PRA1

¿Cómo podemos capturar los datos de la web?

martes, 25 de abril de 2023

Máster Universitario de Ciencia de Datos

Emilio Jesús Ávila González

Contenido

1. Contexto

La información se ha recolectado en el contexto de la búsqueda de viviendas en venta en Palma de Mallorca, España. El sitio web elegido, Idealista, proporciona información relevante sobre inmuebles en venta, incluyendo detalles como el precio, la ubicación, el área, el número de habitaciones y si la propiedad cuenta con ascensor, entre otras características. Esta información es valiosa para potenciales compradores e inversores, así como para análisis de tendencias del mercado inmobiliario en la región. La dirección del sitio web es: https://www.idealista.com/venta-viviendas/palma-de-mallorca-balears-illes/.

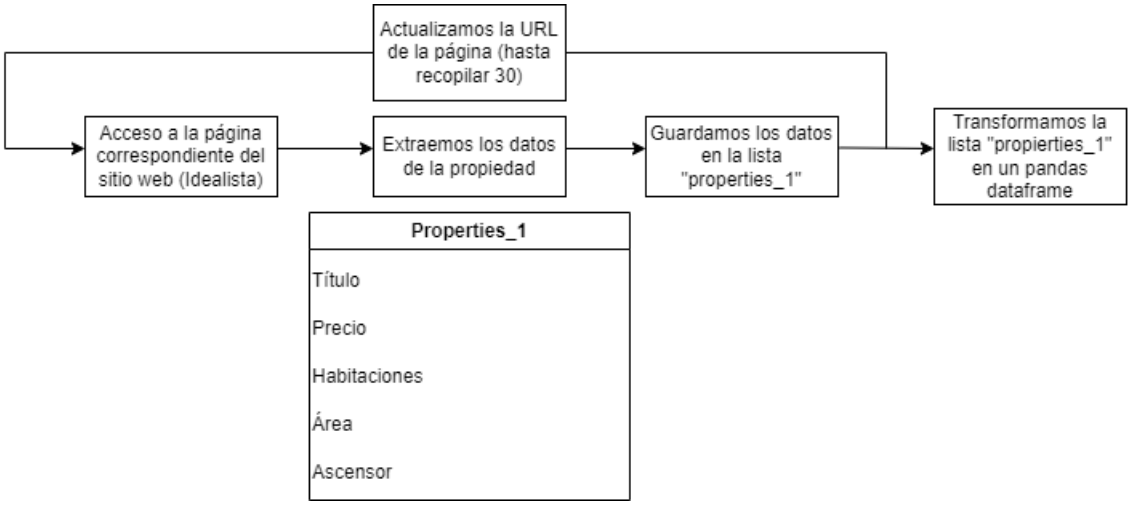
2. Título

Viviendas en venta en Palma de Mallorca - Detalles y características (Idealista).

3. Descripción del dataset

Este conjunto de datos contiene información detallada sobre las viviendas en venta en Palma de Mallorca, extraída del sitio web Idealista. Las propiedades incluidas en el dataset abarcan distintos rangos de precios, tamaños y características. Cada registro del dataset contiene el título del anuncio, el precio de la vivienda, el número de habitaciones, el área en metros cuadrados y si la propiedad cuenta con ascensor o no. Este dataset puede ser de interés para aquellos que buscan comprar una vivienda en Palma de Mallorca, así como para investigadores que deseen analizar el mercado inmobiliario en esta área.

4. Representación gráfica



5. Contenido

El dataset incluye información sobre viviendas en venta en Palma de Mallorca, extraída durante el mes de abril de 2023. Los campos incluidos en el dataset son los siguientes:

- Título: Título del anuncio, que incluye información sobre la ubicación y características de la vivienda (por ejemplo, "Piso en calle de Joan Alcover, Foners, Palma de Mallorca").
- Precio: Precio de la vivienda en euros (por ejemplo, "1.200.000€").
- Habitaciones: Número de habitaciones en la vivienda, expresado con la abreviatura "hab." (por ejemplo, "3 hab.").
- Área: Área de la vivienda en metros cuadrados, expresado con la abreviatura "m²" (por ejemplo, "190 m²").
- Ascensor: Un valor booleano (True o False) que indica si la vivienda cuenta con ascensor o no.

Es importante tener en cuenta que estos datos pueden cambiar con el tiempo, ya que las propiedades pueden ser vendidas, retiradas del mercado o actualizadas con nueva información. Por lo tanto, el dataset representa una instantánea del mercado inmobiliario en Palma de Mallorca durante el período mencionado.

6. Propietario

El propietario del conjunto de datos es Idealista, una plataforma de bienes raíces que proporciona información sobre propiedades en venta y alquiler en España. La información sobre las propiedades se recopila directamente del sitio web de Idealista (https://www.idealista.com). Los datos no incluyen información personal ni violan la privacidad de los usuarios.

Se han encontrado dos análisis previos en Kaggle relacionados con datos extraídos de Idealista. El primero, titulado "Idealista Madrid Real State" (https://www.kaggle.com/datasets/josegabrielgonzalez/idealista-madrid-real-state), utiliza un conjunto de datos sobre propiedades en Madrid. El segundo, titulado "Sevilla Housing" (https://www.kaggle.com/datasets/javieradvani/sevilla-housing), se enfoca en propiedades en Sevilla. Ambos análisis utilizan datos obtenidos mediante web scraping y presentan información sobre características como dirección, precio, cantidad de habitaciones, metros cuadrados, tipo de vivienda y más.

Numerosos estudios y análisis similares en el campo de los bienes raíces, como la predicción de precios de vivienda, la identificación de tendencias del mercado inmobiliario y el análisis de la oferta y demanda de viviendas en diversas regiones, también sirven como base para la importancia y aplicabilidad de este conjunto de datos en el análisis de propiedades inmobiliarias en Palma de Mallorca.

En cuanto a los principios éticos y legales, se han tomado las siguientes medidas:

1. Se ha utilizado un web scraping responsable, extrayendo datos solo de 30 páginas en lugar de toda la plataforma y usando tiempo de espera (sleep) entre las solicitudes para evitar sobrecargar el servidor del sitio web de Idealista.

2. Los datos recopilados se limitan a la información de propiedad pública y no incluyen datos personales o información confidencial de los usuarios.

3. El conjunto de datos se comparte con una licencia adecuada (Creative Commons Attribution) para garantizar el reconocimiento del propietario de los datos y permitir su uso en proyectos de investigación y análisis.

7. Inspiración

Este conjunto de datos sobre propiedades en venta en Palma de Mallorca puede resultar interesante y útil en varios contextos. Dado que el mercado inmobiliario es un sector de gran importancia tanto a nivel económico como social, contar con información detallada y actualizada sobre las propiedades en venta puede permitir responder a diversas preguntas y abordar diferentes retos analíticos. Algunos ejemplos de preguntas y análisis que podrían realizarse con este conjunto de datos son:

1. Predicción de precios de vivienda: Utilizando datos históricos y características de las propiedades, se pueden desarrollar modelos de aprendizaje automático para predecir los precios de las viviendas en el futuro. Estos modelos pueden ser útiles para inversores, compradores y vendedores al tomar decisiones sobre la compra o venta de propiedades.

2. Análisis de tendencias del mercado inmobiliario: Examinar cómo varían los precios, las características de las propiedades y la oferta de viviendas en el tiempo y en diferentes áreas de Palma de Mallorca. Este análisis puede ayudar a identificar áreas de crecimiento potencial y cambios en la demanda de vivienda en función de factores externos, como la economía o el desarrollo urbano.

3. Segmentación del mercado: Identificar segmentos específicos del mercado inmobiliario en Palma de Mallorca en función de características como el tipo de vivienda, el tamaño, la ubicación y otros factores relevantes. Esta información puede ser valiosa para los agentes inmobiliarios, desarrolladores y políticas de vivienda.

4. Evaluación de la accesibilidad y asequibilidad de la vivienda: Analizar la relación entre el precio de las viviendas y factores socioeconómicos, como los ingresos de la población y el costo de vida en Palma de Mallorca. Este tipo de análisis puede ser útil para entender el acceso a la vivienda y apoyar políticas de vivienda justas y equitativas.

Aunque no se encontraron análisis anteriores específicos utilizando este conjunto de datos de Palma de Mallorca, los análisis mencionados previamente en el apartado 6, como "Idealista Madrid Real State" y "Sevilla Housing", abordan temas similares en el campo de los bienes raíces en otras ciudades españolas. Este conjunto de datos de Palma de Mallorca podría contribuir a expandir y enriquecer dichos estudios y abrir nuevas oportunidades para analizar el mercado inmobiliario en la región.

8. Licencia

Para este dataset, se ha seleccionado la licencia "Creative Commons Attribution 4.0 International (CC BY 4.0)".

La elección de esta licencia se basa en los siguientes motivos:

1. Permisividad: La licencia CC BY 4.0 permite a los usuarios compartir (copiar y redistribuir el material en cualquier medio o formato) y adaptar (remezclar, transformar y crear a partir del material) el dataset para cualquier propósito, incluso comercialmente. Esto significa que el conjunto de datos es accesible y reutilizable por una amplia gama de usuarios e investigadores.

2. Atribución: La licencia CC BY 4.0 requiere que los usuarios atribuyan adecuadamente el conjunto de datos a su propietario (Idealista en este caso), proporcionando un enlace a la licencia y señalando si se han realizado cambios en los datos. Esto garantiza que el propietario de los datos reciba el reconocimiento adecuado por su trabajo y contribución al proyecto.

3. Compatibilidad: La licencia CC BY 4.0 es compatible con otras licencias Creative Commons y muchas licencias de código abierto, lo que facilita su uso en conjunto con otros conjuntos de datos y proyectos de investigación.

Al elegir la licencia CC BY 4.0, se promueve el uso responsable y ético del dataset, garantizando el reconocimiento del propietario y permitiendo que otros investigadores y usuarios lo utilicen y modifiquen según sus necesidades.

9. Código

9.1 Introducción

En este proyecto, el objetivo es recolectar datos sobre propiedades en venta en Palma de Mallorca desde el sitio web Idealista. La información recolectada incluye detalles como el título, el precio, el área, la cantidad de habitaciones y la presencia de un ascensor. Hemos implementado un código en Python para extraer esta información y almacenarla en una estructura de datos.

9.2 Descripción y liberías utilizadas

Para llevar a cabo la recolección de datos, se han utilizado las siguientes librerías en Python:

1. BeautifulSoup: Permite analizar y extraer información de páginas HTML. Se usa para encontrar y extraer los detalles de las propiedades dentro del HTML del sitio web.

2. Selenium (webdriver): Facilita la navegación automatizada en páginas web y permite interactuar con el contenido dinámico. Se utiliza para cargar las páginas, establecer un tiempo de espera para permitir la carga de la página y dar tiempo al usuario para resolver el CAPTCHA, si es necesario.

3. undetected_chromedriver: Es una extensión de Selenium que permite evitar la detección de navegadores automatizados por parte de los sitios web. Se emplea para asegurar el acceso a las páginas sin ser bloqueados.

4. time (sleep): Proporciona la función sleep, que permite añadir pausas entre las solicitudes para evitar sobrecargar el sitio web y reducir la probabilidad de ser bloqueados.

5. pandas: Es una librería muy popular en Python para el manejo y análisis de datos en estructuras como DataFrames. En este caso, se utiliza para organizar y manipular los datos extraídos de las propiedades y guardarlos en formato CSV.

9.3 Dificultades y soluciones implementadas

Algunas dificultades encontradas durante el proceso de extracción de datos en el sitio web de Idealista incluyen:

Contenido dinámico: El sitio web utiliza contenido dinámico, lo que dificulta el acceso directo a los datos utilizando librerías como BeautifulSoup o requests. Para abordar esta dificultad, hemos utilizado la librería Selenium, que permite cargar y navegar por páginas web con contenido dinámico.

Prevención de scraping: Idealista puede detectar y bloquear solicitudes provenientes de navegadores automatizados. Para superar esta dificultad, hemos empleado la librería undetected_chromedriver, que permite utilizar Chrome de manera más discreta para evitar ser detectados.

CAPTCHAs: Idealista puede requerir la resolución de CAPTCHAs para acceder al contenido de la página. Hemos implementado un tiempo de espera utilizando la función 'implicitly_wait()' de Selenium, lo que permite al usuario resolver manualmente el CAPTCHA, si es necesario, antes de continuar con la extracción de datos.

Estructura de la página y extracción de datos: El sitio web de Idealista tiene una estructura compleja y es necesario identificar los elementos correctos para extraer la información deseada. Para ello, hemos utilizado la librería BeautifulSoup, que facilita la búsqueda y extracción de elementos dentro del HTML. Mediante el análisis de la estructura de la página, hemos sido capaces de identificar las clases y atributos relevantes para extraer los detalles de las propiedades.

Estas soluciones han permitido abordar las dificultades encontradas en el sitio web y recolectar de manera efectiva la información sobre propiedades en venta en Palma de Mallorca.

10. Dataset

El dataset obtenido es el resultado del proceso de recolección de datos de propiedades en venta en Palma de Mallorca a partir del sitio web Idealista. El dataset se estructura en un archivo CSV, con las siguientes columnas:

Título: Descripción de la propiedad en venta.

Precio: Precio de venta de la propiedad.

Área: Superficie en metros cuadrados de la propiedad.

Habitaciones: Cantidad de habitaciones que posee la propiedad.

Ascensor: Indica si la propiedad cuenta con un ascensor (True) o no (False).

Para almacenar el dataset en un archivo CSV, se ha utilizado la librería pandas en Python. Primero, se ha convertido la estructura de datos que contiene las propiedades en un DataFrame de pandas y, posteriormente, se ha guardado dicho DataFrame en un archivo CSV utilizando la función to_csv.

El dataset resultante se ha publicado en Zenodo, un repositorio abierto que permite el almacenamiento y la compartición de datos de investigación. Se ha incluido una breve descripción del dataset en Zenodo, destacando la información recopilada y su origen. Tras publicar el dataset en Zenodo, se ha obtenido un enlace DOI (https://doi.org/10.5281/zenodo.7854801)que facilita su citación y acceso permanente.

Además, el dataset también se ha incluido en la carpeta /dataset del repositorio del proyecto para facilitar su consulta y uso.

11. Vídeo

Para complementar la memoria escrita, se ha creado una presentación en video que resume y explica los aspectos clave del proyecto, incluyendo el contexto, título, descripción del dataset, representación gráfica, contenido, propietario, inspiración, licencia, código y dataset. El video abarca cada uno de estos aspectos, proporcionando una visión general completa del proyecto y sus resultados.

A continuación, se proporciona el enlace para acceder a la presentación en video:

Enlace a la presentación en video:
https://drive.google.com/file/d/1a8wplTyEnylCLQCI40QI5fh6WBKMcVf8/view?usp=sharing

Contribuciones	Firma
Investigación previa	Emilio Jesús Ávila González
Redacción de las respuestas	Emilio Jesús Ávila González
Desarrollo del código	Emilio Jesús Ávila González