el mismo campo, separada por guiones. En concreto, la información que podemos
    - **Fabricado**(made_in): información del origen de confección del producto.
    - **Temperatura máxima para el lavado** (temp_max_washing): temperatura máxima indicada para el cuidado de la prenda acorde al lavado.
    - **Centrifugado** (spin): indica si el tiempo de centrifugado debe ser nulo, corto o indiferente.
    - **Blanqueante** (whitening): si o no acorde a la posibilidad de utilizar productos blanqueantes, como la lejía, sobre el producto.
    - **Temperatura máxima del planchado** (temp_max_iron): temperatura máxima indicada para el cuidado de la prenda acorde al planchado.
    - **Limpieza en seco** (dry_wash): posibilidad de limpiar la prenda en seco.
    - **Secadora** (drying_machine): posibilidad de usar la secadora para la prenda.
- **Cadena (brand)**: al extraer información desde dos páginas distintas, se diferenciarán en dos marcas: Zara o Mango.

Además, para la marca Mango estará el código pre-preparado para la extracción de la información relacionada a ofertas y materiales. Está disponible la llamada para la extracción de dicha información, sólo se deberá ajustar el código para la generación del CSV para la creación de este campo.

Las temporadas en la moda es un punto muy relevante y que determina el estilo y variedad de productos que encontraremos. En concreto, cómo estamos evaluando las páginas españolas, la colección de otoño-invierno suele haber una gran variedad de abrigos, mientras que en la de primavera-verano hay más cantidad de camisetas.

Por ello, este script obtendrá la información acorde a la época en la que se ejecute el código: de enero a junio será temporada primavera-verano y de julio a diciembre la temporada otoño-invierno.

El dataset presentado, se corresponde a la colección de primavera-verano, ya que se ha extraído la información durante los meses de marzo a abril.

Para la recogida de la información, se han analizado las páginas de Zara (Zara, -) y Mango (Mango, -). Dentro de cada una de ellas, se ha validado, en una primera instancia, que no existiera ningún robot.txt.

Con esta información observada, se realiza una navegación profunda por la página para buscar la localización de los datos requeridos y, de este modo, poder obtenerlos finalmente en el dataset generado.

6. Agradecimientos

Agradecemos al desarrollo tecnológico de nuestro siglo que ha permitido un gran avance en la modernización de nuestra sociedad con ello la competitividad de nuestras empresas cada vez ha ido aumentando, teniendo un impacto directo en el medio ambiente y en la vida de las personas. Dos cadenas de ropa punteras de España en uno de los sectores que más dinero genera a nivel internacional son Mango y ZARA por lo tanto es interesante hacer un análisis de las marcas como los trabajos de fin de grados siguientes:

- Análisis de la estrategia de internalización (Bernal, 2018)
- Grupo Inditex (Contreras-Rivas, 2018)

Por ello, hemos decidido hacer una comparativa de las dos cadenas de tiendas extrayendo datos de sus respectivas páginas web utilizando el web scrapping, obteniendo una gran cantidad de información. Entre estos datos, se destaca, indirectamente, la conciencia y el posible impacto medio ambiental y social que generan estas empresas en el mundo, permitiendo a las personas incautas puedan fiscalizar las acciones de las empresas a las que compran sus ropas. En caso de no ser así, por lo menos, les permitirá saber dónde está dejando su huella en el paso por este mundo.

Algunos ejemplos concretos del impacto medio ambiental que dejan en nuestro planeta estas marcas de ropa están descritos en el siguiente artículo del periódico Economía Digital (Economía Dígital, 2019).

7. Inspiración

En la actualidad, el mundo de la moda se encuentra sujeta a diversas áreas. Desde un punto de vista laboral tendríamos todo un recorrido partiendo de la extracción de materias primeras hasta la venta del propio producto final. Algunos pasos intermedios destacables sería todo el sector textil, diseño de moda, presentaciones (pasarelas de modelaje), etc.

Paralelamente, la moda también es un punto destacable ante el movimiento ecológico y de sostenibilidad del planeta. Como se ha comentado, se requerirá de la extracción de materias primeras para poder realizar el producto final y, dependiendo del tipo de materia, puede ser dañino para el planeta.

La última área para comentar durante este documento, aunque se abarcan muchos más, será la competitividad a la hora de vender los productos finales confeccionados. En los países desarrollados, la presencia y los 'looks' de las personas son temáticas muy comentadas y de gran relevancia. Por ello, existen múltiples marcas que compiten por ser los escogidos por la mayoría de la población.

Teniendo en cuenta el conjunto de datos recopilados junto a las áreas comentadas previamente, este dataset podría darnos respuesta a necesidades cómo:

- Comprar precios entre grupos distintos.
- Compara los cuidados necesarios para la ropa entre grupos distintos.
- Validar los distintos orígenes de donde se genera la ropa.

En los análisis mencionados anteriormente observamos cómo se estudian las cadenas de Mango y Zara desde un punto de vista de negocio. En el artículo de Economía Digital (Economía Dígital, 2019) hemos observado cuánto contaminan algunas de las grandes empresas de ropa de este planeta que, desde el punto de vista de la mayoría de las personas, es excesivo y gustaría buscar alternativas y soluciones.

La principal diferencia entre los estudios anteriores y nuestro dataset se origina en el objetivo final para la utilización de los datos. Desde el punto de vista del artículo, se centran en el modelo de negocio. En cambio, este dataset ofrece la oportunidad de dar respuesta más allá de este modelo, mediante la posibilidad de hacer un estudio del mercado textil o si los cuidados indicados favorecen a la sostenibilidad del planeta.

8. Licencias

Las múltiples capacidades de estudio que puede aportar este dataset favorece a que se pueda utilizar en muchas ramas distintas. Nosotros queremos aportar en todo tipo de investigación y nos gustaría que el conocimiento extraido a partir de esta fuera compartido entre todos.

No obstante, dada nuestra contribución, de tipo altruista, al mundo de la investigación del sector moda, nos gustaría que cualquier utilización de este dataset fuera en beneficio al estudio y no a nivel económico, apostando así por el open data.

Por ello, consideramos que la licencia ideal para ello sería la 'Released Under CC BY-NC-SA 4.0 License'.

9. Código generación dataset

El código se encontrará en el repositorio GIT (González, 2021) creado.

10. Dataset

Zenodo es un repositorio de acceso abierto de propósito general desarrollado bajo el programa europeo OpenAIRE y operado por CERN con la finalidad de compartir datos para realizar investigaciones impulsando el OpenData en el espacio europeo abogando por las licencias libres en su mayoría, pero permitiendo otro tipo de licencia en espacios privados de usuarios. además de estar integrado con github dándole una ventaja para publicar aparte de datasets otro tipo de información importante como gráficos, código etc.

El DOI (Sarabia, 2021)obtenido es:

```
[![DOI](https://zenodo.org/badge/DOI/10.5281/zenodo.4681990.svg)
](https://doi.org/10.5281/zenodo.4681990)
```

11. Contribuciones y firmas

A continuación, se presenta las contribuciones al trabajo:

Contribuciones	Firma
Investigación Previa	ASS, EGB
Redacción de las respuestas	ASS, EGB
Desarrollo código	ASS, EGB

12. Referencias

Bernal, C. A. (01 de 04 de 2018). *ANÁLISIS DE LA ESTRATEGIA DE INTERNACIONALIZACIÓN*. Obtenido de ANÁLISIS DE LA ESTRATEGIA DE INTERNACIONALIZACIÓN: http://diposit.ub.edu/dspace/bitstream/2445/123716/1/TFM-MOI_Almagro_2018.pdf

Contreras-Rivas, J. (1 de 11 de 2018). *GRUPO INDITEX: PLAN DE CRECIMIENTO, ANÁLISIS Y RECOMENDACIONES 2018-2022*. Obtenido de GRUPO INDITEX: PLAN DE CRECIMIENTO, ANÁLISIS Y RECOMENDACIONES 2018-2022: https://pirhua.udep.edu.pe/bitstream/handle/11042/3987/MDE_1870.pdf?sequence=2&isAllowed=y

Economía Dígital. (28 de 10 de 2019). *La moda efímera de Zara y H&M tiene un alto impacto medioambiental*. Obtenido de La moda efímera de Zara y H&M tiene un alto impacto medioambiental: https://www.economiadigital.es/empresas/la-moda-efimera-de-zara-y-h-m-tiene-un-alto-impacto-medioambiental_20005304_102.html

González, A. S. (12 de 04 de 2021). *Git-repository*. Obtenido de Git-repository: https://github.com/ASarabiaSuarez/WebScrapingUOC

Mango. (- de - de -). *Mango Website*. Obtenido de Mango Website: https://shop.mango.es/shop/

Sarabia, E. G. (12 de 04 de 2021). *Dataset Zara&MangoFacts*. Obtenido de Dataset Zara&MangoFacts: https://zenodo.org/record/4681990#.YHSsq_XPyUk

Zara. (- de - de -). *Zara Website*. Obtenido de Zara Website: www.zara.com/es/