¿Cómo podemos capturar los datos de la web?

Elaborar un caso práctico orientado a aprender a identificar datos relevantes para un proyecto analítico y usar las herramientas de extracción de datos. Proyecto trabajado en grupos de dos personas.

Práctica 1

Autores:
Lukaz Martin Doehne & Pablo Vadillo Berganza
Profesora: Laia Subirats Maté
Asignatura: Tipología y ciclo de vida de los datos
Tutora: Elena Villa Estebaranz
20 de noviembre de 2022

0. INTRODUCCIÓN

El objetivo de este documento es el de desarrollar el trabajo realizado durante la primera práctica de la asignatura Tipología y ciclo de vida de los datos. El ejercicio ha consistido en realizar web scraping en una página web para posteriormente extraer los datos en un archivo CSV.

Github: https://github.com/LukazMartin/UOC-WebScraping-project

1. CONTEXTO

El contexto en el que se ha recolectado la información es la obtención de datos que permitan realizar un análisis predictivo del precio de la vivienda en Barcelona a partir de ciertas variables.

La página web elegida es la de la empresa inmobiliaria de origen alemán **Engel & Völkers**, donde se han buscado viviendas residenciales en venta.

Los datos obtenidos son perfectos para desarrollar el análisis planteado ya que se trata de un conjunto de viviendas con un precio ya establecido, lo que nos posibilitará desarrollar un *modelo supervisado*, concretamente un *algoritmo de regresión*.

Dividiendo los datos en 2 conjuntos, uno de entrenamiento y otro de prueba, podremos crear un algoritmo que sea capaz de analizar el nivel de correlación entre las diferentes variables independientes y la dependiente (en este caso, el valor de la vivienda).

La dirección principal del sitio web utilizado es la siguiente:

https://www.engelvoelkers.com/es-es/propiedades/comprar-vivienda/barcelona/

y concretamente, la utilizada a la hora de realizar el web scraping es:

https://www.engelvoelkers.com/es/search/?q=&startIndex=0&businessArea=residential&sortOrder=DESC&sortField=newestProfileCreationTimestamp&pageSize=18&facets=bsnssr%3Aresidential%3Bcntry%3Aspain%3Bobjcttyp%3Acondo%3Brgn%3Abarcelona%3Btyp%3Abuy%3B

2. TÍTULO

El título definido para este dataset es **Barcelona apartments from Engel&Voelkers**.

3. DESCRIPCIÓN DEL DATASET

El dataset está compuesto por **1.254 registros con 12 variables**.

Representa las características de los apartamentos residenciales en venta en Barcelona publicados en la página web de Engel & Völkers.

Los datos han sido capturados en idioma español, si bien se ha optado por utilizar el inglés a la hora de definir los diferentes campos.

7 de las variables son categóricas y 5 numéricas, las cuáles desglosamos en el apartado CONTENIDO.

4. REPRESENTACIÓN GRÁFICA

Una buena representación gráfica facilita la comprensión de la información que hay detrás de un dataset, así como de los posibles análisis que pueden derivar de él.

A continuación presentamos la imagen propuesta como representación gráfica de nuestro dataset,

Creemos que la imagen propuesta resume de manera sencilla la idea detrás del conjunto de datos. Se trata de viviendas en la provincia de Barcelona extraídas de la página de *engel&voelkers*.

5. CONTENIDO

En este apartado procedemos a explicar los diferentes campos que podemos encontrar en el dataset.

Los campos definidos son los siguientes:

- **id** (Identificador): Se trata de un campo *categórico* con un código que identifica las diferentes publicaciones de las viviendas.

  Está compuesto por W- seguido de 6 dígitos. No existen valores en blanco, y cada valor es único. De cara a realizar el análisis, no será de utilidad pero es un campo necesario ya que sirve como identificador de cada uno de los inmuebles y nos permite buscarlo en la página web en caso de incoherencias. No existen valores nulos.

- **title** (Título): Se trata de un campo *categórico*, y recoge los títulos que figuran en cada uno de los anuncios de las viviendas.
  Es meramente informativo, y puede ser utilizado para hacerse una idea acerca de las principales características destacables de la vivienda. No existen valores en blanco.

- **location** (Localización): Se trata de un campo *categórico* y recoge la localización de las viviendas.

  Barrios, calles, comarcas y pueblos. De cara a un posible preprocesado, resultaría interesante subdividir esta información para poder realizar agrupaciones y analizar las posibles correlaciones entre las zonas y el precio de los apartamentos. No existen valores nulos.

- **location_status** (Estado de la ubicación): Se trata de un campo *categórico* y hace referencia a la ubicación de la vivienda.

Se clasifican las ubicaciones de los apartamentos en *Excelente*, *Muy bien*, *Bien* y *Regular*. 167 registros son nulos.

A falta de realizar un análisis de correlación entre esta variable y el precio de la vivienda parece bastante intuitivo que las ubicaciones *Excelentes* presentarán mayores precios que las que son *Bien* ó *Regular*.

- **status** (Estado): Se trata de un campo *categórico* y determina el estado de la vivienda, entendiendo por esto la situación de sus instalaciones.

  Toma los siguientes valores: *Excelente*, *Muy bien*, *Bien*, *Regular*, *Renovado*, *Parcialmente renovado*, *Necesita renovaciones*, *Necesita restauración* y *Otros*. Hay 189 registros nulos.

  La correlación con el precio de la vivienda en este caso, intuitivamente se supondría que cuanto mejor sea el estado del apartamento mayor será el precio de la vivienda.

- **year** (Año): Se trata de un campo *numérico* y sólo toma valores enteros del año de construcción de la vivienda.

  En nuestro dataset, los valores oscilan entre 1700 y 2023. Hay 178 valores nulos.

  La correlación con el precio de la vivienda en este caso, intuitivamente se supondría que cuanto más nuevo el apartamento mayor será el precio de la vivienda.

- **area** (Superficie): Este campo se corresponde con la superficie de la vivienda, en metros cuadrados.

  Al incluir la unidad de medida (m^2), el formato actual es *categórico*, si bien la naturaleza de la variable es *numérica* y en un posible preprocesado sería conveniente transformarla.

  Los valores de los registros van desde $30m^2$ a $110.000m^2$. Esta variable se supone que tendrá mucho peso a la hora de predecir el precio de la vivienda, ya que probablemente exista una correlación directa entre el tamaño de la propiedad y su precio.

- **bathrooms** (Baños): Se trata de un campo *numérico* y hace referencia al número de baños de la vivienda.

  Los valores van desde 1 a 10. Existen 2 registros nulos.

  La correlación con el precio de la vivienda en este caso, intuitivamente se supondría que cuanto más baños tenga el apartamento mayor será el precio de la vivienda.

- **bedrooms** (Habitaciones): Se trata de un campo *numérico* y hace referencia al número de habitaciones del inmueble.

  Los valores van desde 1 y 15. Existen 2 registros nulos.

  La correlación con el precio de la vivienda en este caso, intuitivamente se supondría que cuantas más habitaciones tenga el apartamento mayor será el precio de la vivienda.

- **heating_type** (Tipo de calefacción): Se trata de un campo *categórico* y hace referencia al tipo de calefacción de la vivienda.

  Para realizar un análisis más sencillo se ha concatenado este atributo. Puede tomar los siguientes valores (listados sin ningún orden en particular):

Bombadecalor	BomdadecalorCircuito	BombadecalorGas
BombadecalorGasSolar	BombadecalorPellets	BombadecalorSolar
Calefacciónccentralizada	Central	CentralCalefacciónccentralizada
CentralSueloradiante	Circuito	CircuitoPellets
CorrienteEléctrica	CorrienteEléctricaBombadecalor	CorrienteEléctricaCircuito
CorrienteEléctricaGas	CorrienteEléctricaPellets	CorrienteEléctricaSolar
Estufa	Gas	Gasoil
LongDistanceHeating	nocentralizada	Pellets
PelletsGasSolar	Solar	Sueloradiante

310 registros son nulos. A la hora de modelar no es trivial el valor que aportará esta variable. Puede ser que el precio de la vivienda sea dependiente o independiente con el tipo de calefacción.

Como podemos observar, algunos de los valores son el resultado de concatenar otros dos tipos de calefacción, por lo tanto aquí podríamos seguir dos posibles estrategias:

· Convertir cada uno de los valores (*Gas*, *Solar*...) en campos, de forma que tengamos una variable lógica que tome los valores YES o NO dependiendo de si la cadena de caracteres contiene dicho campo o no. Por ejemplo si el registro tiene el valor *CorrienteEléctricaSolar*, tomará el valor YES en los campos *Corriente Eléctrica* y *Solar* y NO en el resto.

· Convertir cada variable categórica en numérica. Por ejemplo, que todos los campos *CorrienteEléctricaSolar* convertirlos en el valor 0.

- **energy_class** (Clase energética): De nuevo tenemos una variable de naturaleza categórica y hace referencia a la clase de eficiencia energética.

  Toma valores entre A+ y G. Existen 176 registros nulos. Esta etiqueta valora el consumo energético de la vivienda siendo A la más eficiente y G la menos, tal y como se puede observar en la siguiente imagen:

Figura 1: Calificación energética. [1]

[1] Fuente: https://www.engelvoelkers.com/es-es/valencia/certificado-de-eficiencia-energ%C3%A9tica/

Será interesante observar el impacto que pueda llegar a tener la eficiencia energética de una vivienda en su precio.

- **price** (precio): Finalmente, tenemos la variable a predecir, el precio de la vivienda.

  De naturaleza *numérica*, pero en el dataset originado del web scraping figura como *categórica*. Debido a la inclusión de la unidad de medida, en este caso EUR.

  Toma valores entre 75.000€ y 9.500.000€. Hay 8 registros con el valor A *consultar* y 1 con el campo en blanco.

En cuanto al periodo de tiempo de los datos sería desde la creación de la página web de Engel & Völkers hasta la actualidad.

6. PROPIETARIO

El propietario de los datos es, por un lado la propia Engel & Völkers, y por otro lado cada uno de los propietarios de las viviendas en venta.

En cuanto a la inspiración que ha motivado este análisis, los modelos predictivos sobre los precios de las viviendas son muy recurrentes en plataformas como Kaggle ó Github y por lo tanto podemos encontrar múltiples ejemplos de éstos.

A uno de ellos se puede acceder a través del siguiente enlace: https://www.kaggle.com/code/gauravduttakiit/housing-price-prediction-by-linear-regression

En él, se realiza una regresión lineal para predecir el precio de una vivienda a partir del resto de variables independientes, utilizando un dataset de características similares al obtenido en esta práctica. El análisis fue realizado por el usuario de Kaggle Gaurav Dutta.

También podemos encontrar ejemplos de datasets similares al nuestro obtenidos por web scraping, uno de ellos se puede encontrar también en Kaggle a través del siguiente:

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

En este caso, se trata de un dataset obtenido por web scraping en la página web https://www.realtor.com/ por el usuario Ahmed Shahriar Sakib. El usuario plantea varios análisis a realizar a partir del dataset:

Predecir el precio de la vivienda.
Encontrar las ubicaciones con precios más elevados.
Encontrar la correlación entre el precio y otros atributos de las viviendas.
Encontrar si existen tendencias que influyen en los precios de las viviendas.

7. INSPIRACIÓN

Como hemos comentado a lo largo del informe, la principal pregunta a la que queremos dar respuesta con el dataset es a la de si es posible **predecir el valor de los apartamentos en Barcelona a partir de sus características**.

Creemos que el dataset obtenido es realmente interesante, por un lado, porque está compuesto por campos tanto categóricos como numéricos. Por otro lado, contamos con variables que a priori mantienen una relación directa clara con el precio (*superficie*, *número de habitaciones*, *baños*...), pero también contamos con otras variables, como la *localización*, *tipo de calefacción* ó *la eficiencia energética*, que no presentan una correlación tan evidente con el precio y por tanto será interesante comprobar el impacto que tienen en la predicción del valor del inmueble.

Pero por encima de todo, **lo interesante de este dataset es que es real**, se ha obtenido a partir de la página web de una inmobiliaria y por tanto, los campos recogidos son características que los compradores dan importancia en la búsqueda de vivienda. Por lo tanto, encontrar qué campos tienen más peso a la hora de predecir el precio posibilitará el poder realizar inversiones más inteligentes.

También es interesante este análisis si estamos **interesados en invertir en vivienda**. Si construimos un modelo con una precisión elevada podemos introducir los datos de una casa en venta y especular si la casa está sobrevalorada o no. Si predecimos que la casa vale más que su precio real, podríamos invertir en ella ya que en un futuro se podrá vender mejor

En esta línea el análisis a realizar es muy similar a los ejemplos presentados en los apartados anteriores. En el primer ejemplo teníamos una regresión lineal, mientras que en el segundo ejemplo se utilizan diferentes modelos para analizar su precisión. La variable a predecir, es conocida en nuestros datos, por lo que el algoritmo que más se adecua a su predicción es la regresión.

8. LICENCIA

La licencia seleccionada para el dataset resultante es CC BY-SA 4.0 License[2]. Esta licencia presenta los siguientes términos,

- Se debe de dar crédito de manera adecuada, incluir un enlace a la licencia, e indicar si se han realizado cambios. De esta manera, se valora el trabajo ya realizado y se facilita la información acerca de cómo se ha contribuido en las mejoras del trabajo ya hecho.

- Si se crea a partir del material, la nueva contribución también debe ir bajo la misma licencia. De esta manera, la licencia que el autor original seleccionó continuará siendo usada en futuras contribuciones.

A diferencia de la licencia CC BY-NC-SA 4.0, en este caso sí se permite el uso comercial.

9. CÓDIGO

Repositorio: https://github.com/LukazMartin/UOC-WebScraping-project

Dataset: https://github.com/LukazMartin/UOC-WebScraping-project/blob/main/dataset

Source: https://github.com/LukazMartin/UOC-WebScraping-project/blob/main/source

La lógica del código se divide en 2 partes: explorar links y extraer información.

Explorar links

El proceso consiste en:

· Hacer un GET a la url deseada.

[2] Fuente: https://creativecommons.org/licenses/by-sa/4.0/

· Conseguir todos los links de casas a la venta

· Conseguir el link de la siguiente página

Así pues, primero nos enfocamos en acceder a los datos.

```
def __get_page(self, count=0):

    # Some pages don't work without headers. See robots.txt
    if not self.next_url:
        page_ = requests.get(self.url, headers={"User-Agent": self.user_agent})
    else:
        page_ = requests.get(self.next_url, headers={"User-Agent": self.user_agent})

    if not page_.ok:  # Only requests 2XX are valid
        if count >= 5:
            raise Exception(f"Could not get page {self.url}")
        sleep(randint(1, 5))
        self.__get_page(count+1)

    self.page = page_
```

Buscamos seguir unas buenas prácticas para que no se nos deniegue el acceso. Cómo usar User-agent e introducir una espera aleatoria en caso que hayamos saturado el servidor. Si por alguna razón no obtenemos un código 2XX en la respuesta probamos hasta 5 veces más antes de dar el proceso cómo fallado.

Una vez tengamos la página, utilizamos *BeautifulSoup* para filtrar y guardar todos los links que son relevantes. En nuestro caso, son los que contienen la palabra *propiedad* en la url del link.

Por último hemos de acceder a la siguiente página para realizar el mismo proceso. Cuando navegamos en la web es tan fácil como darle a la siguiente página al final. Como podemos ver en la siguiente imagen.

Sin embargo la url en *engelvoelkers.com* está construida dinámicamente. La primera página contiene 16 links a casas y se encontrarán bajo la siguiente url:

`www.engelvoelkers.com/es/search/?q=&`**`startIndex=0`**`&Area…`

La siguiente página tendría la siguiente url:

`www.engelvoelkers.com/es/search/?q=&`**`startIndex=16`**`&Area…`

Por lo tanto para acceder a la siguiente página modificamos el valor *startIndex* de la url.

```
63      def __get_next_link(self):
64          index = self.url.index(self.index_parameter)
65          self.current_index += self.pagination_index
66
67          if self.next_url is None:  # There are 2 advertisements on the first page instead of houses
68              self.current_index -= 2
69
70          self.next_url = self.url[:index + len(self.index_parameter) + 1] + str(self.current_index)
71          self.next_url += self.url[index + len(self.index_parameter) + 2:]
```

En cada página encontramos 18 links a casas a excepción de la primera. Donde hay 2 anuncios en lugar de links a casas por lo que restamos 2 en la línea 68.

Extraer información

Una vez tenemos todos los links importantes en una lista. Los iteramos y analizamos con *BeautifulSoup*.

Para ello hacemos uso de la función *find()* dónde buscamos los atributos relevantes. En particular ha sido un reto filtrar las características que queremos dado que muchas casas tienen diferentes características (ascensor, terraza, parking…) y cada link nos devuelve un diccionario de información diferente. Por ello hemos optado por filtrar la información más necesaria y que salía en la mayoría de casas. A continuación un ejemplo de cómo hemos seleccionado algunas:

```
bedrooms, bathrooms, area, price = None, None, None, None
for i in range(len(fact_titles)):
    category = fact_titles[i].replace(" ", "")
    if category == 'Dormitorios' or category == 'Cuartos':
        bedrooms = fact_values[i]
    elif category == 'Baños':
        bathrooms = fact_values[i]
    elif category == 'Superficiehabitableaprox.' or category == 'Superficieconstruidaaprox.':
        area = fact_values[i]
    elif category == 'Precio':
        price = fact_values[i]
```

Finalmente realizamos un breve preprocesado para guardar los datos en un csv. Esto consiste en quitar comas (,) y espacios (“ “).

10. DATASET

A continuación encontramos el link a Zenodo donde publicamos el dataset y su perteneciente DOI.

Zenodo link: https://zenodo.org/record/7337520#.Y3jKWOzMJ6p

DOI: 10.5281/zenodo.7337520

11. VIDEO

En este último apartado presentamos el vídeo explicativo realizado. Se ha subido al Google Drive asociado a la UOC.

Enlace: https://drive.google.com/file/d/1ShH578AMEN-NsQ9Qgpdy_e0eGqJ_BF4-/view?usp=sharing

Contribuciones	Firma
Investigación Previa	L.M.D., P.V.B.
Redacción de las respuestas	L.M.D., P.V.B.
Desarrollo del código	L.M.D., P.V.B.
Participación en el vídeo	L.M.D., P.V.B.