

Practica 1 de Tipologia y Ciclo de Vida de los Datos

Indice

- - 1. Contexto
 - 2. Título Dataset
 - 3. Descripción
 - 4. Representación gráfica
 - 5. Contenido
 - Campos dataset
 - Periodo temporal
 - Cómo se ha recogido
 - 6. Agradecimientos
 - 7. Inspiración
 - 8. Licencia
 - 9. Código y Dataset
 - 10. Tabla de Firmas
 - 11. Recursos

Miembros del equipo

Esta practica ha sido realizada por Oscar Pedreño y Daniel Villalobos

1. Contexto

Trabajamos en el departamento de *Data* para unos grandes almacenes dedicados exclusivamnete a la moda de hombre. Se nos ha pedido que realicemos un estudio de los precios de la competencia para conocer nuestra situación de mercado respecto a nuestros competidores. Nos centraremos en **El corte Inglés** ya que por volumen de venta nos manejamos sobre las mismas cuotas de mercado.

La página web de donde obtendremos la información principalmente es:

<https://www.elcorteingles.es/moda-hombre/ropa/>

2. Título Dataset

El dataset final que generaremos a partir de los datos recolectados será:

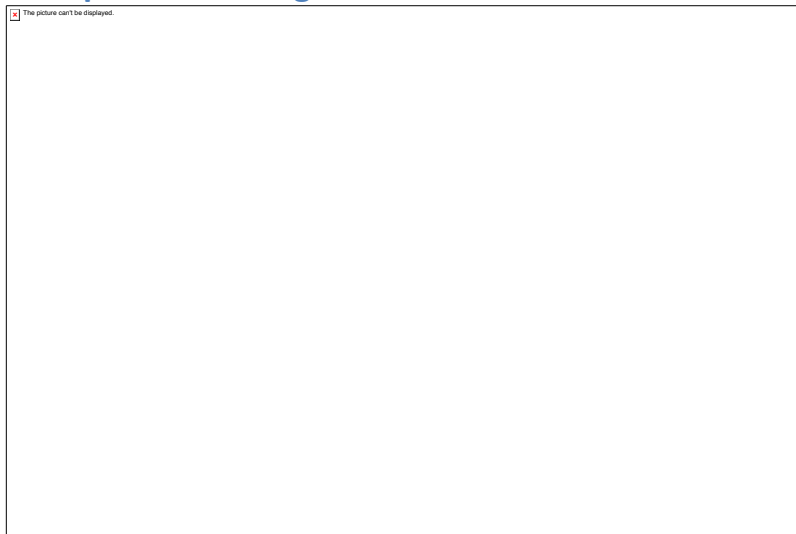
precios_ei_ropa

3. Descripción

Los datos que nos interesa obtener principalmente son, el precio de la prenda, la marca, el nombre de la prenda, la categoría, los colores de la prenda, sus características. Los campos que extraeremos son:

- Fecha
- Id Producto
- Prenda
- Marca
- Categoría
- Precio
- Precio Descuento
- Descuento
- Porcentaje Descuento

4. Representación gráfica



Se debería mostrar la imagen

5. Contenido

Campos dataset

Los campos contenidos en el dataset son:

- Fecha
- Id Producto
- Prenda
- Marca
- Categoría
- Precio

- Precio Descuento
- Descuento
- Porcentaje Descuento

Periodo temporal

Nuestro proyecto, al tratarse de la creación de un catálogo de productos de la página web de El Corte Inglés, extraemos los datos que se encuentran en cada momento en su página web. Uno de los evolutivos que queremos crear en nuestro Bot Scraper es seguir extrayendo los datos de las diferentes páginas y tener un catálogo cada vez mayor.

Cómo se ha recogido

Hemos usado el framework Scrapy para extraer la información de eci moda hombre. Hemos recogido los datos en formato csv para posteriormente extraer información.

6. Agradecimientos

La página web desde donde extraemos los datos es **El Corte Inglés**, unos grandes almacenes españoles reconocidos mundialmente y con una selección de productos muy amplia. Dado que nuestro proyecto se va a centrar en investigar las diferentes propuestas del sector moda, seleccionaremos esa clase de productos.

Otros trabajos que nos han inspirado serían:

<https://github.com/David-Carrasco/Scrapy-Idealista>

<https://github.com/GoTrained/Scrapy-Craigslist>

Se ha utilizado el lenguaje de programación *Python* y la herramienta de *Web Scraping* para extraer la información alojada en las páginas HTML ha sido el framework Scrapy y sus librerías internas.

7. Inspiración

La motivación principal para realizar este proyecto es responder varias preguntas:

- ¿Qué diferencia de precios existe entre la competencia y nosotros?
- ¿Qué marcas trabajan con la competencia?
- ¿Que tipos de productos trabajan la competencia?
- ¿Qué tipos de descuentos hace la competencia?
- ¿Sobre qué productos hace descuentos la competencia?
- ¿En qué épocas del año centran los descuentos la competencia?

Respondiendo estas preguntas seremos capaces de poder entender como funciona la estrategia de negocio de la competencia y trazar una solución creativa y reactiva a los movimientos de la misma.

8. Licencia

La licencia escogida para este trabajo es la licencia CC BY-SA 4.0. Los motivos principales que nos han llevado a elegir esta licencia son:

- Nos permite reconocer el trabajo realizado previamente y las aportaciones que se han realizado respecto el trabajo original.
- Se permite un uso comercial por lo que podría ser utilizado por terceros, pero siempre reconociendo el trabajo original del autor.
- Las contribuciones se deben publicar sobre la misma licencia por lo que se respetan los dos puntos anteriores. De esta manera, se sigue reconociendo al autor original.

9. Código y Dataset

El código y el dataset estan alojados en este mismo repositorio en la carpeta de *scraping*

10. Tabla de Firmas

Contribuciones	Firma
Investigación Previa	OP; DV
Redacción de las Respuesta	OP; DV
Desarrollo de código	OP; DV

11. Recursos

Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data. Subirats, L., Calvo, M. (2019). Introducción al ciclo de vida de los datos. Editorial UOC.

<https://docs.scrapy.org/en/latest/topics/selectors.html>

<https://www.accordbox.com/blog/scrapy-tutorial-7-how-use-xpath-scrapy/>

<https://docs.scrapy.org/en/latest/topics/commands.html>

<https://docs.scrapy.org/en/latest/intro/tutorial.html>