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

En este proyecto hemos decidido recolectar información de la página web de venta y alquiler de viviendas "Idealista". Este portal es utilizado tanto por los usuarios (ya sean particulares o entidades inmobiliarias) interesados en poner en venta/alquiler una vivienda, como por los que buscan adquirirla.

Si nos centramos en la parte que busca comprar o alquilar, este portal se ha convertido en una de las principales fuentes de información a la hora de buscar una nueva vivienda. Es muy común que al menos la primera búsqueda de vivienda empiece en los portales web, donde Idealista es uno de los más populares. Se puede utilizar por ejemplo para tener una idea general de los precios habituales en una ciudad o distrito concretos.

Desde el punto de vista del particular que vende o alquila la vivienda, supone también mayor comodidad el hecho de poder realizar todos los trámites "online", sin necesidad de presenciarse en ningún lugar. En ese sentido, es habitual que las inmobiliarias presenciales busquen en Idealista anuncios de viviendas pertenecientes a particulares para llegar a un acuerdo con ellos y hacer de intermediarios en la transacción, buscando recibir beneficio de este modo.

En este trabajo buscamos recopilar los datos básicos acerca de los anuncios de alquiler de las viviendas de una ciudad. Para nuestro caso hemos elegido la ciudad de Alicante, pudiendo cambiar la ciudad objetivo simplemente modificando la URL principal en nuestro código.

Con este proyecto buscamos que tanto las personas interesadas en poner un anuncio de una vivienda, como los interesados en buscar adquirir una de ellas, accedan rápidamente y de un solo vistazo a los datos básicos de la ciudad que deseen, pudiendo así tener una idea general de los precios que deberán abordar (en caso de ser la parte que compra), o comenzando a valorar el rango de precios en el que se deberán mover al poner su anuncio, en función de las características de la vivienda (en caso de ser la parte que vende).

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

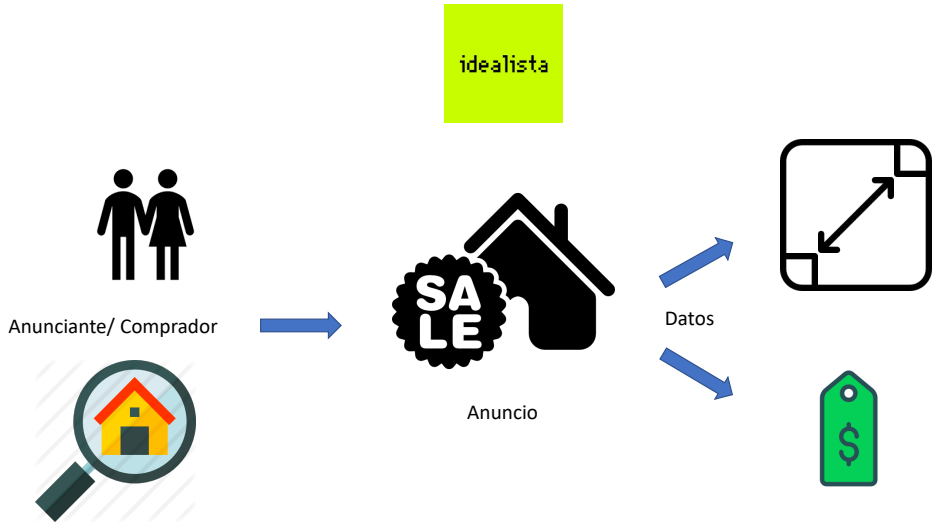
Anuncios de alquiler de la ciudad de Alicante en el portal Idealista: datos básicos.

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 titulo elegido).

Como ya hemos indicado, el dataset busca recopilar los datos básicos de todos los anuncios de alquileres en Alicante que hay actualmente en el portal Idealista. En total hemos localizado 1154 anuncios (a día 07/04/2020), de los cuales hemos extraído los

datos de precio mensual del alquiler (en euros), tamaño (en m^2) y número de habitaciones. Además de estos campos, hemos calculado el campo "Precio por metro cuadrado", muy útil siempre en este tipo de búsquedas, a partir de los valores extraídos desde la web. Existen anuncios que vienen sin el número de habitaciones informados en la vista general; mantendremos el valor nulo para esos casos. Respecto a los campos precio y tamaño, los valores venían informados con unidad, pero hemos decidido preprocesarlos y dejarlos únicamente con el valor numérico para facilitar su tratamiento por ejemplo para el cálculo del precio por metro cuadrado.

4 – Representación grafica. Presentar una imagen o esquema que identifique el dataset visualmente.



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

En este dataset se incluyen los datos básicos de los 1154 anuncios de viviendas en alquiler en Alicante que se encuentran en Idealista. Los campos son:

- Precio: Número de euros mensuales que cuesta alquilar la vivienda.
- Tamaño: Número de metros cuadrados de la vivienda.
- Habitaciones: Número de habitaciones de la vivienda.
- Euros/m^2: Precio por metro cuadrado (en euros) de la vivienda.

Los 3 primeros campos se han extraído haciendo uso de las técnicas de web scraping desde Python. En primer lugar hemos accedido a la primera página de anuncios de viviendas en alquiler de Alicante, y a partir del número total de anuncios y del número de anuncios por página (30), calculamos el número de páginas que debemos recorrer. Dentro de cada página, extraemos los valores que nos interesa de cada anuncio.

Después de preprocesar los valores de algunos campos para quitar la unidad de medida, y de calcular el cuarto campo "Euros/m^2", extraemos el dataset en formato csv.

El período de tiempo es indefinido, puesto que no hemos puesto ningún filtro: el scraper recoge todos los anuncios existentes sin importar la fecha de publicación.

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

El propietario del conjunto de datos es la propia compañía Idealista, puesto que no hemos utilizado ninguna API externa perteneciente a otra entidad para acceder a los datos; directamente los hemos obtenido de la web oficial.

No tenemos constancia de la existencia de otros análisis realizados en base a los datos extraídos del portal de Idealista. En caso de existir, no han sido consultados para la realización de este ejercicio por lo que no es necesario mencionarlos.

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

La principal razón del uso de datos relacionados con viviendas en alquiler es debido a que en la actualidad existe un interés general por el comportamiento del sector por parte de los consumidores. Dicho en otras palabras, es un tema que afecta a gran parte de la población debido al encarecimiento de la vivienda en los últimos años y al difícil acceso del mismo, sobre todo en grandes poblaciones donde el precio es demasiado elevado.

Por estas razones hemos decidido crear un dataset que recoja los datos básicos de viviendas en alquiler y así poder contestar preguntas como cual es el precio medio del metro cuadrado o la evolución del mismo realizando ejecuciones periódicas y viendo la variación del precio en distintos momentos en una zona específica.

8- 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***

Debido a sus características, una posible licencia sería la CC BY-SA 4.0 debido a que dicha licencia refleja reconocer al autor original del conjunto de datos y su posterior uso puede realizarse bajo la misma licencia, apareciendo así la autoría del mismo aunque dicho trabajo varíe o implemente nuevas funcionalidades y resultados. Por otro lado, también permite el uso comercial de los datos generados, aumentando así el interés de explotación de los mismos por parte de terceros.

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

El código usado en esta practica lo podemos encontrar en: https://github.com/svillalc/WebScrapingUoc/blob/master/scraper/scraper.py, no obstante adjuntamos el código implementado realizado en Python.

```
# -*- coding: cp1252 -*-
from bs4 import BeautifulSoup
import requests
import pandas as pd
import math
import time

# Headers para que idealista no bloquee el scraper
headers = {
"Accept":
"text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9",
"Accept-Encoding": "gzip, deflate",
"Accept-Language": "es-ES,es;q=0.9",
"Cache-Control": "no-cache",
"dnt": "1",
"Pragma": "no-cache",
"Upgrade-Insecure-Requests": "1",
"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Safari/537.36"
}

# Creamos el objeto BeautifulSoup
req = requests.get("https://www.idealista.com/alquiler-viviendas/alicante-alacant-alicante/", headers=headers)
soup = BeautifulSoup(req.text, "html.parser")

# Numero de viviendas
required0 = soup.find_all("h1")

nviviendas = []

for i in required0:
nviviendas.append(i.get_text())
numero_viviendas = int(nviviendas[0].split()[0].replace(".", ""))
print(numero_viviendas)

# Numero de paginas a partir del numero total de viviendas
# Idealista muestra 30 viviendas por pagina
numero_paginas = math.ceil(numero_viviendas / 30)

# Inicializamos las columnas del dataset
precios = []
habitaciones = []
```

```
metros = []

# Recogemos la informacion de cada pagina.
# Introducimos un sleep para no saturar el servidor
for i in range(numero_paginas):
req2 = requests.get(
'https://www.idealista.com/alquiler-viviendas/alicante-alacant-alicante/pagina-' + str(i + 1) + '.htm',
headers=headers)
print(i + 1)
soup2 = BeautifulSoup(req2.text, "html.parser")
anuncios2 = soup2.find_all("div", "item-info-container")

# Listas temporales
precios2 = []
habitaciones2 = []
metros2 = []

# Lo recorremos asi para limpiar los datos erroneos antes de incluirlos en nuestras listas
for anuncio2 in anuncios2:
precios2.append(anuncio2.find("span", "item-price h2-simulated").text)
numhabs = anuncio2.find_all("span", "item-detail")[0].text
nummetros = anuncio2.find_all("span", "item-detail")[1].text

if "hab" not in numhabs:
if "m²" not in numhabs:
nummetros, numhabs = None, None
else:
nummetros = numhabs
numhabs = None
elif "m²" not in nummetros:
nummetros = None
habitaciones2.append(numhabs)
metros2.append(nummetros)

# Anyadimos las listas temporales a nuestras listas globales
precios = precios + precios2
habitaciones = habitaciones + habitaciones2
metros = metros + metros2

# Introducimos el sleep
time.sleep(10)


# Formamos el dataset
data = {"Precio (Euros/Mes)": precios,
"Habitaciones": habitaciones,
"Tamanyo (m2)": metros}
df = pd.DataFrame(data)

# Le damos formato a las columnas y calculamos una nueva (precio por metro cuadrado)
df["Precio (Euros/Mes)"] = list(map(lambda x: x.replace(".", "").split("€")[0], df["Precio (Euros/Mes)"]))
df["Tamanyo (m2)"] = list(map(lambda x: x.split()[0], df["Tamanyo (m2)"]))
df["Euros/m2"] = round(df["Precio (Euros/Mes)"].astype(int) / df["Tamanyo (m2)"].astype(int), 2)

# Exportamos el dataset
df.to_csv("dataset.csv")
```

10- Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

El dataset obtenido se encuentra en la ruta https://github.com/svillalc/WebScrapingUoc/blob/master/scraper/dataset.csv

En dicho dataset se recoge un numero variable de registros, debido a que la oferta de viviendas en alquiler en la zona especificada es variante. Por cada vivienda registrada se presentan 4 atributos.

- **Precio**: Precio mensual en euros asignado a la vivienda en concreto.
- **Habitaciones**: Número de habitaciones que presenta la vivienda.
- **Tamanyo:** Tamaño de la vivienda en metros cuadrados.
- **Euros/m2**: Precio del metro cuadrado para la vivienda especificada.