Tipología y ciclo de vida de los datos.

Carlos Chamorro

Práctica 1: Web Scraping

1.Contexto

En este proyecto se confecciona un dataset con información sobre las viviendas en alquiler en Barcelona disponibles en la Web idealista. En esre dataset se incluye la información más relevante de cada vivienda.

2. Título

El nombre del dataset es:

Bcn_alquiler_idealista

3. Descripción del dataset

En este dataset, obtenemos la información principal que se necesita para analizar una vivienda, a saber, un título con el barrio o la calle de la vivienda, precio, metros cuadrados, número de habitaciones, metros cuadrados, planta, ascensor(si tiene o no) y la descripción que el arrendatario ha escrito. Además se le añade una columna llamada "momento_de_descarga" en la que se indica cuando se ha llevado a cabo el scraping. Dicha columna se añade por exigencia de la práctica.

4.Representación gráfica del dataset

El dataset queda representado por la siguiente tabla:

Columna	**Tipo**	**Ejemplo**	**Comentario**
propiedad	String	##### ########## ######	Titulo de la vivenda, suele contener situación georgráfica de la vivienda
precio	Int	3000	
Número de habitaciones	String	3	Puede contener "no data"

m2	Int	94	
planta	string	4	
ascensor	Int	1	1 para si, 0 para no
descrpición	string	#### ####+#####	
“momento_de_descarga”	String	10/04/2022 13:15:13	

5.Contenido

Los datos contenidos en la columnas “propiedad” […] “descripción” se obtienen de un procesado sistemático del texto en HTML de cada una de las páginas (desde la primera hasta la última que permite la web) de la URL 'https://www.idealista.com/alquiler-viviendas/barcelona-barcelona/'. Como vemos la URL inicial ya contempla la provincia y ciudad de Barcelona. Esto podría ser fácilmente automatizado para generar datasets de la ciudad que se desee siempre y cuando Idealista.com guarde datos de la misma. La columna “Momento_de_carga” se obtiene de la librería Datetime.Datetime de Python y se ejecuta en el momento de introducir una fila en el dataset.

6. Agradecimientos

Para realizar este ejercicio se ha pedido permiso a Idealista.com a través de su API https://developers.idealista.com/access-request. Se ha dejado claro que las intenciones son para un proyecto universitario. Sin embargo, no se ha obtenido respuesta hasta el momento por lo que se ha optado por realizar el scraping con el user agent por defecto del paquete resquests. No se ha utilizado un user-agent dinámico puesto que para este proyecto no se realizarán técnicas de dudosa moral, por lo que se realizará el scraping hasta que la página lo permita. La información clave del dataset se ha ocultado sustituyendo los caracteres letras por “#” para no exponer información propiedad de idealista.

7. Inspiración

Las oportunidades laborales, cada vez se concentran más en un puñado de grandes ciudades, provocando una altísima demanda de vivienda en ciudades como Madrid o Barcelona en España, Berlín o Munich en Alemania o Nueva York o Miami en los Estados Unidos de América, por citar algunos ejemplos. En muchos casos, las ciudades no han sabido manejar esta creciente demanda o, simplemente, no han podido, dadas las características geográficas de algunas de estas ciudades. Esto crea un mercado hiperdinámico de

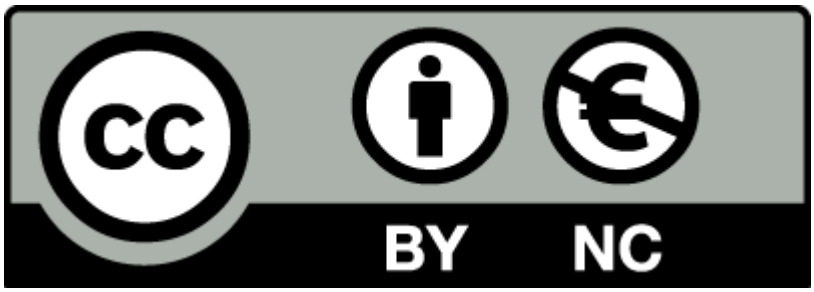
apartamentos y pisos compartidos, donde la rapidez a la hora de solicitar resulta crucial. En el caso de la web Alemana Wg-gesucht.de se considera “tarde” solicitar una habitación pasados los primeros 20 minutos desde que el post se efectúa. Esto obliga a muchos estudiantes y trabajadores a dedicar mucho tiempo, no solo a escribir a los posibles caseros, sino a observar la pantalla del ordenador o del móvil refrescando la página a la espera de una buena oportunidad.

Es mi caso, puesto que trabajo actualmente en Berlín, Alemania, habiendo vivido previamente 4 años en Barcelona. Por eso, una de las primeras y más evidentes funciones del Web Scraping que me vienen a la cabeza es la de poder gestionar, en tiempo real y de forma automatizada, esta tarea.

Debido a la naturaleza en equipo de esta práctica he considerado que sería más oportuno trabajar con una web que no estuviera en alemán, por lo que finalmente, he escogido la Web Idealista.com para obtener un dataset de todas las viviendas en alquiler para la ciudad de Barcelona, la cual, no se queda atrás en lo que a dinamismo del mercado se refiere. Se ha escogido, por tanto, la web idealista.com para dicha tarea.

El objetivo de esta tarea será el de obtener dicho dataset para la ciudad mencionada. Este mismo proyecto invita a ser expandido con diferentes funcionalidades como la de seleccionar la ciudad que será analizada, el filtrado de datos para realizar la búsqueda o, incluso, la automatización de una alarma que avisara de posibles oportunidades, como ejemplos de lo que un proyecto como este podría llegar a convertirse.

8. Licencia



9. Código.

El código queda a disposición en el repositorio de github.

10. Dataset

El dataset queda queda a disposición en el repositorio de github.

11. Video

El video queda a disposición a través del enlace de google drive.