Práctica 1 – Web Scraping

Adrián González González

https://github.com/adrianglez77/Housing-price.git

1.- Contexto

En la presente práctica se ha recolectado información de casas en venta en la provincia de Madrid, España. La web utilizada para realizar Web Scraping ha sido: *"www.tucasa.com",* la cual presenta información relevante sobre la oferta de viviendas en todo el país.

El sitio web escogido, tiene las variables necesarias para poder realizar la práctica que se solicita, ya que contiene mucha información que sería interesante analizar, y con una fecha de actualización muy constante, lo que permite poder ampliar en un futuro el estudio.

La web *"www.tucasa.com"* es una de las principales webs de oferta y demanda de inmuebles en España, con lo que facilita el poder cumplir con los objetivos de esta práctica

2.- Título

El título escogido para el dataset es: **"Housing prices".**

3.- Descripción del dataset

El conjunto de datos que se ha extraído pertenece a las viviendas en venta en las distintas áreas de Madrid. En este dataset, se encuentran cuatro variables (área geográfica, metros, precio por metro, y fecha de actualización), y se han obtenido datos de 200 viviendas distintas.

La idea del estudio del dataset es poder comparar más adelante el precio de las viviendas según el área geográfica de Madrid en la que se encuentre, y ver como varía el precio/metro. También se ha incluido la fecha de actualización, para que el dataset esté con una actualización acorde a la realidad y no con datos históricos.

4.- Representación gráfica

La imagen que representa al dataset es la siguiente, ya que visualmente se centra en una casa, y un posible acuerdo con las llaves (venta de casas).

El conjunto de datos creado tiene un formato CSV, un archivo de texto en el cual los caracteres están separados por comas, haciendo una especie de tabla en filas y columnas. Su visualización con el programa “RStudio” es la siguiente:

	Área Geográfica	Metros	Precio-Metro	Ultima Actualizacion
1	Rinconada (Alcalá de Henares)	68 m2	1.250 €/m2	hoy 14:06
2	Santa Eugenia (Madrid)	86 m2	2.302 €/m2	hoy 14:04
3	Ensanche (Alcalá de Henares)	275 m2	1.600 €/m2	hoy 13:56
4	Ensanche (Alcalá de Henares)	259 m2	1.486 €/m2	hoy 13:54
5	Pueblo Nuevo (Madrid)	85 m2	2.459 €/m2	hoy 13:52
6	Fuente del Berro (Madrid)	47 m2	6.702 €/m2	hoy 13:45
7	Rivas-VaciaMadrid, Covibar	135 m2	1.703 €/m2	hoy 13:42
8	Rivas-VaciaMadrid, Paseo de las Provincias	102 m2	2.696 €/m2	hoy 13:40
9	Pavones (Madrid)	200 m2	3.400 €/m2	hoy 13:37
10	Castillejos (Madrid)	69 m2	2.159 €/m2	hoy 13:30
11	Pinto, Parque Europa	89 m2	2.191 €/m2	hoy 13:26
12	Sevilla la Nueva	277 m2	1.534 €/m2	hoy 13:22
13	Sevilla la Nueva	516 m2	1.328 €/m2	hoy 13:21
14	Nuevo Baztán, Nuevo Baztán	107 m2	1.495 €/m2	hoy 13:20
15	Titulcia	201 m2	871 €/m2	hoy 13:20
16	Nuevo Baztán, Nuevo Baztán	190 m2	1.568 €/m2	hoy 13:19
17	Centro (Casco Antiguo) (Alcorcón), Valderas-Los Castillos-Pa...	75 m2	2.399 €/m2	hoy 13:14
18	Centro (Leganés)	455 m2	1.077 €/m2	hoy 13:03
19	Cerceda	80 m2	2.125 €/m2	hoy 13:01
20	Boadilla del Monte, Valdepastores-Pino Centinela-Las Encinas	140 m2	4.248 €/m2	hoy 13:00
21	Salamanca (Madrid)	141 m2	3.794 €/m2	hoy 12:58

5.- Contenido

Como hemos dicho anteriormente, el dataset tiene 4 campos:

- Área Geográfica: tipo de campo de Texto. Contiene la información relativa a la zona geográfica dentro de Madrid en la que se encuentra el inmueble.
- Metros: tipo de campo de Texto. Contiene la información relativa a los metros cuadrados que posee el inmueble.
- Precio-metro: tipo de campo de Texto. Contiene la información relativa a la relación existente entre el precio y el metro cuadrado.
- Última Actualización: tipo de campo de Texto. Contiene la información relativa a la última actualización del inmueble en la web.

6.- Propietario

El propietario del conjunto de datos es la web www.tucasa.com, que es un portal donde cualquier persona puede publicar anuncios de venta de inmuebles. Este portal se encarga de publicar los anuncios, ofrecer servicios para multiplicar las posibilidades de ventas y ofrece la ayuda para contactar con los interesados o propietarios de inmuebles.

No se ha detectado ningún análisis anterior en esta web de venta de inmuebles. Sí se han encontrado análisis a distintas webs de venta de inmuebles, diferente a la estudiada.

Los principios éticos y legales en el contexto del proyecto que se han seguido han sido:

7.- Inspiración

Este conjunto de datos es interesante de analizar ya que los precios de los inmuebles no paran de crecer en las grandes ciudades como Madrid. En los últimos años se ha producido unos incrementos de los precios de las viviendas, y la demanda también sigue creciendo.

Desde el punto de vista tecnológico, también siguen creciendo los portales webs que se encargan de hacer de “intermediarios” entre un comprador y vendedor.

Por lo tanto, me resulta curioso saber el precio medio, máximos, mínimos, según el área geográfica en la que se sitúe el inmueble dentro de la provincia de Madrid.

También, estaría muy interesante ver la evolución de éstos, y ver el incremento del precio/metro cuadrado en los últimos meses. Ésta sería una tarea para el futuro, ya que actualmente estamos obteniendo los últimos 200 inmuebles publicados, con lo cual, la fecha de publicación es muy reciente y no podremos tener un histórico de los datos.

8.- Licencia

La licencia que se considera adecuada para el dataset resultante es: **Released Under CC BY-NC-SA 4.0 License**

La elección de esta licencia ha sido por diferentes motivos:

- Las licencias CC (Creative Commons), son de derechos de autor, que son modelos de contratos que sirven para otorgar públicamente el derecho de utilizar una publicación protegida por los derechos de autor. Entre menos restricciones implique una licencia, mayores serán las posibilidades de utilizar y distribuir un contenido.

- CC BY-NC-SA: esta licencia permite a los reutilizadores distribuir, remezclar, adaptar y construir a partir del material en cualquier medio o formato únicamente con fines no comerciales, y siempre y cuando se le otorgue la atribución al creador. Si remezcla, adapta o construye sobre el material, debe licenciar el material modificado bajo términos idénticos.

9.- Código

El código con el que se ha obtenido el dataset se encuentra en:

https://github.com/adrianglez77/Housing-price.git

El código ha sido creado en el lenguaje Python, ya que es muy conveniente para este tipo de tareas de Web Scraping, debido a que posee muchas herramientas y librerías para la extracción de datos de la web. A su vez, cuando se trabaje sobre un conjunto de datos muy grande, este lenguaje también facilita el uso de datasets con un gran volumen de datos.

Se han utilizado dos librerías durante el desarrollo del proyecto:

- **Requests**: se utiliza para realizar la petición a la página web donde se extraen los datos.
- **Beautiful Shop**: se utiliza para extraer los datos de htmls.

El programa tendrá una rutina que se encargará de escribir en un fichero ".csv", donde la primera vez se escribirá una cabecera con los nombres de los campos.

Dentro de la rutina, se realizará con la ayuda de "requests" un bucle para realizar peticiones a las 10 primeras páginas de la web.

Un aspecto que destacar es el denominado "**status code**", que hace referencia al código que devuelve la página cuando se realiza la petición. Así, veremos si hay una respuesta positiva o algún error en el cliente o servidor. Estos códigos pueden ser:

- Informational responses (códigos 100-199)
- Successful responses (códigos 200 – 299)
- Redirection messages (códigos 300 – 399)
- Client error responses (códigos 400 – 499)
- Server error responses (códigos 500 – 599)

Al seguir con la rutina, comenzaremos a utilizar "Beautiful Shop" para extraer los datos de la web. Crearemos una lista y con el método **".find_all"** se buscará la etiqueta **"div"** con el la clase del contenedor de los elementos.

Para obtener dentro de este contenedor los datos que necesitamos, crearemos un bucle en el que con el método **".find"** se buscará dentro del "div", los "span", y "li" que queremos obtener. Para eliminar las etiquetas y dejarlos en un mejor formato, se utilizará el **".text"**.

Para finalizar, guardaremos en un array y escribiremos con **"writerow"** en el fichero csv. Por último, aumentaremos el contador inicial, para volver a realizar el procedimiento anteriormente descrito, en la siguiente página.

10.- Dataset

Enlace al dataset publicado en Zenodo:

https://zenodo.org/record/7342443#.Y3uZdHbP2Uk

DOI: 10.5281/zenodo.7342443

11.- Vídeo

https://drive.google.com/file/d/1AyRnkwfvoYW-DhJQH3U6WN6UKOTQ9XFc/view?usp=share_link

Contribuciones

Contribuciones	Firma
Investigación previa	Adrián González González
Redacción de las respuestas	Adrián González González
Desarrollo del código	Adrián González González
Participación en el vídeo	Adrián González González