Práctica 1: Web scraping
Asignatura: Tipología y ciclo de vida de los datos
Universidad: UOC
Alumnos:

- Luís Alejandro León Corcuera (lleonco@uoc.edu)
- Sergi Boadas Vilagran (sboadas@uoc.edu)

Fecha de entrega: 09/11/2020
Enlace repositorio github: https://github.com/Serk-KR/WebScraping

Pràctica 1 - Web Scraping

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

   La información recolectada es de un conjunto de productos, concretamente de la marca Mustang. El conjunto de datos nos dará información detallada del producto como por ejemplo nombre, precio, etc.

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

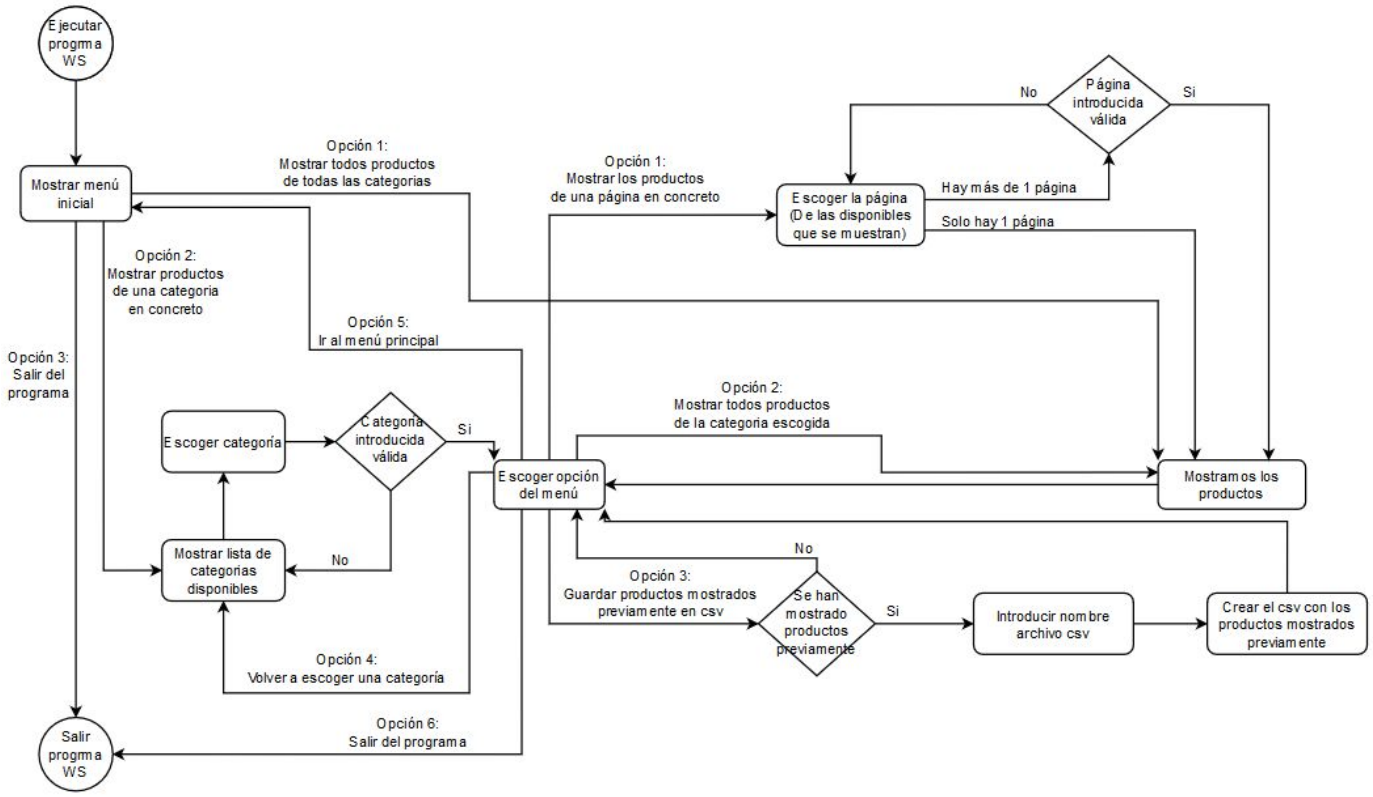
   Productos a la venta online de la marca Mustang

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).**

   Antes de comentar la descripción del dataset se tiene que comentar que dicha información que se extrae va en función de las decisiones del usuario que ejecute el programa. Al ejecutar el programa el usuario introducirá por teclado las opciones que más le interesen según el menú del programa.

   El resultado final del dataset será un conjunto de productos. Estos productos están distribuidos en diferentes categorías: hombre, mujer y niño. Dentro de cada una de éstas puede haber productos de calzado o accesorios. Concretamente el calzado también se distribuye en diferentes tipos, por ejemplo: zapatos, botines, botas, etc. También se da la posibilidad de poder generar un dataset de los productos más vendidos (hombre, mujer o niño) o de una página en concreto (preguntada al usuario).

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



En el diagrama anterior, se puede ver un resumen de las funcionalidades del programa y cómo acceder a ellas

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

   Nombre: nombre del producto que está a la venta
   Categoría: categoría del producto (Hombre, Mujer,…)
   Subcategoría: subcategoría del producto (Zapatos, Bota,...)
   Color: color del producto
   Precio: precio a la venta del producto, en euros
   Venta: "Si" si el producto está disponible para comprar, "No" si no lo está
   Tallas: cadena caracteres que indica las tallas disponibles del producto ("38-40-42...")
   Referencia: codificación única por cada uno de los productos
   Otros: cadena de caracteres, delimitado por "$" donde se da información extra del producto (Material principal, Altura, Material de la suela...). Si todos los productos compartieran las mismas características se hubiesen creado nuevos campos.
   La marca *Mustang* es una marca líder, global y reconocida. Fundada en 1967, durante todo este tiempo se ha ido adaptando a los gustos del consumidor. Información extraída de: https://www.mustang.es/es/info/la_marca/

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

   Los datos recolectados han sido desde la visualización de ventas online de productos de la página online https://www.mustang.es. Antes de analizar los datos de la página se ha inspeccionado el fichero robots.txt, que se encuentra en https://www.mustang.es/robots.txt.
   En las primeras líneas del ficheros se encuentran las siguientes instrucciones:

   *User-agent: **
   *Disallow: /*area_cliente**
   *Disallow: /*checkout**

   El significado de estas instrucciones es el siguiente: se permite inspeccionar la página a todos los usuarios a excepción de las directorios de la área de cliente y la de compra. Es por ello que tenemos permiso para analizar la página.

   Para el resultado final del dataset se han utilizado técnicas de WebScraping con el lenguaje de programación Python.

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

   En el hipotético caso de disponer de un negocio de ventas de ropa, este conjunto de datos nos da la posibilidad de obtener la información más relevante de la página web "https://www.mustang.es/es/". En esta página se pueden encontrar una gran cantidad y variedad de productos de ropa y accesorios.
   El programa que hemos desarrollado nos permite hacer las siguientes acciones/funcionalidades:
   - Mostrar todos los productos de todas las categorías disponibles de la página web
   - Mostrar todos los productos de una categoría en concreto (elegida por el usuario)
   - Mostrar los productos de una categoría en concreto y de una única página (elegida por el usuario en caso que esa categoría tenga más de una página de productos)
   - Crear y guardar información de los productos mostrados previamente a un archivo csv (el cual su nombre, será introducido por el usuario)
   - Salir del programa

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

   La licencia que se ha escogido para este dataset ha sido **Released Under CC BY-SA 4.0 License** ya que con esta licencia nos permite compartir, mediante copia y redistribución del conjunto de datos en cualquier tipo de formato. Además se puede adaptar, modificar y adaptar incluso con medios comerciales.

   Con estos permisos se pueden beneficiar aquellas empresas que con fines comerciales pretendan incentivar a compradores potenciales, analizando los precios, características, etc… manteniendo el reconocimiento inicial del trabajo realizado por los autor/es de los datos.

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

   El código generado para la creación del dataset se encuentra disponible en el repositorio siguiente: **https://github.com/Serk-KR/WebScraping**

   Comentarios que nos gustaría destacar:

   - La opción del link de "Más Vendidos" de la página web "https://mustang.es/es/" que escogimos, resulta que tiene un error y ese enlace del nav_menu, no tiene URL. Así pues nos encontramos que si ejecutamos nuestro programa con esa opción no funciona (no por nuestro programa), sino porque se intenta acceder a un url (inexistente por error de la web).

   - Quitado último elemento del array de categorías del menú, debido a que la última no contenía ninguna url con productos. Es decir, no era de nuestro interés

   - Al inicio de la realización de esta práctica, la página web que escogimos tenía las siguientes categorías: "Mujer", "Hombre", "Niño", "Más vendidos". Cuando estábamos terminando los últimos retoques de la práctica para entregarla, nos dimos cuenta que la empresa que gestiona la página web, ha añadido una nueva categoría de nombre "New in". Este hecho nos ha permitido verificar el buen funcionamiento de la extracción de las categorías por parte del programa, ya que no hemos tenido que tocar nada de código y funcionaba todo correctamente.

   - Después de generar el dataset.csv, hicimos un análisis visual del fichero y observamos que habían campos incorrectos, por ejemplo: "Depotiva" en vez de "Deportivas", etc. Es por ello que a continuación de generar el fichero decidimos hacer una limpieza de estos valores incorrectos.

10. **Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.**

    DOI → **10.5281/zenodo.4249647** - **https://doi.org/10.5281/zenodo.4249647**
    Descripción → Subido un csv (dataset) en la plataforma de Zenodo, el cual contiene información de todos los productos que están a la venta en la página web de la marca española "Mustang". Este csv, permitirá a otros usuarios que se lo puedan descargar y puedan interaccionar con él.

Contribuciones	Signa
Investigación previa	L, S
Redacción de las respuestas	L, S
Desarrollo código	L, S