PRA1: ¿Cómo podemos capturar los datos de la web?

Asignatura: M2.851 - Tipología y ciclo de vida de los datos

Aula: 1

Alumnos: Carlos Santamaría de las Heras y Alba Sanz Horcajo

Fecha: 22/11/2022

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

Se ha recopilado información de inmuebles en venta en Madrid Capital a fecha de 19/11/2022. El objetivo era generar un script que scrapease información para personas interesadas en la compra-venta de inmuebles en la capital. Para ello elegimos un portal web inmobiliario especializado en la compraventa y alquiler de viviendas en España como es Fotocasa (https://www.fotocasa.es/es/). En esta web podemos acceder de manera gratuita a multitud de ofertas de inmuebles nuevos y de segunda mano y encontramos diversos datos acerca de estos (precio, localización, metros cuadrados, ascensor, calefacción…). El único filtro que se ha utilizado en las búsquedas ha sido la localización: Madrid Capital.

2. Título. Definir un título que sea descriptivo para el dataset.

El título elegido para el dataset extraído es “Inmuebles en venta en Madrid Capital”.

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 título elegido.

El dataset está compuesto por 2169 registros y 10 atributos, todos correspondientes a inmuebles en venta en Madrid Capital. De cada inmueble hemos recopilado los metros cuadrados, su precio, la planta en la que se encuentra, el número de baños y habitaciones con las que cuenta, si dispone o no de calefacción, aire acondicionado y/o ascensor y el enlace al anuncio original de Fotocasa. Además hemos extraido el título de cada anuncio donde podemos encontrar información adicional sobre el inmueble como su tipo o su localización.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

El dataset extraído tiene la siguiente estructura:

Título	**Precio**	**M2**	**Planta**	**Baños**	**Habitaciones**	**Calefacción**	**AireAC**	**Ascensor**	**Enlace**
Título Piso 1	*Precio Piso 1*	*M2 Piso 1*	*Planta Piso 1*	*Baños Piso 1*	*Habitaciones Piso 1*	*Calefacción Piso 1*	*AireAC Piso 1*	*Ascensor Piso 1*	*Enlace Piso 1*
…	…	…	…	…	…	…	…	…	…
Título Piso n	*Precio Piso n*	*M2 Piso n*	*Planta Piso n*	*Baños Piso n*	*Habitaciones Piso n*	*Calefacción Piso n*	*AireAC Piso n*	*Ascensor Piso n*	*Enlace Piso n*

En el esquema se ha representado el dataset en formato tabla para facilitar su explicación aunque se entregará en formato .csv, es decir, representando cada registro por una fila y dentro de cada fila cada uno de los atributos separados por comas.

Como representación gráfica del proyecto hemos combinado dos imagenes de "pixabay.com" que representan la venta de pisos en Madrid:

5. Contenido. Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

El dataset contiene los inmuebles en venta en Madrid Capital anunciados en la web Fotocasa a fecha del 19 de noviembre de 2022. Los campos que incluye el dataset son 10:

- *Título*: campo tipo string con el título indicado en el anuncio. Contiene información variable como el tipo de inmueble que es (loft, piso, duplex…), los metros cuadrados o la localización del inmueble, entre otros.
- *Precios*: campo tipo string con el precio del inmueble.
- *M2*: campo tipo string con los metros cuadrados de la vivienda.

- *Planta*: campo tipo string con el nº de planta de la vivienda.
- *Baños*: campo tipo string con el número de baños de la vivienda.
- *Habitaciones*: campo tipo string con el número de habitaciones de la vivienda.
- *Calefacción*: campo tipo string indicando si dispone o no de calefacción.
- *AireAC*: campo tipo string indicando si dispone o no de aire acondicionado.
- *Ascensor*: campo tipo string indicando si dispone o no de ascensor.
- *Enlace*: campo tipo string con el enlace al anuncio.

Al guardar los datos en un archivo '.csv' algunos caracteres como letras con tildes o la letra ñ se han desajustado y cambiado por otros caracteres. Dejamos pendiente el cambio de estos caracteres para la limpieza del dataset en la PRA2.

6. Propietario. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Siguiendo unas buenas prácticas, se ha conseguido el robots.txt de la página, el cual se encuentra en: *https://www.fotocasa.es/robots.txt*

Es importante verificar si la página cuenta con este archivo, ya que es donde el propietario del sitio ha indicado las restricciones a tener en cuenta cuando se pretende rastrearlo. Estas restricciones son solo una sugerencia y nunca una obligación, por lo que en muchos casos es posible recuperar información de páginas en las que el propietario ha expresado su voluntad de

no ser rastreado. No obstante, lo más recomendable es seguir siempre las sugerencias indicadas en robots.txt, con el objetivo de reducir las posibilidades de ser bloqueados y evitar problemas legales futuros.

```
El propietario de la web es:
{
  "domain_name": [
    "WORDPRESS.COM",
    "wordpress.com"
  ],
  "registrar": "MarkMonitor, Inc.",
  "whois_server": "whois.markmonitor.com",
  "referral_url": null,
  "updated_date": "2022-01-30 09:14:20",
  "creation_date": "2000-03-03 12:13:23",
  "expiration_date": [
    "2024-03-03 12:13:23",
    "2024-03-03 00:00:00"
  ],
  "name_servers": [
    "NS1.WORDPRESS.COM",
    "NS2.WORDPRESS.COM",
    "NS3.WORDPRESS.COM",
    "NS4.WORDPRESS.COM",
    "ns2.wordpress.com",
    "ns3.wordpress.com",
    "ns4.wordpress.com",
    "ns1.wordpress.com"
  ],
  "status": [
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited",
    "serverDeleteProhibited https://icann.org/epp#serverDeleteProhibited",
    "serverTransferProhibited https://icann.org/epp#serverTransferProhibited",
    "serverUpdateProhibited https://icann.org/epp#serverUpdateProhibited",
    "clientUpdateProhibited (https://www.icann.org/epp#clientUpdateProhibited)",
    "clientTransferProhibited (https://www.icann.org/epp#clientTransferProhibited)",
    "clientDeleteProhibited (https://www.icann.org/epp#clientDeleteProhibited)",
    "serverUpdateProhibited (https://www.icann.org/epp#serverUpdateProhibited)",
    "serverTransferProhibited (https://www.icann.org/epp#serverTransferProhibited)",
    "serverDeleteProhibited (https://www.icann.org/epp#serverDeleteProhibited)"
  ],
  "emails": [
    "abusecomplaints@markmonitor.com",
    "whoisrequest@markmonitor.com"
  ],
  "dnssec": "unsigned",
  "name": null,
  "org": "Automattic, Inc.",
  "address": null,
  "city": null,
  "state": "CA",
  "registrant_postal_code": null,
  "country": "US"
}
```

Para obtener el propietario del dominio hemos utilizado la función **whois**, sin embargo no ha sido posible obtener dicho propietario ya que al parecer, los registradores de DNS ofrecen la posibilidad de ofuscar/ocultar determinados campos de contacto en un registro de Whois. Esto hace que no aparezcan en ningún servicio que obtenga información para ese dominio.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El conjunto de datos que hemos obtenido facilita mucho la comparación entre las diferentes viviendas que se encuentran en el área centro de Madrid. El fichero que generamos es muy interesante, ya que permite a cualquier persona realizar una búsqueda fácil y rápida permitiendo filtrar por diferentes características como el número de baños, el número de habitaciones o si cuenta con calefacción.

A su vez, permite saber si un domicilio ha variado de precio. Ejecutando el código obtendremos un documento, que si lo comparamos con el obtenido en otro día, nos dirá si se ha encarecido el precio de la vivienda que nos interesa, o si por el contrario, éste ha bajado y es un buen momento para su compra.

De igual manera, resulta útil a la hora de generar métricas que ayuden a tomar la decisión de adquirir una vivienda, como puede ser el precio del metro cuadrado, la media de precio para un tipo concreto de vivienda (por ejemplo, viviendas de tres habitaciones) o la cantidad de oferta de viviendas de un tipo en concreto.

8. Licencia. Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- **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.**
- **Otra (especificar cuál).**

La licencia que consideramos adecuada para el dataset resultante es la CC0: "Released Under CC0: Public Domain License". Esta licencia permite renunciar a los sus derechos de autor y derechos en la mayor medida permitida por la ley. De esta forma otras personas pueden desarrollar, mejorar y reutilizar estos datos. Esta licencia no requiere de atribución, va a permitir reutilizar y transformar los datos así como darles un uso comercial.

9. Código. Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

- **El código deberá ubicarse en la carpeta /source del repositorio.**
- **Se deben indicar las librerías y versiones utilizadas. P. ej., en Python pueden obtenerse mediante el comando pip3 freeze > requirements.txt**
- **En el documento PDF se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo las habéis resuelto.**

Se proporciona un archivo llamado “requirements.txt” con todas las librerías instaladas en el entorno utilizado y su versión.

El código utilizado para realizar el scraping de Fotocasa se ha llevado a cabo con python.
Iteramos por cada una de las 70 primeras páginas de la búsqueda de Madrid Capital en todas las zonas. Utilizando la librería selenium hacemos una búsqueda en Chrome para la página correspondiente a la iteración. Después, guardamos el código HTML y utilizando beautifulsoup lo parseamos. Aprovechando que toda la información que necesitamos se encuentra bajo la etiqueta “article” iteramos por cada elemento con esta etiqueta. Dentro de esta iteración extraemos cada campo que nos interesa indicando la clase con la que se identifica y los vamos añadiendo a la lista correspondiente a su campo. Por último, cerramos el explorador chrome, creamos un dataframe con las listas que hemos generado y exportamos el dataframe a un archivo .csv.

Las dificultades que hemos encontrado han sido:

- Inicialmente solo conseguimos descargar la información correspondiente a los primeros anuncios de cada página. Esto se debía a que solo se extraía la información de los anuncios que habíamos visualizado. Para solucionar este problema introdujimos un “scroll down” cada vez que cargamos una nueva página.
- El campo título nos parecía muy interesante porque en la mayoría de los anuncios muestra información de dónde está localizada la vivienda. Nos encontramos con el inconveniente de que dentro de este campo string utilizan la coma y esto interfería a la

hora de leer los datos en el archivo csv. Para solucionarlo, antes de guardar el campo en la lista de títulos utilizamos regex para sustituir la coma por la barra lateral.

- También nos encontramos con la dificultad de que no todos los anuncios cuentan con las mismas características. Esto suponía un problema ya que si, por ejemplo, una vivienda no tenía calefacción este campo no existía y a la lista de calefacción no se añadía ningún elemento, lo que desajustaba la lista (cuya posición indica la vivienda a la que corresponde). Para solucionarlo añadimos una sentencia condicional que, en caso de que no tenga ese campo añada a la lista correspondiente un string indicándolo ("Sin calefacción", "Sin ascensor"...).
- Otra dificultad fue encontrar la etiqueta "article" que es común a todos los anuncios. Inicialmente utilizabamos otras etiquetas que variaban en función del tipo de artículo (premium, advance, minimal…) lo que generaba que algunos anuncios se quedasen sin extraer.

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta /dataset del repositorio. Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

El dataset está publicado en Zenodo y disponible en el enlace DOI: https://doi.org/10.5281/zenodo.7348662.

11. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

El video se encuentra en el siguiente enlace google drive:

https://drive.google.com/drive/folders/1X6e9daiSq4ftyPD6GsVgTPAoKa62MASK?usp=sharing

Contribuciones

Contribuciones	Firma
Investigación previa	CSH, ASH
Redacción de las respuestas	CSH, ASH
Desarrollo del código	CSH, ASH
Participación en el video	CSH, ASH

Bibliografía

- https://creativecommons.org/share-your-work/public-domain/cc0/
- https://www.youtube.com/watch?v=o8s9z6icgPY
- https://www.youtube.com/watch?v=RjfqdJEwWyU&t=334s
- https://stackoverflow.com/questions/13960326/how-can-i-parse-a-website-using-selenium-and-beautifulsoup-in-python
- https://stackoverflow.com/questions/12451997/beautifulsoup-gettext-from-between-p-not-picking-up-subsequent-paragraphs
- https://stackoverflow.com/questions/5041008/how-to-find-elements-by-class
- https://stackoverflow.com/questions/30942041/slow-scrolling-down-the-page-using-selenium
- https://www.geeksforgeeks.org/find_element_by_class_name-driver-method-selenium-python/
- https://stackoverflow.com/questions/22411910/beautiful-soup-article-scraping
- https://webmasters.stackexchange.com/questions/704/how-can-i-hide-whois-information