


Prueba técnica

Desarrollo:

Para empezar, el primer paso que se realizó fue el de revisar la estructura de la página de la que se debe obtener la información, para ver si solo se tenía que obtener la información de las etiquetas, o si se tenía que entrar a cada una de las propiedades publicadas, en busca de esa información.

Para este caso, se podía obtener la información desde las pantallas de previsualización, o sea, solo se tenía que recorrer página por página del buscador, para extraer lo necesario: precio, divisa, metros cuadrados de construcción y terreno, numero de recamaras, y numero de baños.



Super destacado

1 / 14

MN 15,000,000

Allure Condos de Lujo Entrega Inmediata en Puerto ...


KM 1.5 BLVD KUKULCAN, Cancún, Quintana Roo, Zona Hotelera, ...

309 m² terreno 309 m² construidos 3 Recámaras 3 Baños

Exclusivo Y maravilloso desarrollo con vista al mar dentro de puerto cancún, con una inigualable calidad en su construcción Y acabados. Amenidades, club de playa Y seguridad máxima para vivir tranquilo con la familia. Listos para vivir. Desde \$9, 900, 000 pesos. con amenidades Y acabados de lujo, club de playa, club de golf, marina de puerto cancún, spa,...

Publicado hace 10 días

Contactar



Super destacado

1 / 34

USD 2,700,000

Penthouse en Venta, Bay View Grand, Zona Hotelera...


Bay View Grand, Zona Hotelera, Cancún

550 m² terreno 550 m² construidos 4 Recámaras 4 Baños

Penthouse a la Venta en Bay View Grand, Zona Hotelera, Cancún, en una de los mejores condominios de la Zona Hotelera en donde gozaras de paz y tranquilidad así como de acceso exclusivo a la playa y hermosas vistas al Mar Caribe. El penthouse completamente amueblado y remodelado cuenta con Cuatro Recámaras, Cuatro y medio Baños, Cocina...

Publicado hace 2 días

Contactar



Super destacado

1 / 15

USD 2,500,000

Departamento en Venta en Cancun Zona Hotelera

Departamento en Venta en Cancun Zona Hotelera, Zona Hotelera...

540 m² terreno 4 Recámaras 4 Baños

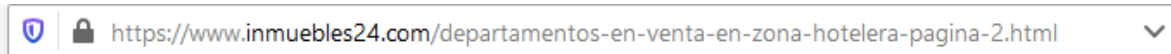
Departamento en venta en Cancun Zona hoteleraCondominio Portofino, Zona hotelera4 recamaras, 4 baños completos, sala, comedor, cocina, cuarto de servicio con baño completo, cuarto de lavado. Areas comunes: alberca, gimnasio, spa, cancha de tenis y area de niños con juegos. Magnifica inversion!! B-jpr815

Publicado hace 17 días

Contactar

Adicional a la información que se menciona anteriormente, para guardar en la base de datos, se puede obtener el enlace de cada propiedad, el título, la ubicación, la descripción, número de contacto, e incluso, cuanto tiempo tiene la propiedad de ser publicada.

Después se tuvo que observar el comportamiento de cómo se iría desplazando al cambiar de página, ver si este comportamiento se podría implementar haciendo un incremento en el numero de páginas, o si existía algún patrón que seguir para realizar el “barrido horizontal”, para el caso de la página, se observó el siguiente patrón, el cual es algo sencillo de implementar.



Otro punto importante a considerar, es el saber hasta cuantas páginas o propiedades, había que extraer, por lo que analizando un poco más la página, se encontró este indicador:

858 Departamentos en venta en Zona Hotelera, Cancún

Por lo que un paso que también se debe seguir es el de extraer ese número, para saber el límite de cuantas propiedades será extraída la información.

Se seleccionó la herramienta “*requests*”, el cual es un paquete de Python, el cual nos permite hacer la automatización de la extracción de la información, sumado a *BeautifulSoup*, para realizar la búsqueda de los elementos necesarios dentro de la página. Se hizo esta selección ya que es el paquete de Python con el que más me encuentro familiarizado, además de que permite observar si la petición que se intenta realizar procede, o no. Y *BeautifulSoup*, ya que había que hacer un “parser” para facilitar la búsqueda de los elementos.

Después de hacer el análisis de la página, se procedió a averiguar si esta nos iba a permitir el hacer uso de un *robot* para hacer peticiones. Esta pequeña prueba, nos arroja un código 200, lo cual nos indica que se podrá extraer información de forma satisfactoria.

```
https://www.inmuebles24.com/departamentos-en-venta-en-zona-hotelera.html
<Response [200]>
```

Otro punto que se consideró, es el precio del dólar para hacer la conversión a pesos mexicanos, para esto, también se automatizó la extracción de este valor, para que se actualice día con día. Para la fecha de las pruebas que se realizaron, el resultado es el siguiente:

Precio del dolar el dia de hoy \$18.54

Regresando a la página, se inspeccionó para revisar la estructura, y poder automatizar la extracción de la información, para todos los elementos de previsualización, la estructura era similar, por lo que primero, se obtuvieron todas estas, y después, de cada una, se obtenían todos los elementos necesarios.

- Del precio, se obtuvo el texto que se muestra, incluida la divisa, y se implementó un método para corroborar si el precio estaba en dólares, o pesos mexicanos. Cuando se detectara que estaba en dólares, se realiza la conversión haciendo uso del precio del dólar que se extrajo al inicio. Se realizó de esta manera, ya que considero que es la más práctica y así se puede asegurar que todos los precios se encuentren en la misma divisa.
- El título y la ubicación estaban dentro del mismo apartado, por lo que se podían obtener casi a la par.
- Posiblemente el más complejo, fue la lista de las características, ya que no siempre se encuentran todas, por lo que se implementó un método para poder extraer esta información estén o no todos los elementos. Además se implementó un método que “limpia” el texto, ya que este se extraía aun con algunas etiquetas HTML.
- Como se mencionó anteriormente, se obtuvo información adicional, esto con el fin de almacenarlo en la base de datos, para futuras consultas, esta información es: el enlace de la publicación, la descripción de la publicación, el tiempo que lleva publicada, y el número de contacto el cual va con un enlace de WhatsApp, por lo que se tuvo que “limpiar” para solo obtener el número de contacto.

Después se hacía un incremento en el número de la página, para que se mostraran nuevos elementos, y se seguía el procedimiento anterior.

Antes de hacer un incremento, o un cambio de página, se revisaba si el número de elementos extraídos era menor o igual que el número máximo de elementos, esto para evitar que arrojara una excepción y el programa se detuviera.

Finalmente, los elementos son guardados en un archivo “csv”, simulando la base de datos. Para este caso en particular, al no ser un numero muy grande de elementos, estos podían ser almacenados

hasta el final del proceso, sin embargo, si el número de elementos hubiese sido mayor, posiblemente, una mejor propuesta sería guardar los elementos extraídos, después de cierta cantidad, para evitar la pérdida de información, o tener que repetir todo el proceso.

Preguntas:

¿Qué ideas explorarías si el sitio que estás extrayendo bloquea la herramienta que estás usando?

- La primera solución podía ser intentar con otra herramienta, tal como scrapy, selenium, u otra que se pueda implementar la misma tarea. Si aun no se puede, podría ser que la página bloqueó nuestra IP, por lo que se podría hacer uso de algún proxy, o tratar de enmascarar o cambiar la IP del dispositivo con el que se pretende realizar la extracción
Si esto sigue sin dar frutos, como última herramienta que propondría, *pyautogui*, el cual es un paquete de Python que puede simular un comportamiento humano, y se puede extraer información de forma automatizada, sin embargo, esto será más lento que haciendo uso de las herramientas anteriormente mencionadas.

¿Cómo podrías monitorear el flujo de la información desde el sitio web que estás extrayendo hasta la base de datos?

- Esto podría realizarse haciendo una consulta a la base de datos, contando los elementos que hay guardados comparándolos con la cuenta de elementos de los que se ha extraído la información y este número idealmente debería cuadrar, de no ser así, puede que exista algún problema en el trayecto de la información.

¿De qué forma podrías modificar tu herramienta para ampliar la extracción a todo el sitio? Y ¿para extraer otro sitio distinto?

- La programación del script que les envió, debería ser capaz de extraer información de toda la página inmuebles24, lo único que habría que cambiar, sería el enlace objetivo del cual se requiere extraer información, para cambiar el enlace, el cambio que se tendría que hacer es crear un método que sea exclusivamente para esto, primero, para encontrar cuantas

propiedades hay ofertadas en el enlace que se le envía, y posteriormente, modificarlo para realizar el “barrido horizontal”.

- Para extraer de otro sitio, tal vez se tengan que realizar varios cambios, ya que no todas las páginas son similares, sin embargo, al haber realizado el script con POO, considero que podría ser posible implementar un método que sea para agregar pasos de extracción, al cual se le hagan llegar como parámetros los elementos se quiere extraer, y que el script realice esta operación. Aunque como lo menciono, esto sería un área de oportunidad que puede llevar un poco más de tiempo en desarrollar.

Análisis:

Posterior a la extracción y almacenamiento de los datos, se procedió a realizar el análisis de estos.

Primero, se cargó el csv con la información almacenada, con solo las columnas que nos interesan, posteriormente, se revisó si había que hacer un poco de pre - procesamiento de los datos, y resultó que sí, ya que había algunas propiedades que tenían un precio de \$1 peso, y otras que salían demasiado del rango de los precios:

	construidos_m2	precio_pesos
count	751.000000	8.760000e+02
mean	309.796272	1.570933e+07
std	1380.890067	2.067650e+07
min	1.000000	1.000000e+00
25%	160.000000	6.889950e+06
50%	212.000000	1.205100e+07
75%	290.000000	1.763454e+07
max	34190.000000	5.047000e+08

Se optó por dejar los valores que se encontraban entre el percentil 5% y el percentil 95%, esto para “normalizar” un poco el set de datos. Después se hizo la operación para obtener el precio por metro cuadrado de cada propiedad.

	construidos_m2	precio_pesos	precio_m2
0	NaN	9900000.0	NaN
1	NaN	12051000.0	NaN
2	214.0	9000000.0	42056.07
3	96.0	4740000.0	49375.00
4	200.0	11197000.0	55985.00

Finalmente se obtuvieron los resultados requeridos:

mediana precios : \$	12,001,804.50
promedio precios : \$	13,550,759.00
promedio precio/m2 : \$	55,002.00
promedio precio/m2 : \$	61,246.79