Proyecto Web_Scraping y proceso de obtención del dataset hometogo.csv

Elisabeth Anna López Simpson
Jesús Antonio Blay Tamarit

https://github.com/ElisabethUoc/Web_Scraping

En este proyecto, además de la wiki del repositorio, se incluyen los siguientes archivos:

- README.md (pequeña introducción al proyecto)
- main.py (script para la obtención del dataset)
- hometogo.csv (dataset obtenido con la ejecución de *main.py*)
- proyecto_EALS_JABT.pdf (detalle del proyecto y proceso de obtención de nuestro dataset)
- LICENSE

1. Contexto.

Este proyecto de scraping tiene como finalidad conseguir un dataset con el objetivo de poder filtrar y resumir las distintas alternativas para encontrar el mejor destino de alquiler en un lugar y fechas pre-establecidas por el usuario, con la información obtenida del portal hometogo.com. El motivo de elegir dicha web es que se trata de un agregador de alquileres vacacionales, que aglutina la información de la mayoría de fuentes y portales de este tipo, como Booking, Airbnb o Tripadvisor, entre muchas otras.

Sin embargo, dada la finalidad del proyecto, nuestro principal objetivo era que este proyecto fuese reutilizable (asumiendo que toda modificación de la web hometogo.com requerirá, en principio, una revisión y actualización del script main.py), para que nosotros mismos y todos aquellos interesados, puedan adaptarlo a otras fechas y características, con el fin de poder obtener el dataset que se adapte a la búsqueda y a las circunstancias personales de cada usuario.

2. Título del dataset: hometogo.csv

Hemos optado por conservar como nombre del dataset el mismo de la web desde la que hemos obtenido la información, ya que resume muy bien el objetivo que perseguíamos con este proyecto (encontrar una casa a la que ir de vacaciones), pero especialmente por tratarse de un dataset dinámico que no necesitaremos conservar en el tiempo, sino que será útil para un momento y circunstancias concretas. Por tanto, se deberá ejecutar el script cada vez que se pretenda obtener un dataset

actualizado y adaptado a las preferencias de cada momento y usuario (destino, época, precio máximo, etc).

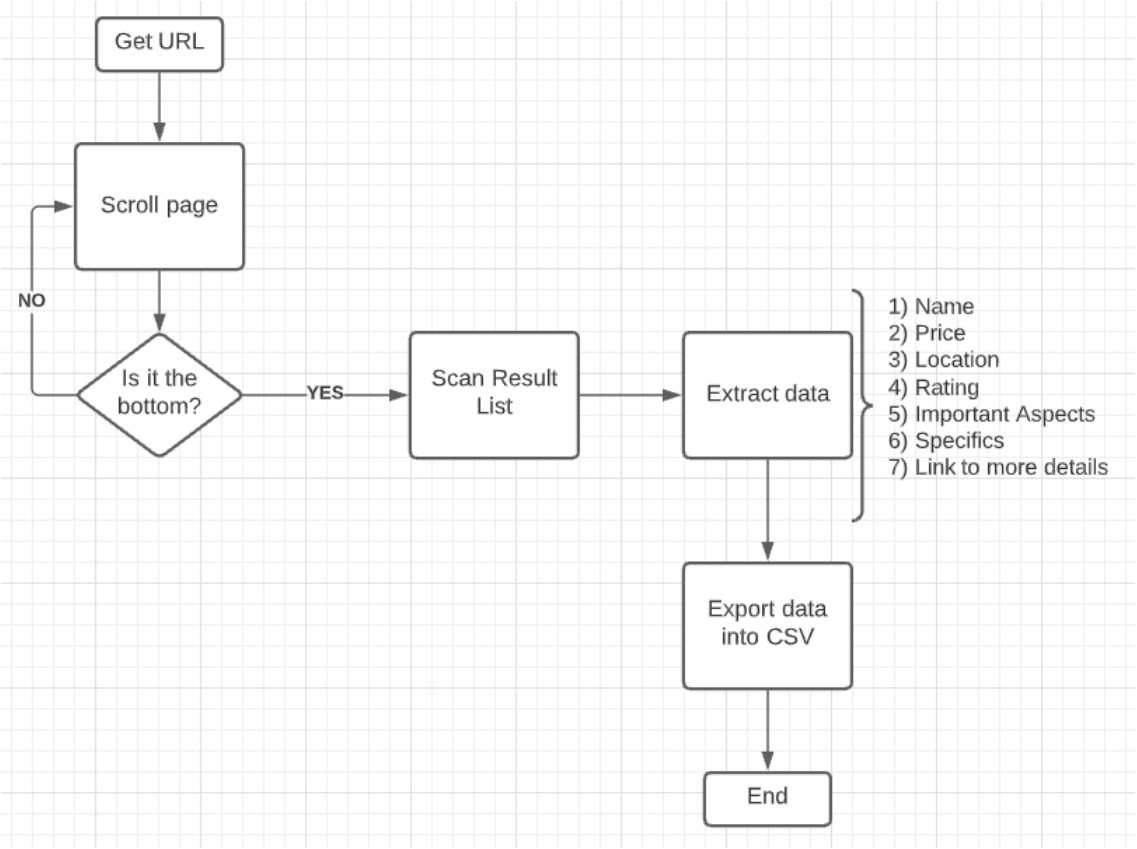
3. Descripción del dataset.

Como ya hemos explicado, hemos creado un dataset que nos proporciona la información que consideramos relevante a la hora de elegir un alquiler vacacional.
A través de la URL, identificamos dónde queremos ir de vacaciones (en nuestro caso, Suiza), las fechas para las que buscamos el alojamiento (31/05 - 06/06) y filtrando por un precio (175 euros la noche)
https://www.hometogo.com/search/5460aea910151?adults=2&arrival=2021-05-31&duration=6&maxPricePerNight=175EUR

Una vez tenemos definidos los parámetros que queremos para la búsqueda, extraemos la información que consideramos relevante para la toma de la decisión de dónde alojarnos en nuestras vacaciones. En el punto 5 explicaremos cuál es la información que nos interesa extraer.

4. Flujo de ejecución del script para la obtención del dataset.

A continuación, se adjunta la representación visual de dicho flujo:



5. Contenido.

Tras realizar el *scraping* a la url especificada y valorar la información que nos interesa de cada inmueble, nos quedamos con los campos que se detallan a continuación, por ser la información que encontramos relevante para después comparar y filtrar entre todos los inmuebles disponibles que integran el dataset obtenido.

Hay que tener en cuenta qué dada nuestra finalidad, así como el tipo de portal y su contenido, el dataset obtenido contiene todos aquellos inmuebles que se encuentran disponibles en el momento de la búsqueda (ejecución del script `main.py`), pero que podrían no estar ya disponibles en un muy breve espacio de tiempo si los inmuebles son alquilados por terceros.

Los campos o atributos de interés de nuestro dataset son los siguientes:

- **name**: nombre o descripción con la que el propietario publica el inmueble.
- **price**: precio para el periodo de selección (en nuestro caso del 31/05 al 6/06).
- **location**: localización del inmueble.
- **rating**: puntuación media obtenida por el inmueble a partir de sus valoraciones
- **important_info**: se incluye la característica más destacada; el literal 'Free cancellation' en los casos en los que esta opción está disponible para el inmueble o la valoración media sobre 5 en su defecto.
- **specifications**: descripción y características generales del inmueble.
- **link**: enlace al inmueble dentro de la web de *hometogo*, conservando los criterios de búsqueda.

En la siguiente imagen podemos ver cuáles son los campos del HTML del que hemos extraído cada uno de los campos anteriores:

6. Agradecimientos.

Dado que ninguno de los dos integrantes de este grupo tenemos un perfil técnico, hemos dedicado bastante tiempo a la búsqueda de proyectos parecidos realizados sobre la web de Airbnb.

Además de eso, estuvimos revisando el proyecto realizado por un compañero de trabajo de Elisabeth, que había hecho un proyecto de web scraping para su uso personal, para buscar inmuebles para comprar en páginas como idealista.

La revisión de este proyecto, fue lo que nos inspiró para hacer un proyecto que pudiéramos utilizar en el futuro.

https://github.com/rubenkerkhofs/Airbnb-scraping
https://github.com/paupalou/housefinder

7. Inspiración.

Al empezar con el proyecto, algo que teníamos claro era que queríamos realizar un proyecto que pudiéramos utilizar en el futuro.

Como hemos comentado, no tenemos un perfil técnico y ninguno de los dos había realizado web scraping con anterioridad, por tanto, sabíamos que la ejecución de este proyecto supondría un reto para nosotros. Puesto que le íbamos a dedicar una cantidad de horas considerables, dada la falta de experiencia, queríamos que estas horas fueran a parar a algo que sea de utilidad.

Hicimos varias calls y un brainstorming antes de decidirnos por scrapear una web de alquiler vacacional.

¿Por qué elegimos el alquiler vacacional? Pues por dos motivos básicos:

El primero, que a ambos nos encanta viajar, y no es algo que hayamos podido hacer últimamente por razones obvias.

Y el segundo, también dada la situación, que parece más razonable y seguro, el alquiler vacacional que no irse a un hotel a pasar unas vacaciones.

Elegimos la web de Hometogo porque es una empresa que además de recoger alojamientos de otras webs como aribnb tiene también producto propio que están intentando promocionar y nos pareció que tenía una oferta de alojamientos más amplia y a mejores precios.

8. Licencia.

Hemos añadido en el repositorio el fichero LICENSE para especificar la licencia que aplica a todo el repositorio. Optamos por una licencia "Released Under CC BY-SA 4.0" para apoyar el uso de datos abiertos y también para animar a la eventual actualización del script **main.py** en caso de que fuese necesario debido a la actualización o modificación de la página hometogo.com.

La licencia escogida aplica igualmente al dataset incluido en este repositorio, hometogo.csv, de manera que cualquier usuario es libre de compartir - copiar y redistribuir el material en cualquier medio o formato- y de Adaptar - remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente-, bajo los siguientes términos:

Attribution - Se debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerse en cualquier forma razonable, pero no de forma tal que sugiera que el usuario o su uso tienen el apoyo del licenciante.

ShareAlike - Si se remezcla, transforma o crea a partir del material, debe distribuirse su contribución bajo la misma licencia original.

9. Código.

El código de nuestro proyecto puede encontrarse https://github.com/ElisabethUoc/Web_Scraping en "main.py".

Y el fichero csv con los datos extraídos se encuentra en el mismo repositorio en "hometogo.csv" o en Zenodo: https://zenodo.org/record/4671950#.YG7R7OgzaUk

Contribuciones	Firma
Investigación Previa	EALS, JABT
Redacción de las respuestas	EALS, JABT
Desarrollo código	EALS, JABT