TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Práctica 1: Web Scraping

Autores: Elisa Fernández Maraver y Francisco Javier Cea Barceló

Contexto

Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Se ha decidido realizar el ejercicio de recolección de datos utilizando métodos de Web Scraping en el portal luxemburgués de anuncios inmobiliarios https://www.wortimmo.lu/. La razón de esta elección se deriva del hecho de que uno de los integrantes del grupo reside actualmente en Luxemburgo, y de cara al próximo año 2020 deberá encontrar un nuevo apartamento de alquiler al que mudarse.

Cualquier proceso de búsqueda de piso de alquiler supone un desafío a la vez que una inversión de tiempo considerable. Con la intención de maximizar el uso del tiempo invertido en buscar el piso idóneo se ha querido agrupar en un solo fichero la información más relevante a los anuncios de alquiler en todo Luxemburgo. Esto permitiría un estudio posterior del dataset para encontrar las mejores ofertas en el mercado.

Dataset

Definir un título para el dataset. Elegir un título que sea descriptivo.

El nombre del dataset en el que se almacenarán todos los datos recopilados es **mm_dd_YYYY_Pisos-Alquiler_Luxemburgo.csv**, de esta forma, si se consultan distintos días puede tenerse un seguimiento de la última vez que se consultaron los pisos y la fecha en la que estaban disponibles.

Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído

El dataset extraído recoge un listado con los anuncios de apartamentos ofertados para su alquiler en todo el país de Luxemburgo, ordenado por orden de publicación. El conjunto de datos se almacena en un fichero .csv que contiene 9 columnas que aportan información relevante para cada anuncio.

Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

A continuación, se muestra un ejemplo del dataset obtenido para los cinco últimos anuncios en función de los filtros elegidos:

Description	Price (€)	Area (m²)	Num. Rooms	Zone	Num. Parkings	Agency	Contact	URL
Apartment with 3 bedroom(s) to rent in Luxembo...	2400.0	109.0	3	Luxembourg-Kirchberg	1	Sylvie Becker Sàrl	tel:+352 26 33 46 46	https://www.wortimmo.lu/en/rent/apartment/cent...
Apartment with 2 bedroom(s) to rent in Luxembo...	2500.0	110.0	2	Luxembourg-Kirchberg	1	MKA	tel:+352 26 31 00 08	https://www.wortimmo.lu/en/rent/apartment/cent...
Apartment with 3 bedroom(s) to rent in Luxembo...	2300.0	83.0	3	Luxembourg-Kirchberg	1	REMAX Real Estate Solutions	tel:+352 28 84 02	https://www.wortimmo.lu/en/new/apartment/centr...
Apartment with 2 bedroom(s) to rent in Luxembo...	1950.0	80.0	2	Luxembourg-Cents	1	Immo Frank	tel:+352 691 19 65 20	https://www.wortimmo.lu/en/rent/apartment/cent...
Apartment with 4 bedroom(s) to rent in Luxembo...	3600.0	200.0	4	Luxembourg-Centre	1	Weckbecker S.A.	tel:+352 22 25 92	https://www.wortimmo.lu/en/rent/apartment/cent...

***Figura 1**: Alquiler en todo Luxemburgo hasta un radio de 2Km de un Apartamento con un mínimo de dos dormitorios y que no supere los 4500€.*

Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Para cada registro del dataset, que se corresponde con un anuncio inmobiliario de alquiler en todo el país de Luxemburgo, se recogen los siguientes campos:

- **Description**: nombre del anuncio.
- **Price**: precio de la vivienda consultada (€).
- **Area**: superficie en m^2.
- **Num. Rooms**: número de habitaciones.
- **Num. Parkings**: número de parkings.
- **Zone**: Zona en la que se encuentra (ciudad, pueblo, etc).
- **Agency**: agencia responsable de la propiedad.
- **Contact**: teléfono de contacto.
- **URL**: link al anuncio mencionado.

Los datos se recogen en por orden de publicación (más recientes primero) y en función de la disponibilidad de la propiedad anunciada en un determinado momento. Estos anuncios son publicados por las propias agencias en la página y recogidos mediante técnicas de *Web Scraping* en Python, hasta obtener un csv con la recopilación de los mismos.

La idea es que el usuario pueda elegir una serie de filtros para realizar su búsqueda y obtenga un dataset con la información principal de los anuncios que cumplen dichas especificaciones, simplificando así la búsqueda. Además, se puede definir a priori el número de anuncios a visualizar. Así, si se está haciendo una búsqueda continuada cada cierto periodo de tiempo, únicamente se visualizarán los anuncios más nuevos.

Los filtros que pueden aplicarse son:

- Tipo de transacción: compra o alquiler (obligatorio).
- Localización: por país, región o ciudad (obligatorio).
- Radio: el radio de búsqueda dentro de la localización.
- Tipo de propiedad: para simplificar la búsqueda sólo se puede elegir un tipo de propiedad en cada dataset (apartamento, duplex, granja, etc) (obligatorio).
- Precio mínimo y máximo.

- Número de dormitorios mínimo y máximo.
- Superficie mínima y máxima de la propiedad.

En el caso de la localización y el tipo de propiedad se utilizan una serie de códigos predefinidos en el archivo *readme.md* que se puede encontrar en [1].

Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecemos al portal inmobiliario luxemburgués **Wortimmo.lu**, el cual ofrece gratuitamente, entre otros servicios, un lugar donde las distintas agencias inmobiliarias pueden anunciar viviendas disponibles para su alquiler. A continuación, se detallan las menciones legales del sitio web:

- Editor del sitio www.wortimmo.lu: Régie Immobilière s.à r.l.
- Sociedad de responsabilidad limitada
- Sede: 2 rue Christophe Plantin, L-2339 Luxemburgo
- Inscrita en el Registro de Empresas y Comercio de Luxemburgo con el número B190938
- Número de IVA: LU 273 585 88
- Autorización de establecimiento n°10052559/0 emitida por el Ministerio de Economía de Luxemburgo (19-21 Boulevard Royal, 2449 Luxemburgo) el 20 de octubre de 2014.
- Teléfono de contacto: (+352) 24 84 95 1
- Email de contacto: info@wortimmo.lu

Inspiración

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El conjunto de datos recopilado aporta una información muy valiosa a cualquier persona que se encuentre actualmente en búsqueda de un piso de alquiler en la región de Luxemburgo. La alta demanda y una oferta no tan amplia acompañada de unos precios de alquiler bastante elevados, hacen que este proceso de búsqueda pueda llegar a ser bastante complicado a la vez que intensivo en cuanto a tiempo se refiere. Por ello, el dataset pretende facilitar todo este proceso aportando la información relevante para, primero, analizar el precio de la vivienda de alquiler en las distintas zonas del país y, segundo, poder encontrar más fácilmente aquellas ofertas que se encuentren por debajo del precio de mercado y que puedan constituir una oportunidad atractiva.

Además, debido a que el resultado final es un dataset, posibilita realizar determinados estudios de mercado y análisis de gran utilizad en otras áreas de aplicación.

Licencia

Seleccione una de estas licencias para su dataset y explique el motivo de su selección.

Antes de comenzar con este ejercicio, se procedió a analizar el fichero robots.txt del sitio web para estudiar las posibles restricciones de uso/acceso impuestas por el mismo. Tras la evaluación se concluyó que en el mismo no se incluye ninguna restricción. Para la publicación de este conjunto de datos se ha

decidido utilizar una licencia **Released Under CC BY-SA 4.0.** Ésta se adapta perfectamente al ámbito del trabajo realizado con la recolección de datos del portal inmobiliario bajo las siguientes cláusulas:

- Se permite a otros remezclar, ajustar y construir sobre su trabajo incluso con fines comerciales, siempre que lo acrediten y otorguen licencias de sus nuevas creaciones bajo los mismos términos.

Código fuente y dataset

Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. Presentar el dataset en formato CSV

El código fuente escrito para la extracción de datos, así como el dataset generado en csv, se ha publicado en un repositorio de Github al que puede accederse en el siguiente enlace [1]

Bibliografía

[1] https://github.com/elisafm4/wortimmo_scraping

[2] Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

[3] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

[3] Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.