Tipología y ciclo de vida de los datos

PRÁCTICA 1: Web scrapping

Autores: Carmen Lobato Cassinello y Raúl Vicente Ferrer

Noviembre 2021

Contents

1. Contexto.

Desde hace unos años a la actualidad se está viviendo una gran emigración de los jóvenes españoles. Uno de los destinos más solicitados es Alemania. Esto ha dado lugar a multitud de debates, pero entre ellos hay uno recurrente, la diferencia de precios entre los dos países. Esta cuestión crea reticencia en muchas personas a la hora de dar el paso de emigrar. De los gastos a los que hay que hacer frente a diario, el mayor es el alquiler de una vivienda. Este dato nos puede hacer una idea de lo caro que es vivir en una determinada ciudad. Los portales más populares para acceder al alquiler de una vivienda en España y Alemania, son Idealista y WG-Gesucht respectivamente. Todos tenemos una opinión respecto a esto pero no suele estar fundamentada en una buena muestra de datos. Por esto, nuestro trabajo de *Web Scraping* se ha centrado en recopilar información sobre el precio del metro cuadrado de alquiler en las principales ciudades de ambos paises.

2. Título.

El título escogido es **HouseRentingSpainGermany.csv** que resume perfectamente el contenido del archivo.

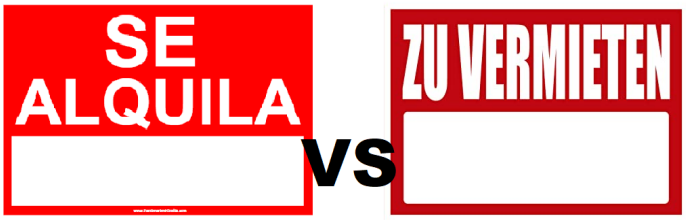
3. Descripción del dataset.

El conjunto de datos recoge variables que nos pueden ayudar a extraer información relacionada con el precio por metro cuadrado de alquiler de vivienda en las mayores ciudades de España y Alemania:

- España: Madrid, Barcelona, Valencia, Sevilla y Zaragoza.
- Alemania: Berlín, Hamburgo, Munich, Colonia y Frankfurt.

Cada uno de estos registros se corresponde a una única vivienda.

4. Representación gráfica.

- 

5. Contenido.

Para seleccionar las páginas web de las que extraer información, hemos aplicado las buenas prácticas recogidas en el material de la asignatura:

- Examinar el archivo *robots.txt* para asegurarnos de que el propietario de los datos no restringe el acceso a la información.
- Examinar las Webs y su estructura para identificar la ubicación de la información que queremos obtener.
- Análisis del tamaño y tecnología de la página web para evaluar la cantidad total de información a recoger.
- Utilización de la libreria BeautifulSoup para evitar el parseado manual.
- Hemos establecido un tiempo de espera entre peticiones para no los servidores web.

Por otro lado, respecto a la información a recoger, el dataset contiene diferentes variables que se listan y detallan a continuación:

- **Ciudad**: Nombre de la ciudad donde se ubica la vivienda.
- **Descripción**: Descripción de la vivienda en el anuncio.
- **Precio**: Precio mensual del alquiler de la vivienda.
- **Superficie**: metros cuadrados de la vivienda.

6. Agradecimientos.

Los datos han sido recolectados de las páginas webs Idealista y WG-Gesucht. Para ello se ha hecho uso del lenguaje de programación *Python* y de técnicas de *Web Scraping* para extraer la información alojada en las páginas HTML.

7. Inspiración.

Este conjunto de datos resulta interesante por muchos motivos. De entrada puede acabar con el debate de si es más caro vivir en España o en Alemania, y ayudarnos a tomar la decisión de emigrar o no. También puede emplearse para comparar ciudades dentro de uno de los dos países. Con una visión más amplia, se podría emplear este dataset para establecer políticas en determinadas ciudades que ayuden a regular los precios. Ya que teniendo información detallada se pueden establecer comparaciones fiables. Incluso con futuras versiones del dataset se podría estudiar como diferentes medidas han afectado a los precios.

8. Licencia.

9. Código.

El código utilizado para la extracción de la información requerida mediante *Web Scrapping* se puede acceder mediante la siguiente URL.

10. Dataset.

El dataset se ha publicado en formato CSV en Zenodo DOI. 10.5281/zenodo.5643913 con la siguiente descripción:

The dataset contains information extracted from the websites "Idealista" and "WG-Gesucht" about different prices of square meter rent in the main cities of Spain and Germany.The dataset contains data of November 2021.

Tabla contribuciones

Contribuciones	**Firma**
Investigación Previa	Carmen Lobato Cassinello, Raúl Vicente Ferrer
Redacción de las respuestas	Carmen Lobato Cassinello, Raúl Vicente Ferrer
Desarrollo del código	Carmen Lobato Cassinello, Raúl Vicente Ferrer