

Práctica 1 [PRA 1] - Web Scraping

Marcos Iturbe Marengo

1. Contexto

El alto costo de los alquileres es un problema que afecta a muchas personas y comunidades. Los inquilinos a menudo tienen dificultades para encontrar un lugar adecuado para vivir a precios razonables, lo que puede tener un impacto significativo en su calidad de vida. En este contexto, el web scraping puede ser una herramienta valiosa para abordar esta situación. La [Nueva Ley de la Vivienda](#) en España, parece ser la medida que calme esta situación. Con herramientas como la que proponemos a continuación se pretende poder extraer información sobre los precios de los alquileres y poder monitorizarlos.

2. Título

El título para este *dataset* es “**AlquiESP(v0.04/2023)** - Pisos de alquiler en España disponibles en Pisos.com”

La idea es ir agregando versiones para poder monitorizar la evolución de los precios.

3. Descripción del dataset

El *dataset* contiene la información de todos los alquileres de España disponibles en www.pisos.com. Para ello se han recorrido todas las provincias disponibles de la plataforma y se ha obtenido todos los anuncios de cada provincia.

El dataset tiene 8 columnas y 18870 registros

Toda esta información se ha recolectado utilizando técnicas de *Web Scraping*.

4. Representación gráfica

A continuación mostramos un esquema del *dataset*:

	titulo	url	zona	dirección
0	Piso en calle Ciudad de Alicante, 6	https://www.pisos.com/alquilar/piso-el_playazo...	Alquilar en Almería (382)	El Playazo (Vera)
1	Piso en Aguadulce Norte	https://www.pisos.com/alquilar/piso-aguadulce_...	Alquilar en Almería (382)	Aguadulce Norte (Distrito Aguadulce. Roquetas ...
2	Piso en Aguadulce Sur	https://www.pisos.com/alquilar/piso-aguadulce_...	Alquilar en Almería (382)	Aguadulce Sur (Distrito Aguadulce. Roquetas de...
3	Piso en calle Azorín, 26	https://www.pisos.com/alquilar/piso-oliveros_a...	Alquilar en Almería (382)	Oliveros-Altamira-Barrio Alto (Almería Capital)
4	Piso en calle del Galeón	https://www.pisos.com/alquilar/piso-el_ejido_a...	Alquilar en Almería (382)	Almerimar (El Ejido)
5	Piso en calle Ciudad de Alicante, 1	https://www.pisos.com/alquilar/piso-el_playazo...	Alquilar en Almería (382)	El Playazo (Vera)
6	Piso en calle de El Greco	https://www.pisos.com/alquilar/piso-barrio_de_...	Alquilar en Almería (382)	Barrio de Archilla (Vicar)
7	Piso en Los Ángeles-Cruz de Caravaca-Piedras R...	https://www.pisos.com/alquilar/piso-los_angele...	Alquilar en Almería (382)	Los Ángeles-Cruz de Caravaca-Piedras Redondas ...
8	Piso en Carretera de Ciudad de Cádiz	https://www.pisos.com/alquilar/piso-urbanizaci...	Alquilar en Almería (382)	Urbanización de Roquetas-Las Marinas (Roquetas...
9	Piso en calle Zapillo Almería	https://www.pisos.com/alquilar/piso-ciudad_jar...	Alquilar en Almería (382)	Ciudad Jardín-Tagarete-Zapillo (Almería Capital)

precio	detalles	descripcion	fecha extraccion
130 €/día	['2 habs.', '1 baño', '60 m²', 'Bajo']	Alquilo, temporada, vivienda en planta baja, c...	2023-04-25 11:42:57
800 €/mes	['3 habs.', '2 baños', '100 m²', '6ª planta']	Piso con muebles. Altura piso , piso superfici...	2023-04-25 11:42:58
750 €/mes	['2 habs.', '1 baño', '87 m²', '3ª planta']	Piso con muebles. Altura piso 3º, piso superfi...	2023-04-25 11:42:58
650 €/mes	['2 habs.', '2 baños', '80 m²', '7ª planta']	Piso en centro de la ciudad de almería frente ...	2023-04-25 11:42:58
700 €/mes	['3 habs.', '2 baños', '140 m²']	Amplio y luminoso apartamento en alquiler en a...	2023-04-25 11:42:58
A consultar	['1 hab.', '1 baño', '52 m²', '2ª planta']	Situado en la urbanización jardines de nuevo v...	2023-04-25 11:42:58
650 €/mes	['4 habs.', '1 baño', '200 m²', '1ª planta']	Alquilamos este gran piso en la localidad de v...	2023-04-25 11:42:58
700 €/mes	['80 m²']	Una vivienda casi para estrenar, de esas que t...	2023-04-25 11:42:59
1.000 €/sem	['2 habs.', '1 baño', '90 m²', '1ª planta']	Este piso se encuentra en avenida ciudad de cá...	2023-04-25 11:42:59
800 €/mes	['3 habs.', '1 baño', '80 m²']	Por favor leer bien todo el anuncio. Se alquil...	2023-04-25 11:42:59

En la siguiente imagen podemos observar de donde se ha sacado cada columna. Faltaría la “fecha extracción” que se genera automáticamente y la columna “url” que no es visible en el html renderizado pero sí se encuentra presente en el nodo html.

The screenshot shows a search results page for 'Alquiler de pisos en Vizcaya - Bizkaia'. The page has a sidebar on the left with filters for 'Tipo de alquiler' (set to 'Todos'), 'Tipo de inmueble' (set to 'Casas y pisos'), and 'Tipo de casa o piso' (with 'Todos' selected). The main content area displays two listings. Red arrows point to specific elements in the listings, labeled with data column names: 'zona' points to the search title; 'precio' points to the price '1.500 €/mes'; 'título' points to the title 'Piso en calle Fontecha y Salazar, 1'; 'dirección' points to the address 'Castaños (Distrito Uribarra, Bilbao)'; 'detalles' points to the details '3 habs. 2 baños 130 m² 6ª planta'; and 'descripción' points to the description 'Ático bilbao alquila en exclusiva este magnifico piso en pleno campo de volantin, 130 m² exterior con terraza y...'. Each listing also includes a 'Contactar' button, an 'Avisame si baja' button, and a 'WhatsApp' button.

5. Contenido

El contenido del *dataset* es todos los pisos en alquiler disponibles en España a través de la página web www.pisos.com. La información se ha obtenido el día 25 de abril. La carga comenzó a las 11:42:57 y terminó a las 13:47:42. El *dataset* incluye las siguientes columnas:

- **título** - Título del anuncio de la vivienda en alquiler
 - Ejemplo: *Piso en Carretera de Ciudad de Cádiz*
- **url** - Link al anuncio concreto
 - Ejemplo: https://www.pisos.com/alquilar/piso-urbanizacion_de_roquetas_las_marinas04740-33340257518_100500/
- **zona** - Provincia a la que pertenece el anuncio
 - Ejemplo: *Alquilar en Almería (382)*
- **dirección** - Dirección de la vivienda en cuestión
 - Ejemplo: *Urbanización de Roquetas-Las Marinas (Roquetas de Mar)*
- **precio** - Precio del alquiler de la vivienda
 - Ejemplo: *1.000 €/sem*
- **detalles** - Lista de detalles. Dependiendo de la vivienda puede no contenerlos todos pero son los siguientes:
 - habitaciones
 - baños
 - m²
 - planta
 - Ejemplo: ['2 habs.', '1 baño', '60 m²', 'Bajo']

- **descripción** - Descripción de la vivienda
 - Ejemplo: *Este piso se encuentra en avenida ciudad de cádiz, 04740, roquetas de mar, almería, situado en el distrito de roquetas de mar, en ...*
- **fecha extracción** - Fecha y hora de extracción del dato
 - Ejemplo: 2023-04-25 11:42:59

6. Propietario

Los datos son propiedad de HABITATSOFT, S.L. Esta empresa es la propietaria de la página web www.pisos.com de donde se ha sacado la información.

Redactando este apartado me he dado cuenta que ya existía una práctica de otro año sobre la misma plataforma: <https://github.com/paucampana/pisosScraper> sin embargo en esta práctica no se obtienen los datos de toda España, solo de una zona.

También se han revisado análisis sobre otras plataformas como Idealista: <https://github.com/David-Carrasco/Scrapy-Idealista>

Para la extracción de esta información no se requiere la aceptación de ninguna política de uso. Además toda la información es pública y no es necesario ningún registro. Por último, el uso de estos datos es sin fines de lucro.

7. Inspiración

La creación de este *dataset* viene inspirada por la noticia de la entrada en vigor de la Nueva Ley de la Vivienda. Este *dataset* pretende responder a las siguientes preguntas entre otras:

- ¿Cuál es la media del precio del alquiler en España?
- ¿Y por zona?
- ¿Cómo evolucionan los precios de los alquileres?

Las anteriores aproximaciones no responden a estas preguntas ya que solo se centran en una zona concreta.

8. Licencia

Se decide utilizar la siguiente licencia:



Atribución - Debes dar el crédito correspondiente, proporcionar un enlace a la licencia y señalar si se han realizado cambios. Puede hacerse de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante te respalda a ti o a tu uso.

No comercial - No puedes utilizar el material con fines comerciales.

Sin obras derivadas - Si remezclas, transformas o construyes a partir del material, no puedes distribuir el material modificado.

9. Código

El código se ha implementado en Python utilizando la librería de Selenium. Se ha utilizado el [undetected-chromedriver](#) que internamente utiliza el siguiente User-Agent:
*Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/112.0.0.0 Safari/537.36*

El código está disponible en el siguiente repositorio:

```
# Importación de librerías
from datetime import datetime
from selenium.webdriver.common.by import By
import undetected_chromedriver as uc
import pandas as pd

# Función que crea el Web driver
def set_driver():
    global driver
    driver = uc.Chrome(use_subprocess=True)

# Función para exportar un diccionario a un CSV externo
def to_csv(dictionary, file_name):
    pd.DataFrame(dictionary).to_csv(file_name, index=False)

# Función para aceptar las cookies
def aceptar_cookies():
    driver.find_element(By.CSS_SELECTOR, 'button#didomi-notice-agree-button').click()

# Función para extraer todas los links de las provincias disponibles
```

```

def extraer_provincias(url):
    driver.get(url)
    aceptar_cookies()
    provinces_div = driver.find_element(By.CLASS_NAME, 'seolinks')
    # se excluyen los subitems para no repetir viviendas
    provinces = provinces_div.find_elements(By.CSS_SELECTOR, 'a.seolinks-zones-item:not(.seolinks-zones-subitem)')
    results = {}
    for province in provinces:
        results[province.text] = {
            "url": province.get_attribute('href'),
        }
    print(results)
    return results

# Función recursiva para extraer la información de cada vivienda de los enlaces de las provincias
def extraer_apartamentos(url, zona):
    driver.get(url)
    article_nodes = driver.find_elements(By.CSS_SELECTOR, 'div.ad-preview')
    apartments=[]
    for article in article_nodes:
        apartment={}
        try:
            apartment['titulo']=article.find_element(By.CSS_SELECTOR, 'a.ad-preview__title').text
            apartment['url']=article.find_element(By.CSS_SELECTOR, 'a.ad-preview__title').get_attribute('href')
            apartment['zona']=zona
            apartment['dirección']=article.find_element(By.CSS_SELECTOR, 'p.p-sm').text
            apartment['precio'] = article.find_element(By.CSS_SELECTOR, 'span.ad-preview__price').text
            detalles_nodes= article.find_elements(By.CSS_SELECTOR, 'p.ad-preview__char.p-sm')
            detalles=[]
            for detalle in detalles_nodes:
                detalles.append(detalle.text)
            apartment['detalles']= detalles
            apartment['descripcion'] = article.find_element(By.CSS_SELECTOR, 'p.ad-preview__description').text
            apartment['fecha extraccion']=datetime.today().strftime('%Y-%m-%d %H:%M:%S')
            print(apartment)
        except:
            pass
        apartments.append(apartment)
    # Verifica si hay un botón de paginación y si lo hay lo recorre
    if driver.find_elements(By.CLASS_NAME, "pagination__next"):
        button = driver.find_element(By.CLASS_NAME, "pagination__next")
        url = button.find_element(By.TAG_NAME, "a").get_attribute("href")
        apartments += extraer_apartamentos(url, zona)
    return apartments

def main():
    set_driver()
    print(f'User agent used is :{driver.execute_script("return navigator.userAgent;")}')
    provincias =extraer_provincias("https://www.pisos.com/alquiler_viviendas/")
    result= []
    for provincia in provincias.keys():
        result+= extraer_apartamentos(provincias[provincia]['url'], provincia)
    to_csv(result,"apartamentos.csv")

```

```
if __name__ == "__main__":  
    main()
```

10. Dataset

El *dataset* ha sido publicado en Zenodo con el siguiente **DOI: 10.5281/zenodo.7863788**

Es accesible mediante el siguiente enlace:

<https://doi.org/10.5281/zenodo.7863788>

11. Vídeo

Esta disponible en el siguiente enlace:

<https://drive.google.com/file/d/1XkME7Odv3oT11SgNgQj5DMzIQzsnp4-O/view?usp=sharing>