TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRÁCTICA 1: WEB SCRAPING

ALEJANDRO MARTÍNEZ OTAL| WEB SCRAPING | 11/11/2019

Índice

Contexto

Con este proyecto, a partir de un portal de alquiler de pisos se puede extraer información sobre el precio de alquiler de habitaciones junto otros parámetros como número de habitaciones, superficie útil, etc.

Título dataset

El título elegido para el dataset es : ''Métricas para el alquiler de habitaciones por provincia(s)''.

Descripción dataset

El dataset en concreto para este ejemplo contiene todas las habitaciones disponibles de alquilar en provincias [1] para la fecha 11/11/2019. Este contiene algunos campos que se explican más adelante en el documento como precio, superficie, número de baños....

Representación gráfica

Imagen 1 El alquiler de piso siempre ha sido un tema polémico que preocupa, especialmente a jóvenes.

Contenido

El dataset contiene para cada habitación en alquiler la siguiente información:

- Id: Identificador único del piso proporcionado por la propia web.
- Location: Localidad dentro de la provincia elegida dónde se encuentra el piso.
- Precio: Cantidad en € mensual para el alquiler de la habitación.
- Descripción: Se almacena una breve descripción.
- Superficieutil: Número de m^2 de la habitación.
- Numbanos: Número de baños con los que cuenta el inmueble.
- Estadoconservación: Estado en el que se encuentra el inmueble.
- Gastosincluidosalquiler: Si el precio de la columna 'Precio' incluye los gastos.
- Numhabitaciones: Número de habitaciones con los que cuenta el inmueble.
- Número de inquilinos: Número de personas viviendo en el piso.
- Edad mínima: Edad mínima para entrar.
- Género: Género que debe cumplir la persona interesada en entrar.

Motivación

El precio de las viviendas siempre ha sido un tema polémico durante los últimos años desde la crisis mobiliaria.

Obviamente, el precio del alquiler de las habitaciones está ligado al precio de las viviendas. La aparición de nuevas formas de alquiler proporcionadas por plataformas como Airbnb han aparecido como alternativas al alquiler tradicional también ha influenciado en su precio.

Los resultados de la extracción se pueden utilizar directamente para encontrar una habitación según ciertas preferencias o utilizar como un juego de datos para estudiar el mercado.

Agradecimientos

Todos los datos han sido extraídos de la página web www.pisos.com mediante una herramienta de *Web Scraping* programada en Python.

Agradecer al portal, que aparentemente, después de comprobar el fichero **robots.txt** no decide bloquear el *user-agent* que utiliza la librería de Python *request* por defecto.

Licencia

Para mi dataset he decidido optar por la licencia '' CCo: Public Domain License" [2] .

Personalmente, soy un firme defensor de los proyectos open-source, dónde el esfuerzo conjunto de la comunidad permite desarrollar grandes proyectos.

A pesar de que mi código me ha llevado un dedicación y esfuerzo, toda la información que he adquirido ya es pública, simplemente la entrego en un formato estructurado en filas y columnas.

El hecho de haber elegido esta licencia implica:

- La persona que decide usar esta licencia renuncia a todos sus derechos de autor.
- Cualquier persona podría copiar, distribuir y realizar trabajos (incluso con fines comerciales) sin pedir permiso.
- La persona que ha licenciado el producto con esta licencia no ofrece ningún tipo de garantías sobre los resultados que este otorgue o posibles implicaciones legales sobre el mal uso de este.
- El creador de este trabajo no tiene porqué ser mencionado al usar el código.

Código

El código utilizado se puede encontrar en https://github.com/AlejandroUPC/pisos_scrapper.

Algunos comentarios:

- Todas las funciones están complementadas con su Docstring dónde se explica brevemente su objetivo/función.
- Se recomiendo leer el fichero **README.md** que es encuentra en la capeta raíz del repositorio.
- Con tal de evitar posibles restricciones por parte del propietario de la web en la configuración (**configuratoin/main_configuration.py**) se pueden activar la opción de añadir un *delay* entre *requests* y los segundos, se pide por favor que se utilice con valores lógicos.
- Al final, en la carpeta *output_files* y si la opción está habilitada, genera algunos gráficos orientativos para sacar tempranas conclusiones.

Dataset

El dataset se puede encontrar en https://github.com/AlejandroUPC/pisos_scrapper bajo la carpeta **output_files/** nombre **Global_d%-%m-%Y.csv**.

Bibliografía

[1] Once, "Lista de Provincias con Prefijo," [Online]. Available: ftp://ftp.once.es/pub/utt/bibliotecnia/Miscelanea/Mapas/Mapa_auton%F3mico_de_Espa%F1a.docx.

[2] C. Commons, "CCo 1.0 Universal," Creative Commons, [Online]. Available: https://creativecommons.org/publicdomain/zero/1.0/.

[3] Laia Subirats Maté and Mireia Calvo González, Web scraping, UOC, 2019.

Contribuciones

Contribuciones	Firma
Investigación previa	Alejandro Martínez Otal
Redacción de las respuestas	Alejandro Martínez Otal
Desarrollo código	Alejandro Martínez Otal