Integrantes: José Javier Martí Camarasa (JJMC)
Luis Villazón Esteban (LVE)

Asignatura: Tipología y Ciclo de Vida de los Datos

Resolución PRAC1 9-Nov-2020:

1. *Wiki con los componentes del Grupo*

   https://github.com/Villaz/idealistaScraper/wiki

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

Actualmente la compra-venta y alquiler de vivienda es un tema de actualidad, teniendo especial interés los precios y la disponibilidad de vivienda en las diferentes zonas de la geografía Española.

Para ello se ha decidido recolectar información de las viviendas ofertadas en territorios específicos para así poder estudiar cual es la evolución del precio y la cantidad de las ofertas a lo largo del tiempo.

Para conseguir este objetivo se ha decidido extraer la información del portal inmobiliario Idealista, dado que Idealista es considerado como el portal de viviendas más importante y grande del país posiblemente se puede extraer del mismo la información más completa y veraz acerca de cómo se encuentra el parque inmobiliario en un instante de tiempo específico.

3. *Definir un título para el Dataset. Elegir un título que sea descriptivo.*

El título del dataset depende de la tipología y zona geográfica donde se realiza el proceso de scraping. El formato seguirá la siguiente estructura:

venta-viviendas_oviedo-asturias.csv: Conteniendo los datos de cada inmueble.

venta-viviendas_oviedo-asturias_images.csv :Conteniendo la ruta de las imágenes de cada inmueble.

4. Descripción del Dataset.

Dataset de viviendas:

Contiene los datos principales de cada uno de los anuncios publicados en idealista. Tal como se ha comentado en el apartado anterior el nombre del mismo es "tipo_transaccion"-"tipologia"_"ciudad"-"provincia".csv Por lo tanto podrexistir tantos CSVs como distintas bsquedas existan.

Donde:

<tipo_transaccion>-_-<tipología>-_-<ciudad>.csv

<tipo_transaccion>-_-<tipología>-_-<ciudad>-_-<provincia_images>.csv

tipo_transaccion: puede ser venta o alquiler.

Tipología: Actualmente solo puede ser **viviendas**, pero se puede extender a cualquier tipologia existente en el portal de Idealista.

Ciudad :Ciudad donde se realiza la búsqueda.

Provincia: Provincia a la que pertenece la ciudad.

Dataset imágenes:

Contiene las URLs de las imagenes encontradas para cada uno de los inmuebles capturados. Una vez que se tiene este dataframe hay que ejecutar un nuevo scraper para descargar las imágenes a un repositorio propio.

El formato del dataset es un CSV con cabecera y utilizando una coma "," para separar los atributos.

Los atributos que contiene el dataset son:

url: URL donde se encuentra la imagen a descargar.

code: Código del inmueble al que pertenece la imagen.

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

Con la imagen detallada abajo, con el objetivo de recolectar información del parque inmobiliario, podemos establecer una relación entre la ilustración y los dos datasets mencionados y descritos en el apartado anterior, "*Dataset Viviendas*" con la información relativa a cada inmueble y el "*Dataset Imágenes*" donde guardamos la información de todas las imágenes relacionadas con cada uno de los inmuebles de "*Dataset Viviendas*",

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	code	link	address	barrio	distrito	ciudad	lat	lon	price	area	has_elevator	floor	exterior	rooms
2	91431452	https://www.idealista.com/inmuebl	Olivares	Barrio Olivares	Distrito Buenavista-Ería-Montec	Oviedo	43.3619258	-5.8781634	590000	40				4
3	90992307	https://www.idealista.com/inmuebl	Calle Arzobispo Guisasola	Barrio Auditorio-Seminario-Pa	Distrito Centro-Casco Histórico	Oviedo	43.357596	-5.841457	209000	116	True	9.0	True	3
4	90084540	https://www.idealista.com/inmuebl	Calle Asturias	Barrio Parque San Francisco-U	Distrito Centro-Casco Histórico	Oviedo			750000	219	True	4.0	True	5
5	91431452	https://www.idealista.com/inmuebl	Olivares	Barrio Olivares	Distrito Buenavista-Ería-Montec	Oviedo	43.361926	-5.878163	590000	40				4
6	91431452	https://www.idealista.com/inmuebl	Olivares	Barrio Olivares	Distrito Buenavista-Ería-Montec	Oviedo			590000	40				4
7	90992307	https://www.idealista.com/inmuebl	Calle Arzobispo Guisasola	Barrio Auditorio-Seminario-Pa	Distrito Centro-Casco Histórico	Oviedo	43.357596	-5.841457	209000	116	True	9.0	True	3
8	90084540	https://www.idealista.	Calle Asturias	Barrio Parque San	Distrito Centro-Casc	Oviedo	43.364161	-5.853979	750000	219	True	4.0	True	5

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

Hemos obtenido dos **Datasets**. Un dataset corresponde a los datos del inmueble y otro dataset con las imágenes de inmueble, **se relacionan a través del código de anuncio:**

EJ: https://www.idealista.com/inmueble/91431452/

Dataset venta-viviendas_oviedo-asturias.csv :

Code [int]: Código del inmueble.

Link [string]: Enlace al anuncio.

Address [string]: Dirección del inmueble. Suele ser inexacta por motivos de la inmobiliaria.+

Barrio [string]: Barrio donde se ubica el inmueble.

Distrito [string]: Distrito donde se ubica el inmueble.

Ciudad [string]: Ciudad del inmueble.

Lat y Lon [float]: Coordenadas donde se encuentra el inmueble.

Price [float]: Precio del inmueble.

Area [int]: Superficie del inmueble en metros cuadrados.

Has_elevator [boolean]: India si tiene ascensor o no. {True,””}

Floor[int]: Altura respecto a la finca

Exterior[boolean]: Si el inmueble es exterior o no . {True,””}

Rooms [int]: Num de habitaciones

Dataset: venta-viviendas_oviedo-asturias_images.csv

Code [int]: Código del inmueble.

URL: Diferentes imágenes del inmueble

La **recolección de la información** se realizó el día 27.10.2020. Estos datos se recogieron mediante técnicas de Web scraping utilizando la librería Selenium con Python.

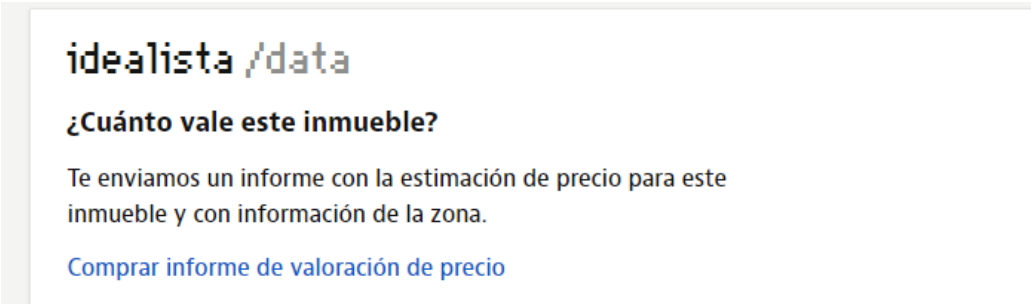
7. *Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).*

Idealista
Es una plataforma web y Aplicación móvil, de propiedades que nació para cambiar la forma en que se venden y arriendan inmuebles en España.

Da servicio tanto a inmobiliarias , agentes de la propiedad y a usuarios.

Quienes busquen una nueva vivienda o quieran invertir en propiedades encontrarán en Idealista, además de las ofertas disponibles**, valiosa información para tomar las mejores decisiones**: promedios de precios, características de los barrios, comparaciones, etc.

Todo esto a través de:



Idealista también ofrece una propuesta novedosa para las inmobiliarias o agentes que quieran vender una propiedad, **ya que el modelo de negocios se basa en entregar contactos de calidad.**

8. *inspiración. Explique por que es interesante este conjunto de datos y que preguntas se pretenden responder.*

Se ha decidido recolectar información de las viviendas ofertadas en territorios específicos para así poder estudiar cual es la evolución del precio y la cantidad de las ofertas a lo largo del tiempo, extrayendo la información del portal inmobiliario Idealista, dado que Idealista es considerado como el portal de viviendas más importante y grande del país.

9. *Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:*

Released Under CC0: Public Domain License

- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License.

Hemos elegido la licencia **CC BY-NC-SA 4.0 License-**

El motivo de la selección es una licencia de código abierto, al ser un trabajo práctico hemos pensado que es importante que esta información se pueda estudiar, compartir con la finalidad de que muchas usuarios tengan acceso y se puedan beneficiar.

Esta información se podrá utilizar bajo una propósito No comercial, y en caso de modificar los datasets, se deben de distribuir bajo la misma licencia que los originales.

Se debe de dar crédito a los creadores, en el caso de uso de esta información y mencionar y guardar registro de las modificaciones en caso de que las hubiera

10. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Adjuntamos la URL del GitHub donde está el código,

https://github.com/Villaz/idealistaScraper

11. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI)con una breve descripción.

EL **DOI** obtenido por la publicación del dataset es "**10.5281/zenodo.4147720**".

Descripcion: "Property price research in Oviedo Asturias, carried out with Selenium and Python, mainly to obtain price evolution and amount of offers

Contribuciones	**Firma**
Investigación previa	**JJMC, LVE**
Redacción de las respuestas	**JJMC, LVE**
Desarrollo código	**JJMC, LVE**

12. Recursos:

a. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
b. Masip, D. El lenguaje Python. Editorial UOC.
c. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
d. https://www.idealista.com
e. How to Use Selenium to Web-Scrape with Example