Memoria

PRÁCTICA 1

Autora: Fanfan Yang

Noviembre 2022

Esta memoria trata de contestar las once preguntas planteadas para la resolución de la práctica 1 de la asignatura Tipología y ciclo de vida de los datos.

---

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

En esta práctica se ha decidido de estudiar la plataforma de inmobiliario "TECNOCASA" de España. En dicha plataforma, existe dos tipos de usuarios: el inmobiliario y los compradores, el inmobiliario es quien publica información sobre la venta de los inmuebles, ellos publican los datos de los inmuebles con los precios fijados; y los compradores son los que buscan la información publicada en dicha plataforma para encontrar el inmueble que ellos tienen más interés de comprar.

Como hay tanta información en la plataforma (sobre todo en las comunidades grandes como Madrid), se complica mucho a los compradores a la hora de tomar la decisión, no son capaces de entender si el precio es demasiado alto o bajo para el inmueble de su interés, y la dificultad de encontrar un inmueble con el presupuesto y cumple las expectativas del comprador es bastante alta. Por estos motivos sale el proyecto este, con el fin de poder analizar y comparar los inmuebles de la Comunidad de Madrid, ayudar a los compradores a encontrar el inmueble que les adapta mejor en dicha comunidad.

---

2. Título. Definir un título que sea descriptivo para el dataset.

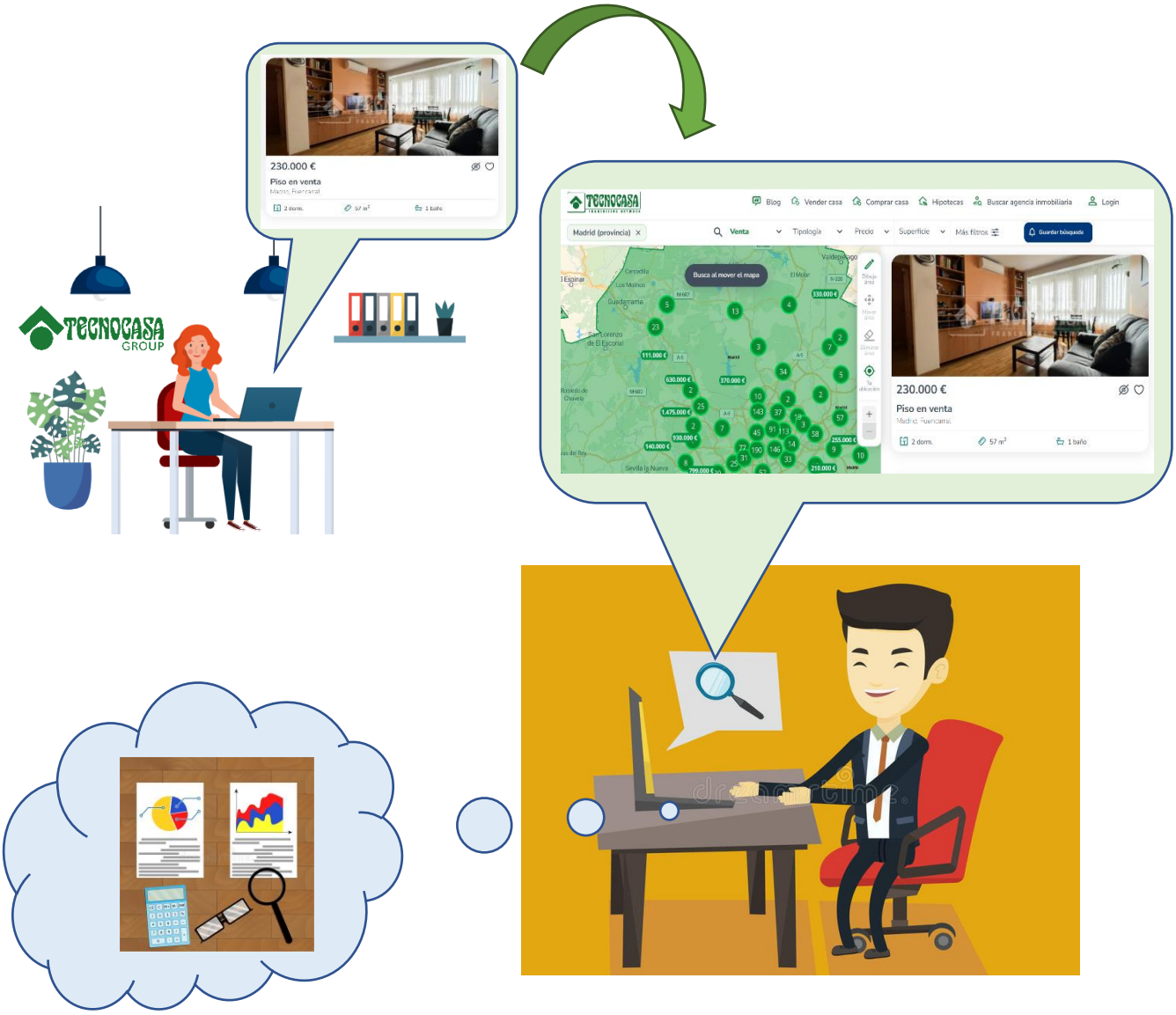
Análisis y comparación de los inmuebles de la Comunidad de Madrid a través de la plataforma TECNOCASA.

---

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Como el objetivo es de comparar y analizar los inmuebles de la Comunidad de Madrid con el fin de ayudar a los compradores a tomar la decisión, se ha extraído principalmente los datos relevantes de los inmuebles tales como el precio, la superficie, la dirección ubicada, el número de dormitorios y de baños y el título del inmueble que nos sirve de referencia. Los cuales son los que más influyen a la hora de comprar un inmueble. De esta manera, los compradores pueden tener una vista más clara de los inmuebles a través de comparar, como ejemplo, los inmuebles de una misma zona o comparar los inmuebles del mismo precio de distintas zonas para poder encontrar el mejor que adapte a sus necesidades.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido. Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

El dataset incluye todos los inmuebles de la Comunidad de Madrid que encuentran actualmente publicados en la plataforma. El periodo de tiempo se calcula desde la fecha de publicación hasta hoy en día, como la fecha de publicación es desconocida, no se puede saber el periodo de tiempo de los datos. Pero el inmobiliario solo mantiene publicado los inmuebles que se encuentran actualmente disponibles, se supone que el periodo de tiempo no se ha superado de un año ya que según la estadística el tiempo medio de venta de inmuebles es inferior a un año.

El dataset está compuesto por seis siguientes atributos:

- Address: un texto corto que describe la dirección donde está ubicado el inmueble.
- Baths: formado por un número y la palabra “baños”, describe el número de baños que tiene el inmueble.
- Price: corresponde al precio en euros del inmueble.
- Rooms: formado por un número y la palabra “dorm.”, describe el número de dormitorios que tiene el inmueble.
- Surface: formado por un número y la palabra “m”, describe la superficie en metros cuadrados que tiene el inmueble.
- Title: es un texto tipo String que describe el inmueble con caracteres generales, nos puede servir de referencia (como ID de cada inmueble).

Los datos son extraídos a través de web scraping en Python sobre las paginas individuales de cada inmueble. Para eso lo que hace primero es abrir la página donde está el listado de los inmuebles de la Comunidad de Madrid y de forma automática se recorre a la pagina de cada inmueble para recolectar la información de cada inmueble hasta el ultimo inmueble del listado. Para poder recolectar información de las paginas individuales de los inmuebles se tuvo que utilizar Selenium porque son datos dinámicos.

6. Propietario. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario del conjunto de datos es la empresa TECNOCASA, los datos vienen de su plataforma la cual donde se publica la información de los inmuebles para que los usuarios que acceden a la plataforma pueden visualizar estos datos.

Análisis de Tecnocasa en concreto no se la encuentra por internet, pero casos similares sí que existen varios:

https://www.youtube.com/watch?v=1gQyRuIpgUU

https://github.com/David-Carrasco/Scrapy-Idealista

---

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Es interesante este conjunto de datos porque puede ayudar a los compradores a encontrar el inmueble que les interesa más comparando con el resto de los inmuebles que les puede servir de referencia a la hora de hacer contra ofertas o evitar los inmuebles que tienen precios relativamente altos según las características que tienen.

Pretende contestar preguntas como ¿con el presupuesto que tiene qué inmuebles puede comprar?, ¿el inmueble es caro o es barato según sus características?, ¿existe algún inmueble que cumple todas las expectativas para ese comprador con un precio competitivo? El objetivo es ayudar a los compradores a tomar la decisión.

---

8. Licencia. Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección.

Released Under CC BY-SA 4.0 License.

Se ha seleccionado esta licencia CC (Creative Commons) porque permite a los usuarios (licenciatarios) usar obras protegidas por derecho de autor sin solicitar el permiso del autor de la obra. Y el tipo BY-SA porque de esta forma el beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y cite la obra de la forma especificada por el autor o el licenciante y también de distribuir obras derivadas bajo una licencia idéntica a la licencia que regula la obra original. Se considera como licencia libre.

---

9. Código. Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

- El código deberá ubicarse en la carpeta /source del repositorio.

- Se deben indicar las librerías y versiones utilizadas. P. ej., en Python pueden obtenerse mediante el comando

pip3 freeze > requirements.txt

- En el documento PDF se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo las habéis resuelto.

```
# -*- coding: utf-8 -*-


import scrapy
from scrapy_selenium import SeleniumRequest
from source.items import Tecnocasa



class TecnoCasaSpider(scrapy.Spider):
    name = 'source'


    def start_requests(self):
        url = 'https://www.tecnocasa.es/venta/inmuebles/comunidad-de-madrid/madrid.html'
        yield SeleniumRequest(url=url, callback=self.parse)


    def parse(self, response):
        # XPATH selector to get each book url
        for link in response.xpath('//*[@class="estate-card"]/a/@href').extract():
            yield SeleniumRequest(url=link, callback=self.parse_house)
        next_page = response.xpath('//*[@class="pagination"]//a[.=">"]/@href').extract()

        if next_page:
            yield SeleniumRequest(url=next_page[0], callback=self.parse)
```

```
    def parse_house(self, response):

        # Function to scrap each field of data

        title     =     response.xpath("//*/h1[@class='estate-
title']/text()").extract_first()

        address     =     response.xpath("//*/h2[@class='estate-
subtitle']/text()").extract_first()

        price     =     response.xpath("//*/span[@class='current-
price']/text()").extract_first()

        surface     =     response.xpath("//*[@class='estate-card-
data-element                                    estate-card-
surface']/span/text()").extract_first()

        rooms = response.xpath("//*[@class='estate-card-data-
element estate-card-rooms']/span/text()").extract_first()

        baths = response.xpath("//*[@class='estate-card-data-
element estate-card-bathrooms']/span/text()").extract_first()


        tecnocasa = Tecnocasa()

        tecnocasa['Title'] = title.replace('\n', "").strip()

        tecnocasa['Address']      =      address.replace('\n',
"").replace(",", "-").strip()

        tecnocasa['Price'] = price.replace('\n', "").strip()

        tecnocasa['Surface']      =      surface.replace('\n',
"").strip()

        tecnocasa['Rooms'] = rooms.replace('\n', "").strip()

        tecnocasa['Baths'] = baths.replace('\n', "").strip()


        yield tecnocasa
```

La principal dificultad de este proyecto se encuentra en el tipo de los datos que se intenta extraer, la información relevante de los inmuebles está guardada como datos dinámicos basados en Javascript, por lo que hay que utilizar Selenium cada vez que se recorre la página individual de los inmuebles lo que causa un mayor tiempo de scraping.

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta /dataset del repositorio.

Enlace: https://zenodo.org/record/7369857#.Y4PzxnbMI2w

11. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

Contribuciones:

Solo hay una única integrante (100% de contribuciones).