PRAC1

Alumno: Carlos Garabatos Fernández

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. **Contexto.**

   En esta práctica se aplican técnicas de web scraping mediante el lenguaje de programación Python para extraer así datos de la web y generar un dataset. En este caso se extraen los precios de diferentes productos textiles de la página web https://latiendadevalentina.com/.

   Aunque en mi caso en particular responde a una finalidad meramente educativa, podríamos imaginar que nuestro objetivo fuese el análisis de nuestra competencia. Por lo que el objeto de esta actividad sería la creación de un dataset a partir de los datos contenidos de una web de venta de textil.

   Estudio inicial:

   • Ver archivo robots.txt:
   User-agent: *

   Tecnología usada:
   Web: Prestashop

   Domain Name: LATIENDADEVALENTINA.COM
   Registry Domain ID: 1958360246_DOMAIN_COM-VRSN

   Tiempo de carga
   1.61 segundo(s)

   Número de registros de texto previstos: 435

2. **Definir un título para el dataset**.

   valentinaTextil

3. **Descripción del dataset.**

   Productos textiles de Valentina: Se extraen los precios de diferentes productos textiles de la página web https://latiendadevalentina.com/.

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

5. **Contenido**.

   Se extraen solo los datos de las secciones de zapatos y moda, porque solo nos interesa el producto textil, y además muchos productos se repiten en otras secciones.

   Numero de registros: 435. Se regulan los tiempos, para no sobrecargar a 30 minutos de extracción total.

   Los datos extraídos son:

   - Categoría: Las categorías de producto han variado durante el proceso de desarrollo. El motivo de esto es que existen unas primarias, como moda y zapatos, y otras que se generan a partir de los productos de estas como homewear u outlet. Por ese motivo solamente se extraen las categorías textiles moda y zapato.
   - Subcategoría. Subcategoría del producto.
   - Nombre: Nombre del artículo.
   - Color: Color o colores del producto.
   - Url: Dirección web del producto, para poder consultar un producto específico si fuese necesario. Y, además, ayudara en la fase de corrección.
   - Talla: Rango de tallas.

- Precio: El precio del detalle de producto también ha evolucionado durante el desarrollo, por lo que actualmente se extrae del producto reducido y no del detallado.
- Descripción: Se extrae del producto detalle, y permite analizar y corregir el enfoque final del producto.

6. **Agradecimientos.**

   No se han incluido las fotos para respetar los derechos de autor declarados en la página.

   Vulneración de DERECHOS DE AUTOR:
   En el caso de utilización o uso de CUALQUIERA de las fotos o imágenes pertenecientes a Valentina, se procederá a la denuncia inmediata del autor del hurto por vulneración de derechos de autor.

   Agradezco que la tienda de Valentina no se haya opuesto al análisis educativo de sus datos.

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

   Es un supuesto comercial podría tener valor estratégico como fuente de análisis de la competencia. Esto nos permitirá obtener el conocimiento de que productos, precios, colores y tallas trabaja la competencia. Y no permitirá elaborar estrategias de bajadas de precios, y además reforzar nuestro stock de algunos de los productos más vendidos, para suplir las carencias de stock de la competencia en tallas o colores.

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

○Released Under CC0: Public Domain License: puedan ser redistribuidos y manipulados de manera completamente libre y sin restricciones, ya sea comercial o no comercialmente.

○Released Under CC BY-NC-SA 4.0 License. Atribución-No Comercial-Compartir Igual

○Released Under CC BY-SA 4.0 License. Atribución-Compartir Igual

○ Database released under Open Database License, individual contents under Database Contents License. Permite compartir sus datos libremente sin preocuparse por los problemas relacionados con los derechos de autor o la propiedad

○ Other (specified above)

○ Unknown License

En este supuesto los datos no son de mi propiedad, la creación de la licencia debe remitirse al propietario de los datos. Por ese motivo escogería:

○Released Under CC BY-NC-SA 4.0 License.

Las razones son:

BY: El beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y **cite la obra** de la forma especificada por el autor o el licenciante.
NC: El beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas para fines **no comerciales**.
SA: El beneficiario de la licencia tiene el derecho de distribuir obras derivadas bajo una licencia idéntica a la licencia que **regula la obra original.**

9. **Código.**

   Se adjunta el código Python: **cgarabatos.py**

   Enlace al repositorio:
   https://github.com/cgarabato/prac1

10. **Dataset.**

    Se presentar el dataset en formato CSV. Los campos se han formateado con comillas separados por comas. El fichero se llama**: valentinaTextil.csv**

11. **Problemáticas o inconvenientes.**

    Durante el desarrollo del programa la web ha sufrido diferentes modificaciones en su estructura y en sus condiciones de uso.

    Como por ejemplo el diseño del campo de precios en la descripción de los productos y la sustitución de la sección homewear. Esta última, posterior a la entrega parcial del código.