Práctica 1: Tipología y Ciclo de Vida de los Datos

Jokin Cuesta Arrillaga

Andrés Escamilla Baca

Contexto

A diferencia de la última década, el entorno económico actual se está tornando en una situación altamente inflacionaria. El último dato publicado por el INE es de una subida del IPC del 7,6% en tasa interanual. En este escenario, la vivienda se ha convertido en el valor refugio de muchos ciudadanos españoles, siendo está el destino de los ahorros para una gran mayoría.

Idealista.com es el portal inmobiliario número 1 en España y es una referencia fundamental para la compra de vivienda, ya sea con motivo habitacional o como inversión. Aunque el mismo portal proporciona varios servicios de tasación y evolución de indicadores relativos a la vivienda, en la mayoría de los casos son de pago.

En este escenario, el conocimiento de la evolución de la vivienda cobra mucho sentido, y nuestra herramienta de web scraping viene a solucionar las necesidades de recopilación de información para tener una base de datos propia.

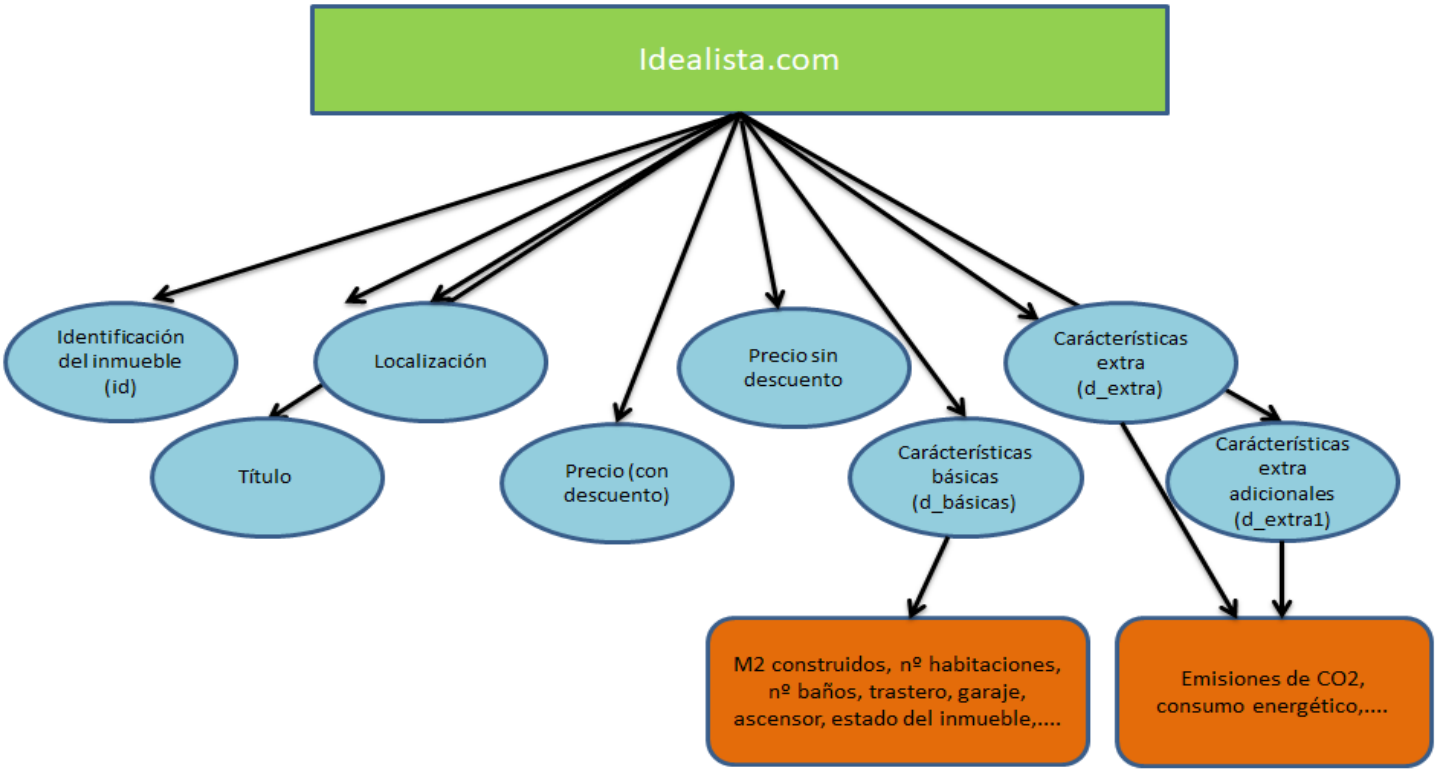
Título

Información sobre viviendas en venta situadas en Sevilla por la web Idealista.com.

Descripción del dataset

Este dataset consta de 30 viviendas de Sevilla (código postal 41003) que se sitúan en venta desde el portal web de Idealista. En él, se puede apreciar toda la información básica que ofrecen los anuncios: ubicación, precio, bajada de precio, elementos básicos de la vivienda y el consumo energético.

Representación gráfica



Contenido

El dataset generado contiene las variables más importantes que caracterizan a una vivienda en venta, desde su precio, situación, número de habitaciones, metros cuadrados construidos y útiles, así como los servicios añadidos con los que cuenta: ascensor, garaje, trastero, sótano,...

El dataset final obtenido de la práctica de Web Scraping en la web de Idealista contará de estas columnas:

COLUMNA	DESCRIPCIÓN	EJEMPLO
id	str con el id del inmueble	97301650
título	str con el título de la vivienda	Piso en venta en Encarnación-Las Setas

localizacion	str con el lugar de la vivienda.	Centro, Sevilla
precio	int con el precio rebajado (si existe) de la vivienda en euros.	560000
precio_sin_descuento	int con el precio original de la vivienda en euros	600000
d_basica	lista de strings con los detalles básico de la vivienda	[“220 m2 construidos”, “3 habitaciones”, “3 baños”, “Segunda mano/buen estado”, “Armarios empotrados”, “Trastero”, “Construido en 1978”, “Calefacción individual”]
d_extra	lista de strings con algunos detalles extra del equipamiento de la vivienda	[“Aire acondicionado”]
d_extra1	lista de strings con algunos detalles extra de la vivienda acerca del consumo energético.	[“Consumo: 79 kWh/m2 año”, “Emisiones: 14 kg CO2/m2 año”]

Agradecimientos

Estos datos han sido obtenidos de la web de Idealista. Idealista es una compañía española fundada el 4 de octubre de 2000 que ofrece a través de Internet entre otros los servicios de portal inmobiliario en España, Italia y Portugal. Además de anunciar pisos, Idealista utiliza su conjunto de datos para realizar distintos ejercicios como son estimaciones de precios, valoraciones del mercado actual, evolución del mercado inmobiliario durante los últimos años. El punto negativo de este servicio es el ser de pago.

El dataset obtenido tiene como objetivo realizar un análisis del mercado inmobiliario en Sevilla. Hemos utilizado como referencia el vídeo de Youtube de Miguel Ángel Gisbert *Cómo Encontrar CASAS Baratas - Parte 1: Scrapear datos de IDEALISTA*, en el cuál trata de scrapear datos de Idealista en Madrid.

No hemos encontrado ningún análisis parecido a lo que proponemos en términos de ubicación, pero sí que existen proyectos similares con otras ciudades como puede ser el de Boston, realizado por Víctor Román en su trabajo: *Proyecto Machine Learning: Predicción de Precios de Viviendas en Boston con Regresión* posteado en Medium.

Inspiración

Como comentado en el contexto del trabajo, el trabajo se hace interesante para realizar un estudio de mercado inmobiliario, bien para comprar una vivienda habitual o realizar una inversión. Por tanto, esta práctica de Web Scraping nos ayuda entonces a recopilar información para conocer, por ejemplo, si debemos invertir en una propiedad, o determinar el potencial de alquiler de una ciudad o distrito.

Siempre se debe hacer un estudio de mercado inmobiliario, ya sea al comprar o vender una propiedad, ya que ayuda a comprender el mercado actual, cuánto valen propiedades similares, si se trata de una propiedad de inversión, cuánto se puede cobrar por el alquiler, etc.

El dataset obtenido serviría, por ejemplo, para valorar mediante algoritmos de Machine Learning de regresión un precio de cotización y ayudar a los compradores a ver si el precio de venta es demasiado alto, bajo o razonable. Este estudio coincide con el anteriormente mencionado de Víctor Román.

Se ha escogido Sevilla por ser la ciudad natal de uno de los componentes del grupo, pero se podría fácilmente extrapolar el proyecto a distintas ciudades que suben sus anuncios a la web de Idealista.

Las preguntas a las que pretende dar respuesta este conjunto de datos, entre otros, se listan a continuación:

- ¿Es el precio de la casa adecuado a su valor de mercado?
- ¿Qué zonas son las más cotizadas?
- ¿Qué elementos hacen que el valor de una casa suba?
- ¿Dónde es la rentabilidad más alta?
- ¿Cómo ha ido evolucionando el precio de las casas?
- ¿Cuáles son las tendencias actuales del mercado?
- ¿Qué productos o servicios están disponibles en el mercado?

Como punto a mejorar, vemos que al introducir un código postal, sólo obtenemos los datos de las casas de la primera página que ofrece la web, es decir, treinta casas. Si bien nosotros buscamos tener todas las casas que ofrece Idealista, deberíamos introducir en el código una paginación mediante Web Scraping, es decir, que el programa recorra todas las páginas de Idealista para recolectar los datos de todas las casas. Ante el plus de dificultad y el riesgo a banear el IP, hecho que nos ha ocurrido, hemos decidido simplificar el proyecto y scrapear únicamente los datos de la primera página.

Licencia

La licencia escogida para la presentación de este conjunto de datos ha sido CC BY-SA 4.0 License. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

- Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.
- Se permite un uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.
- Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma. Esto hace que el trabajo del autor original continúe distribuyendose bajo los términos que él mismo planteó.

Contribuciones	**Firma**	
Investigación previa	J. C. A.	A.E.B.
Redacción de las respuestas	J. C. A.	A.E.B.
Desarrollo del código	J. C. A.	A.E.B.