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.

En una época en la que la movilidad internacional está cada vez más presente en la carrera profesional de los trabajadores, se observa un incremento de la emigración de los jóvenes españoles. Uno de los destinos más solicitados es Alemania, lo cual ha dado lugar a multitud de debates a lo largo de los últimos años; entre ellos, uno de los más recurrentes es la diferencia de precios entre ambos países. Esta cuestión genera reticencia en muchas personas a la hora de dar el paso de emigrar.

Ante esta situación, resulta relevante conocer el valor de la vivienda en ambos países: de los gastos a los que hay que hacer frente a diario, el mayor será el alquiler del lugar de residencia. Este dato nos puede dar una idea de lo caro que será vivir en una determinada ciudad.

Gracias a digitalización de las agencias y a la aparición de páginas de consolidación de información, la información sobre el precio del alquiler está al alcance de todos. Los portales más populares para la búsqueda de alquiler de una vivienda en España y Alemania, los cuales serán utilizados para nuestro trabajo de *Web Scraping* son *Idealista* (https://www.idealista.com/) y *WG-Gesucht* (https://www.wg-gesucht.de/es/), respectivamente.

Utilizaremos las principales ciudades de cada país para recopilar información sobre el precio de alquiler por metro cuadrado. La actualización periódica del fichero generado permitiría comparar el precio mensual de vivienda en cada uno de estos países, facilitando la comparación del coste de vida, y analizar la posible evolución del mismo a lo largo del tiempo.

2. Título.

El título escogido es **HouseRentingSpainGermany.csv**: como se ha comentado anteriormente, el proyecto seleccionado hace uso de las agencias digitales de alquiler más populares en España y Alemania para obtener información sobre el precio mensual de la vivienda en ambos países.

3. Descripción del dataset.

El conjunto de datos generado recoge información el precio de alquiler por metro cuadrado en las viviendas de 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.

Los datos del dataset son de tipo carácter o numérico, quedando excluidos otros tipos de variables. Adicionalmente, si la página de origen no contiene información para algún campo, esta variable aparecerá vacía en el conjunto de datos; se ha decidido mantener estas entradas, dejando a decisión del usuario lo que hacer con ellas. No es necesaria ninguna acción de limpieza adicional.

4. Representación gráfica.

A continuación se muestra la representación gráfica del conjunto de datos escogido:

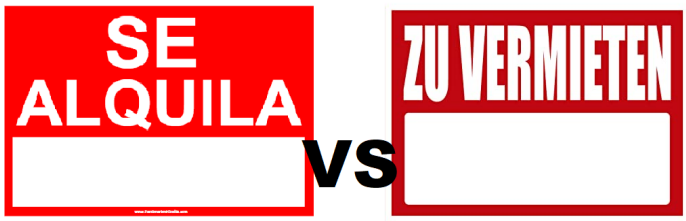

Figure 1: Alquiler vs Vermieten

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 saturar 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.

Las páginas mencionadas, o los terceros con los que tienen acuerdos, son los propietarios de todo el contenido que aparece en sus webs; adicionalmente, en estas páginas se realiza un trabajo de procesamiento de datos

previo a la muestra de contenido a los usuarios. Idealista y WG-Gesucht serán, por tanto, propietarios de los datos generados en esta práctica.

En relación a análisis anteriores, las propias páginas realizan estudios inmobiliarios con los datos que recogen y procesan, los cuales están abiertos al público para su consulta.

7. Inspiración.

Este conjunto de datos resulta interesante porque contribuye a eliminar la subjetividad a la hora de debatir si es más caro vivir en España o en Alemania; los datos están ahí, se pueden recopilar y estudiar, y su análisis puede ayudarnos a tomar la decisión de emigrar o no. También pueden emplearse para comparar el precio de alquiler en diferentes ciudades dentro de uno de los dos países.

Con una visión más amplia, se podría emplear este dataset para analizar la evolución de los precios a lo largo del tiempo y establecer políticas que ayuden a regularlos: teniendo información detallada se pueden establecer comparaciones fiables. Con futuras versiones del dataset se podría estudiar cómo diferentes medidas establecidas han afectado a los precios.

8. Licencia.

Se ha decidido hacer uso de la licencia **Released Under CC BY-NC-SA 4.0 License**, la cual ofrece libertades a los usuarios a la par que derechos a los propietarios. Se permite su distribución, se reconoce al autor de la obra y se permite editar el código fuente. No obstante, no se permite lucrarse económicamente con el mismo ni privatizar el software con una licencia que altere las libertades anteriormente expuestas. Los propietarios de los datos serán siempre Idealista y WG-Gesucht, y así deberá hacerse saber en cualquier utilización de los mismos.

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 as 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