Práctica 1: Tipología y ciclo de vida de los datos

Fecha
15/04/2019
Objeto
Creación de un dataset de inmuebles a partir de los datos contenidos en Fotocasa y Pisos
Autores
Irene Fernández Héctor Hernández

1. **Contexto**. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Hoy en día es más fácil que nunca buscar inmuebles. Lejos quedan los tiempos en que se debía utilizar revistas de anuncios.

Actualmente varios portales web incluyen cientos de miles de anuncios, con la posibilidad de filtrar por todas las características imaginables. No obstante, si se accede a un portal inmobiliario con un objetivo distinto a la búsqueda de un inmueble en particular, la interfaz propia de estos sitios web hace farragoso la obtención de información. Técnicas como el web scraping facilitan tareas como ésta, evitando la navegación manual por las decenas de páginas que conforman los resultados de una localidad concreta.

El portal Fotocasa.es y Pisos.com han sido elegidos para la realización de esta práctica por los siguientes motivos:

- Política flexible con los web crawlers
- Estructuración de la información
- Baja tasa de bloqueos y/o saturación del sitio web

Otros portales como Idealista directamente bloquean el uso de robots para la obtención de información. De hecho, Idealista proporciona una API para el acceso a sus datos, pero es de pago.
A continuación se muestra una tabla donde se resumen estas conclusiones.

Página web	Motivación	Conclusión
Idealista.com	Es una de las principales web de compra venta y alquiler de vivienda en España	Política estricta antirobots y API de pago. Se descarta.
Fotocasa.es	Es una de las principales web de compra venta y alquiler de vivienda en España, por detrás de Idealista	Es posible la extracción de información. No tan organizado como se esperaba. Links de búsqueda engorrosos
Pisos.com	Menos viviendas que en Fotocasa e Idealista, pero muy bien estructurada y fácilmente extraíble	Sin problemas para la extracción de información. Links sencillos e información estructurada.

2. **Título del dataset**. Elegir un título que sea descriptivo.

Como será un dataset que junta información sobre inmuebles de varias fuentes, se denominará ParqueInmobiliario.

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 conjunto de datos extraído presenta la información detallada sobre los inmuebles que componen el parque inmobiliario a la venta en Madrid y Barcelona. Para ello, se utiliza la información públicamente disponible en dos de los principales portales web inmobiliarios en España, Fotocasa.es y Pisos.com.

4. **Representación gráfica**. Presentar una imagen o esquema que identifique el dataset visualmente

A continuación se muestra una imagen que sitúa mediante coordenadas los distintos inmuebles en un mapa en la web de Fotocasa.es

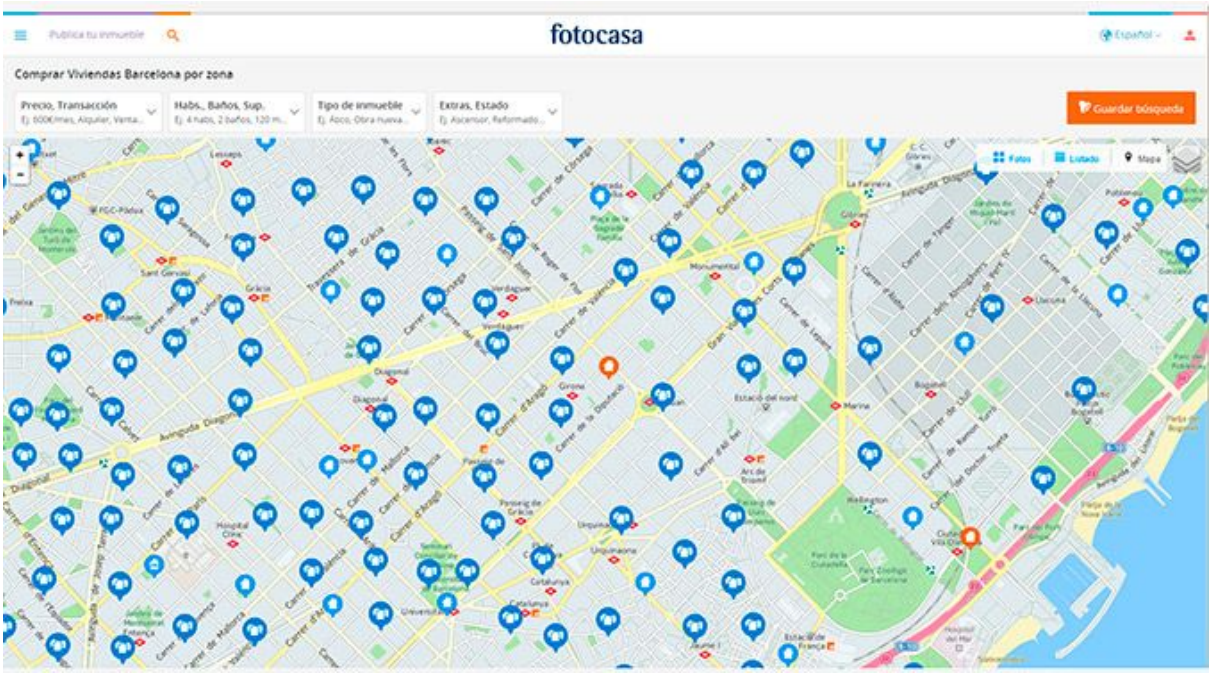

Figura 1 - Portal inmobiliario Fotocasa.es

Esta segunda imagen hace referencia al buscador por tipo de acción a realizar (Alquiler, compra), tipo de inmueble y zona geográfica, presente en la página Pisos.com

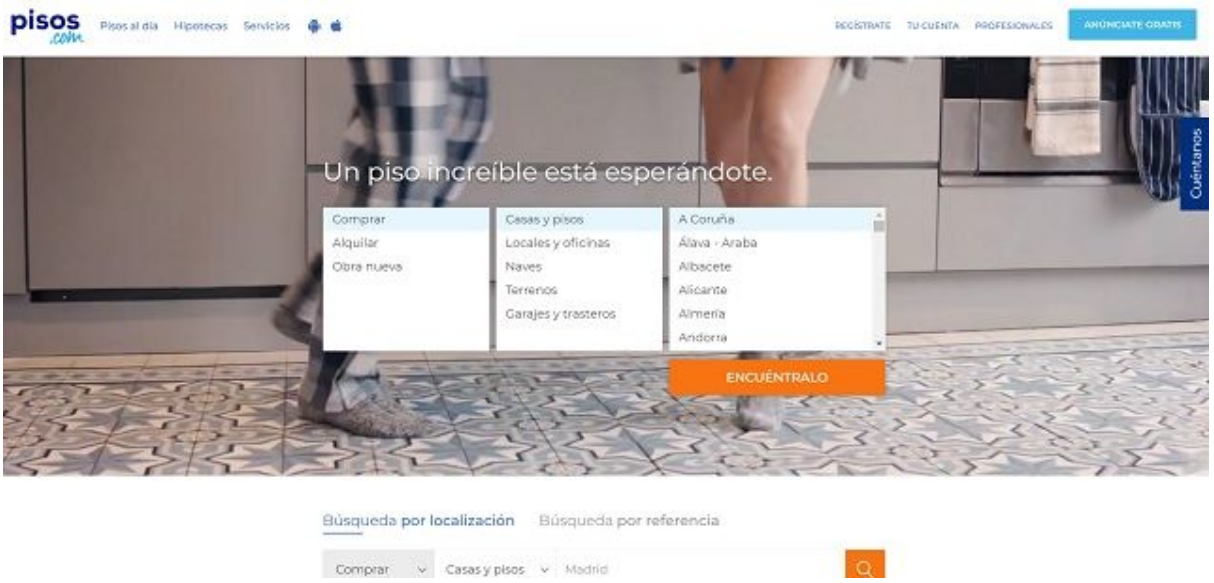

Figura 2 - Portal inmobiliario Pisos.com

La tercera y cuarta imagen muestra mediante el método head e ifo respectivamente de la librería Pandas para el lenguaje de programación Python, cómo se verían los datos una vez extraídos y almacenados en el csv

	ciudad	comunidad	habitaciones	link	particular	precio	referencia	superficie	web	banos
1	Madrid Capital	Madrid	3	https://www.fotocasa.es/es/comprar/vivienda/ma...	Profesional	92.0	150802392	53	Fotocasa	
1	Madrid Capital	Madrid	2	https://www.fotocasa.es/es/comprar/vivienda/ma...	Profesional	89.9	150377752	63	Fotocasa	
1	Madrid Capital	Madrid	1	https://www.fotocasa.es/es/comprar/vivienda/ma...	Profesional	165.0	150801881	31	Fotocasa	
1	Algete	Madrid	3	https://www.fotocasa.es/es/comprar/vivienda/al...	Profesional	163.0	147625495	98	Fotocasa	
4	Meco	Madrid	6	https://www.fotocasa.es/es/comprar/vivienda/me...	Profesional	598.0	148243983	390	Fotocasa	

Figura 3 - Ejemplo de visualización del dataset mediante la librería Pandas

```
<class 'pandas.core.frame.DataFrame'>
Index: 713 entries, 1 to 1
Data columns (total 9 columns):
ciudad          713 non-null object
comunidad       713 non-null object
habitaciones    712 non-null object
link            713 non-null object
particular      713 non-null object
precio          711 non-null object
referencia      713 non-null object
superficie      713 non-null object
web             713 non-null object
dtypes: object(9)
memory usage: 27.9+ KB
```

Figura 4 - Salida tras aplicar el método info de la librería Pandas

5. **Contenido**. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El dataset 'ParqueInmobiliario' incluye los siguientes campos:

Campo	**Tipo de dato**	**Descripción**
Referencia	Numérico, de tipo real	Identificador de anuncio inmobiliario
Precio	Numérico, de tipo entero	Precio del inmueble
Particular	Booleano (Sí / No)	Identifica si el inmueble lo oferta un particular o un profesional inmobiliario
Habitaciones	Numérico, de tipo entero	Número de habitaciones del inmueble
Banos	Numérico, de tipo entero	Número de baños del inmueble
Superficie	Numérico, de tipo real	Superficie del inmueble en metros cuadrados
Link	Carácter (URL)	URL del anuncio del inmueble
Web	Carácter	Portal web donde se aloja el anuncio del inmueble
Ciudad	Carácter	Ciudad donde se haya ubicado el inmueble
Comunidad	Carácter	Comunidad autónoma donde se haya ubicado el inmueble

Los datos han sido extraídos durante el mes de Abril de 2019.

Para la construcción del dataset 'ParqueInmobiliario' se ha utilizado el framework de Python Scrapy.

Los datos han sido extraídos tanto de la web fotocasa.es como de pisos.com.

6. **Agradecimientos**. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Fotocasa.es es un portal inmobiliario especializado en la compraventa y alquiler de viviendas de segunda mano y de obra nueva en España. Compite en el negocio de los anuncios clasificados de venta y alquiler de viviendas en España. Permite publicar y buscar anuncios de inmuebles a través de internet.

Fotocasa.es elabora desde enero de 2005 un indicador sobre la evolución del precio de la vivienda de segunda mano en España, así como con un blog con noticias sobre la actualidad inmobiliaria.

Hoy en día Fotocasa.es pertenece a Schibsted Classified Media Spain (SCM Spain), que además cuenta con los portales Infojobs, Segundamano.es, Coches.net o Milanuncios.

Pisos.com es un portal inmobiliario fundado el 19 de enero de 2009[1] y que pertenece a la empresa española HabitatSoft SL. Pisos.com forma parte de los servicios web de anuncios clasificados del grupo español de comunicación multimedia Vocento. (https://www.abc.es/hemeroteca/historico-19-01-2009/abc/Economia/vocento-lanza-pisoscom-para-liderar-el-mercado-inmobiliario-de-clasificados_912553632619.html) El target de usuarios del portal se establece en un perfil de personas de entre 18 y 55 años, de clase media-alta y que desean encontrar una vivienda, ya sea para comprar, vender, alquilar o compartir.
En mayo de 2009, Pisos.com lanzó su Gabinete de Estudios, encargado de realizar informes sobre el sector, fundamentalmente en lo que se refiere a precios medios de venta y al alquiler de vivienda de segunda mano. La empresa también mantiene un canal de actualidad inmobiliaria, otro dedicado al hogar y una comunidad virtual donde los usuarios pueden exponer sus dudas sobre vivienda.
En septiembre de 2010, Pisos.com lanzó el portal Pisocompartido.com orientado al mercado de las habitaciones en alquiler (https://www.europapress.es/economia/construccion-y-vivienda-00342/noticia-economia-vivienda-pisoscom-integra-oferta-pisocompartidocom-mas-160000-anuncios-20100914122122.html)

7. **Inspiración**. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Actualmente en ciudades españolas como Madrid y Barcelona hay mucho interés en la compraventa de vivienda. Debido a la creciente burbuja inmobiliaria (https://www.elmundo.es/economia/2018/12/21/5c1bd277fc6c831d488b45ba.html), tanto inversores como particulares deben estar muy atentos a las fluctuaciones de este mercado para conseguir rentabilidad o simplemente un lugar donde habitar adecuado a su capacidad adquisitiva (https://cincodias.elpais.com/cincodias/2019/03/26/economia/1553619322_728918.html).
Es por ello, que un dataset como 'ParqueInmobiliario' puede ayudar en la resolución de problemas como los siguientes:

- Particulares:
  - Evaluar el precio medio de la zona antes de poner a la venta un inmueble o comenzar la negociación para su adquisición
- Inmobiliarias:
  - Conocer los precios de los inmuebles ofertados por la competencia
- Ayuntamientos:
  - Establecer el precio de inmuebles de protección oficial
- Analistas:
  - Conocer cómo influencian en el precio de un inmueble características como la superficie, el número de habitaciones, el número de baños, etc.

8. **Licencia**. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

○ 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
○ Other (specified above)
○ Unknown License

La licencia elegida para el dataset es Released Under CC BY-NC-SA 4.0 License.

El motivo de esta elección es que los portales web propietarios de los datos prohíben expresamente en su política de datos la explotación comercial de los mismos.

9. **Código**. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

La totalidad del código se encuentra en el repositorio https://github.com/EdelBlau/PEC_TPC, junto con una explicación de los distintos ficheros.
La solución se ha desarrollado en Python, haciendo uso del framework Scrapy para la generación de los spiders.

10. **Dataset**. Presentar el dataset en formato CSV

El dataset se encuentra en el mismo repositorio que el código.

Contribuciones	**Firma**
Investigación previa	IFM / HHM
Redacción de las preguntas	IFM / HHM
Desarrollo de código	IFM / HHM