COMPONENTES DEL GRUPO

Diana Cózar Salas

PREGUNTAS Y RESPUESTAS

1. Título del dataset. Poned un título que sea descriptivo.

Books Store Data

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.

Books store catalog metrics and prices detail

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente

4. Contexto. ¿Cuál es la materia del conjunto de datos?

La materia del conjunto de datos es el catálogo de una librería online. Se han realizado técnicas de web scraping para obtener el catálogo de libros a la venta y métricas significativas como el número de libros por categoría, precio mínimo/máximo y medio de cada categoría; además de un listado detallado de libros, sus precios y la categoría a la que pertenecen.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

Se han generado dos ficheros de texto en formato CSV con la información extraída de la web:

- **catalogData:**

  Extraído mediante web scraping a través de navegar por la página principal y los links del catálogo. Así se obtiene un listado de categorías de libros que se venden en la web y sus datos significativos. Los campos del fichero son los siguientes:

  1. **category:** categoría del catálogo (Travel, Mistery, etc…)
  2. **url:** url de la página principal de esta categoría
  3. **n_elements:** número de libros que se venden en esta categoría (obtenido navegando a través de las páginas principales de cada categoría y almacenando el número de resultados)
  4. **min_price:** precio del libro más económico de esta categoría (obtenido agregando los datos del fichero de detalle)
  5. **max_price:** precio del libro más caro de esta categoría (obtenido agregando los datos del fichero de detalle)
  6. **mean_price:** precio medio de los libros de esta categoría (obtenido agregando los datos del fichero de detalle)

- **catalogPrices:**

  Extraído mediante web scraping a partir de los links obtenidos en el catálogo. Una vez situados en una categoría, navegamos a través de todas las páginas de esta categoría para ir almacenando los libros que muestra cada página (máximo 20 libros por página) y sus precios.
  Los campos del fichero son los siguientes:

  1. **book:** título del libro
  2. **price:** precio del libro
  3. **category:** categoría del libro

Los datos se extraen y se almacenan a tiempo real de la web, así que el período de tiempo de los datos es el del momento de la extracción.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Este proyecto de web scraping se ha realizado sobre una web específica para ello, con datos no reales y por tanto públicos.
La web utilizada y otras webs con el mismo propósito forman parte del proyecto educativo Web Scraping Sandbox (http://toscrape.com/)

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Este conjunto de datos es interesante dado que permite recopilar información acerca de los libros que contiene una web de ventas, sus precios y su catálogo de venta (clasificación de los libros según su tipología).

Estos datos si fueran de una web real de venta de libros de la competencia podrían resolver preguntas como:

- ¿A qué precio vende el mismo libro un competidor?
- ¿Qué oferta de libros tiene? (número de libros por categoría)
- ¿Cuál es el precio medio/mínimo/máximo de sus libros en cada categoría? (a qué mercado va dedicado)

Este tipo de información es muy útil cuando queremos analizar el mercado y así equiparar nuestros productos con la competencia para ser más competitivos a nivel de ventas o lanzar al mercado productos innovadores (u ofertas innovadoras).

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

○ Released Under CC0: Public Domain License
○ Released Under CC BY-NC-SA 4.0 License
○ Released Under CC BY-SA 4.0 License
○ Database released under Open Database License, individual contents under Database Contents License
○ Other (specified above)
○ Unknown License

Al ser datos totalmente públicos ya que son parte de un proyecto educativo, elegiríamos la licencia CC0 (ningún derecho reservado).
En caso que este mismo código se utilizase para la extracción de datos en una web librería real, elegiríamos la licencia CC BY-NC-SA 4.0 ya que podríamos extraer los datos pero para fines no comerciales.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

Incluido en el repositorio de GitHub

10. Dataset: Dataset en formato CSV

Incluido en el repositorio de GitHub