

PRACTICA 1

TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS

DAVID DE VEGA MARTIN

- ddevega -

Descripción de la práctica a realizar

El objetivo de esta actividad será la creación de un data set a partir de los datos contenidos en un sitio web. El idioma del sitio web elegido deberá ser español, inglés o catalán. Se deberán resolver los siguientes apartados:

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

Durante toda mi vida laboral he trabajado en el sector de la construcción, concretamente en la rama de promoción y construcción inmobiliaria de cartera propia. Siempre he deseado contar con una ***herramienta propia que sirva para establecer cuáles son los verdaderos precios de oferta tanto del mercado de obra nueva como el de segunda mano, de modo directo y que capturen un área geográfica para un periodo concreto, el actual.*** Sin estar sujeto a las estadísticas oficiales del ramo, que son muy escasas, no alcanzan el suficiente nivel de detalle y que generalmente no capturan fehacientemente la situación real del mercado, esto es, cuáles son los precios de venta más cercanos posibles a la realidad y reflejan situaciones ya pasadas.

Para esto he desarrollado variantes de los scripts que presentamos en esta práctica y que pudiesen ser aplicables a los portales inmobiliarios más importantes del país como podría ser el caso de idealista o de fotocasa, casa-trovit. Lamentablemente estas páginas tienen unos sistemas anti-scraping muy desarrollados, que no he podido burlar y que han derivado en mi baneo permanente.

Así las cosas, como aficionado al deporte, concretamente al ciclismo, el siguiente foco de interés fue obtener los catálogos completos de productos de grandes constructores como giant y trek. Al igual que con el caso de los portales inmobiliarios obtuve un baneo como resultado.

Así desarrollado el primer script de los que presento adaptado convenientemente, a unas 20 páginas de distintas temáticas me encuentro que estoy en el punto de partida.

En ese momento recuerdo Tecnocasa, la red de franquicias inmobiliarias, y encuentro que el script desarrollado puede adaptarse nuevamente y retomar el proyecto ideal que era desarrollar un spider que permitiese recorrer cualquier área geográfica del país, cuestión con la que Tecnocasa, por su implantación a nivel

nacional cumpliría sobradamente. De hecho, este mismo spider podría adaptarse con muy poco esfuerzo al resto de países en los que Tecnocasa opera como sería el caso de Italia. Y con ligeras modificaciones en un par de campos, permitiría, asimismo, evaluar el mercado de alquiler.

La web de Tecnocasa presenta una estructura uniforme donde todos los franquiciados suben la información de las propiedades que tienen en su cartera.

Hay que destacar que el principal reto al que nos enfrentamos en esta práctica es precisamente ese. La estructura de la web es la misma para todos los franquiciados. **El problema reside en que la entrada de información carece de normas fijas, lo que da lugar a que la codificación varíe en función del franquiciado.**

En aras de obtener la mayor integridad posible de la información extraída veremos que algunos campos pueden, en ocasiones, ser redundantes y que recogen información duplicada. **En las pruebas efectuadas, ambos scripts alcanzan un alto compromiso de concisión, la información que pueda perderse por diferencias de codificación, que no puede evitarse en la fase de preprocesado, puede recomponerse en la fase de procesado,** si bien esto excede los planteamientos de esta PEC.

Este spider permitiría, obtener todos los datos relevantes de una propiedad inmobiliaria situada en un área geográfica concreta para el momento actual.

Esta información serviría para realizar estadísticas de todo tipo, como tamaño medio de la vivienda en términos absolutos, como tamaños por número de habitaciones, al incluir la antigüedad se podría calcular la edad media del parque de vivienda de segunda mano disponible, asimismo con el precio del m² tanto absoluto como para cada tipo de casa podemos realizar tasaciones en función de la oferta de la zona, factor este empleado hasta por los propios arquitectos a la hora de justificar valoraciones para la concesión de hipotecas.



3. Título. Definir un título que sea descriptivo para el data set.

Los scripts desarrollados son dos:

El primero entregado en concepto de preentrega y desarrollado con la librería Scrapy.

El segundo como sugerencia de ampliación de la PEC, tras la preentrega para mostrar que el autor puede emplear métodos más avanzados como Selenium y que es capaz de recorrer webs mediante clics.

Así llegados a este punto los scripts se denominarían:

-Script portal inmobiliario Tecnocasa con Scrapy.

-Script portal inmobiliario Tecnocasa con Selenium

Este título es un tanto abstracto ya que esta es la función del spider y lo que varía es el área de aplicación de este, que sería el área geográfica sobre el que se aplicaría, cuestión que es tan simple como cambiar la url-semilla, de ambos. Llegado el caso también podría emplearse para evaluar el mercado del alquiler.

Por estas razones no podemos dar un nombre más concreto y conciso.

Hay que destacar que ambos dada una url-semilla realizan el proceso completo de extracción y envío del resultado a un archivo csv, con la única diferencia de que el de la librería Scrapy está pensado para ejecutarse desde la consola de comandos razón por la cual en la última línea de este se adjunta el comando a ejecutar.

El script de selenium se puede ejecutar directamente desde el IDE que empleemos, en mi caso PyCharm.

Empleando PIP freeze podemos observar que las versiones con las que se han desarrollado ambos scripts son:

- Scrapy 2.5.0.
- Selenium 3.141.0

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

Los dos scripts desarrollados para la práctica realizan un scraping vertical, esto es, entrada en el detalle de cada propiedad, desde el listado general de propiedades para el área geográfica designada en la url-semilla. Asimismo, se efectúa un scraping horizontal, esto es, se recorre la totalidad del listado general del área seleccionada. Ambos scraping se efectúan a un único nivel, ya que la información a extraer así lo requiere.

No es posible efectuar un scraping horizontal único, ya que la información a extraer no se encontraba en su totalidad y la que se encontraba disponible no permitía una identificación unívoca de las propiedades. Esto justifica el uso combinado de ambas. Cabe mencionar que no se requería más niveles de profundidad porque simplemente no existen en el formato de esta web.

La información que empleamos para la extracción de datos se encuentra en la caja de "Características del Inmueble" dentro de cada una de las propiedades.

Estas características se encuentran alojadas mayoritariamente en un contenedor div. Nuevamente mencionamos que el formato con el que se han rellenado la información recogida en este contenedor no es único, ya que la proporcionan los franquiciados, en un formato bastante flexible y con pocas normas de relleno.

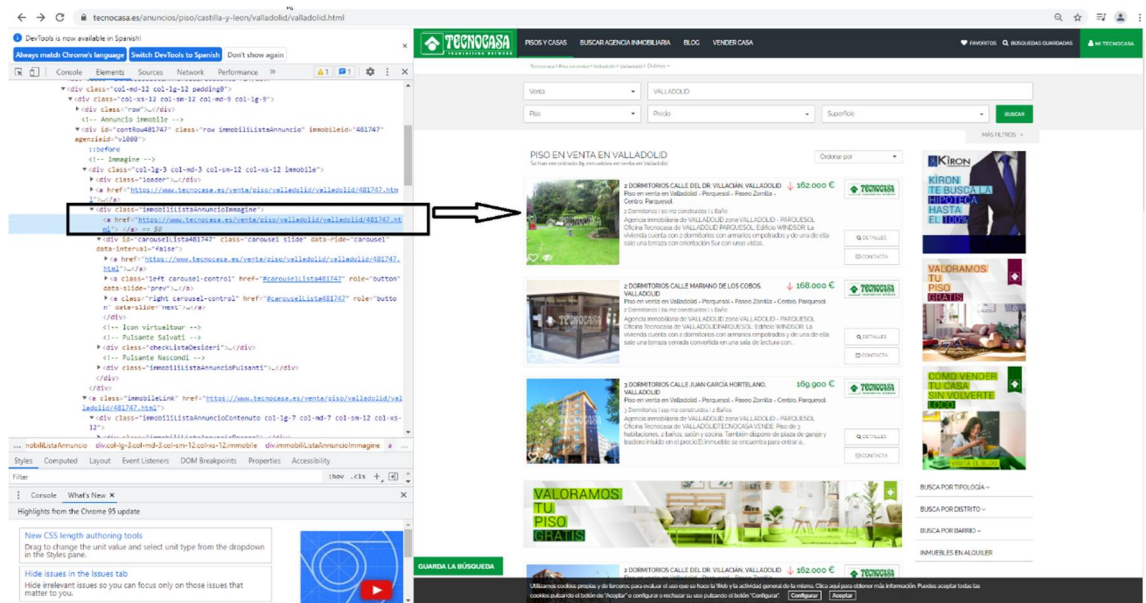
Esto hace que algunos campos cambien de lugar, no siendo un defecto de nuestro script. De todos modos, mediante muchas pruebas hemos conseguido una configuración con la que únicamente el 5-7% de registros contienen errores.

Dichos errores se pueden subsanar en el preprocesado del archivo CSV resultante mediante campos añadidos como el de ubicación en el caso del script de selenium. El script desarrollado para scrapy no cuenta con esa mejora, que sería fácilmente implementable, y que no realizamos por falta de tiempo. Simplemente sería añadir el xpath de ese campo.

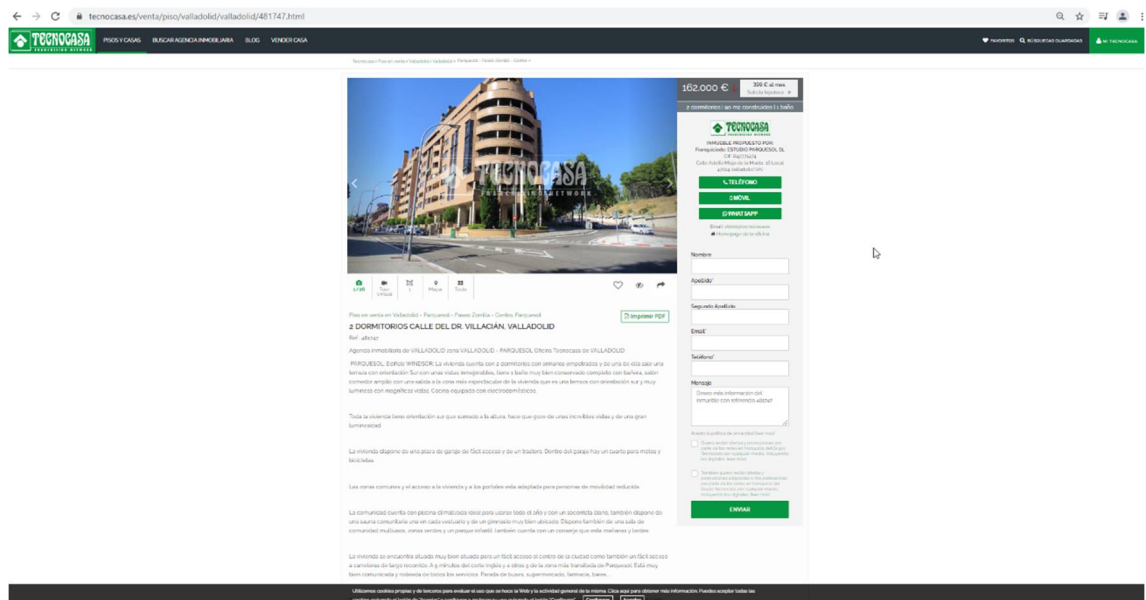
Esquema general de la página:

Partimos del listado general de la provincia que hayamos definido en la url-semilla del script.

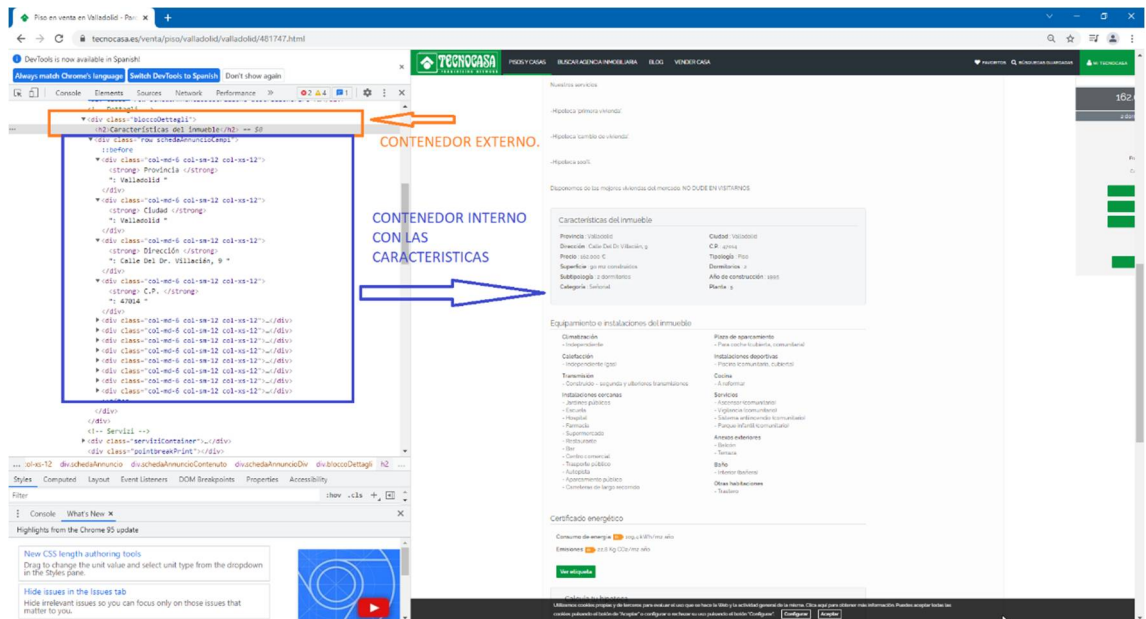
Mediante el contenedor de la clase div accedemos a un elemento a con clase href que aloja los links de las propiedades, que nos permitirá hacer scraping vertical.



Entramos en el detalle de la propiedad:



Tras la descripción de la propiedad encontramos una caja llamada Características del Inmueble, que es de donde obtendremos todos los elementos que queremos extraer.



El contenedor `div class="bloqueDettagli"` alberga a su vez otro contenedor `div class="col-md-6 col-sm-12 col-xs-12"` que guarda todas las características que queremos extraer.

Efectuado esto, hemos realizado el scraping horizontal de la primera página. Ahora tenemos que hacer el scraping horizontal, la paginación para todo el listado.

Esto lo haremos definiendo un patrón para la url así encontramos que el patrón de paginación es

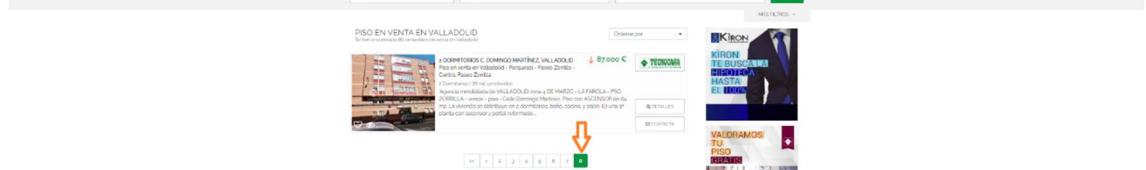
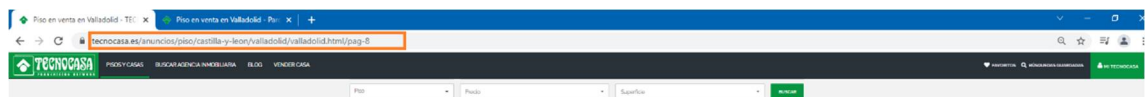
<https://www.tecnocasa.es/venta/piso/valladolid/valladolid/481747.html>
(pagina1)



<https://www.tecnocasa.es/anuncios/piso/castilla-y-leon/valladolid/valladolid.html/pag-2>



<https://www.tecnocasa.es/anuncios/piso/castilla-y-leon/valladolid/valladolid.html/pag-8>



Como podemos ver el patrón es claro:

<https://www.tecnocasa.es/venta/piso/valladolid/valladolid/481747.html>

+ '/' + pag n

El modo de hacerlo difiere enormemente en los dos scripts.

En el script de Scrapy lo incorporaremos como una Rule... regla.

En el script de Selenium, deberíamos hacerlo de modo automático haciendo clic en el botón "siguiente". El problema está en que si presionamos sobre el chevron lateral ">" es un elemento que a su vez contiene un elemento

```
<a> href="https://www.tecnocasa.es/anuncios/piso/castilla-y-leon/valladolid/valladolid.html/pag-8">>></a> == $0.
```

No he podido extraer el link con limpieza. Razón por la cual para el script de Selenium he generado un bucle for a partir de la página 1 añade los links de todas las páginas hasta el final.

Esquema Script Scrapy:

<1> Clase Artículo

< contiene las características a descargar >

<2> Clase TecnocasaCrawler

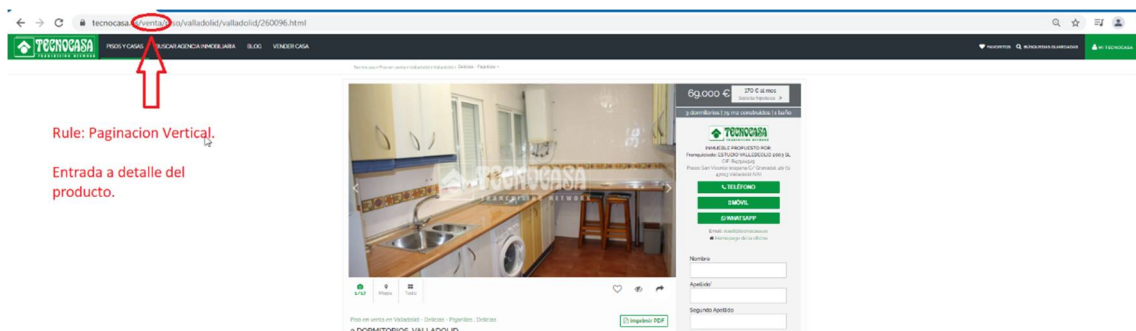
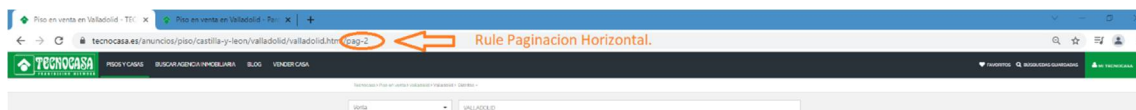
< contiene:

-url semilla.

-Rules:

paginación horizontal → pag

Verticalidad → venta



<3>Funciones de preprocesado

<4>Función de parseo de elementos.

Extraemos las características vía XPath

Enviamos al archivo CSV los resultados tras preprocesamiento.

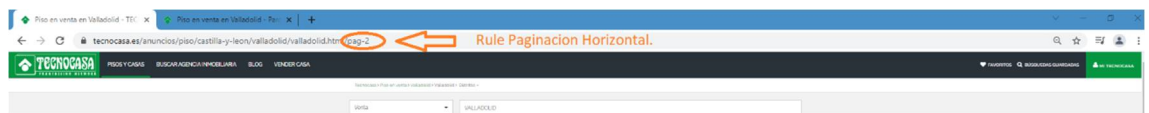
Esquema Script Selenium:

<1> Bucle FOR < obtención links para paginación horizontal>

En un string en la parte superior izquierda del listado general tenemos el total de inmuebles. Extraemos el número de inmuebles y dividiendo entre 12 anuncios por página obtenemos el número de páginas a parsear.

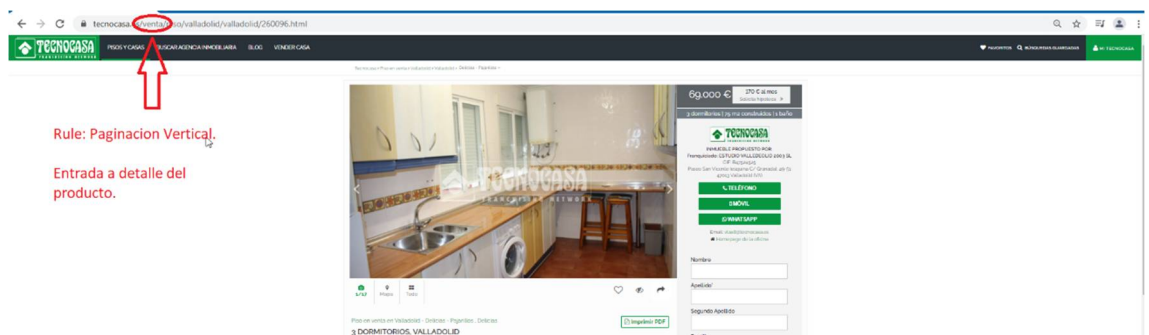
<2> Bucle externo FOR < paginación horizontal>

<3> Bucle FOR (recopila los links para scripting vertical, entrada a detalle)



<4> Bucle interno FOR < paginación horizontal>

Extraemos las características de los elementos.



<5> Funciones de Pre-Procesado.

<6> Creación de dataframe panda. Envío df a archivo CSV.

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

Tanto el script desarrollado en scrapy como el desarrollado para selenium cuentan con los campos básicos para la completa identificación de una propiedad inmobiliaria, para el área que se haya seleccionado en la url semilla:

- Provincia.
- Ciudad.
- Precio.
- Código Postal (CP).
- Superficie: Expresada en m² construidos.
- Dormitorios: número de dormitorios
- Tipología/Subtipología: Dependiendo del script. Sirve para identificar el elemento piso, local comercial, estudio etc.
- Año Construcción: Para determinar la antigüedad.
- Categoría: Factor que permite clasificar el tipo de finca en el que se encuentra ubicada la finca puede ser popular, media, señorial etc.

Por ultimo mencionar que en el script de selenium, preparado para un objetivo algo más ambicioso que el original, y ya con una pequeña ventaja en cuanto a curva de aprendizaje se incorpora el campo ubicación, que sería el compendio de Provincia, Ciudad, Barrio y Calle que debidamente preprocesado, esto es eliminado leyenda como local comercial en o piso en venta en ... etc. recoge todos los elementos de ubicación en aquellos casos confusos , en los que por errores de codificación de los franquiciados no obtenemos una descripción que permita identificar unívocamente el activo inmobiliario.

Este campo tiene especial importancia para las tareas de preprocesado posterior a la obtención del CSV, y que permitiría recuperar el 5-7% de registros con deficiencias de identificación que hemos encontrado en las pruebas realizadas.

El dataset se corresponde con la oferta que el franquiciado perteneciente al área geográfica especificada en la url semilla del spider tiene en el momento de ejecutarla.

La información de cada uno de los campos la adquirimos de los contenedores div de la ficha de la propiedad fundamentalmente.

La finca en la que se encuentra la vivienda ha sido completamente rehabilitada recientemente, mejorando su accesibilidad desde el portal, ascensor completamente nuevo, escaleras modificadas, ventanas de aluminio climatit en las zonas comunes, fachada y tejado también actualizados. Los gastos de comunidad ascienden a 53€/mes.

En Tecnocasa facilitamos el acceso a la financiación de nuestros clientes, a través de KIRON, la empresa de intermediación financiera del grupo Tecnocasa, con convenios nacionales con las principales entidades de España. Podemos llegar a conseguir hasta el 100% del valor de compraventa. Por eso, invitamos a todas las personas a las que podría interesarles esta vivienda, a contactar con nosotros para realizar un estudio financiero personalizado y concertar una visita sin ningún tipo de compromiso.

por parte de las redes en franquicia del Grupo Tecnocasa por cualquier medio, incluyendo los digitales. (Leer más)

ENVIAR

Características del inmueble

Provincia : Salamanca	Ciudad : Salamanca
Dirección : Calle Miguel de Unamuno, 43	C.P. : 37004
Precio : 108.000 €	Tipología : Piso
Superficie : 79 m2 construidos	Dormitorios : 3
Subtipología : 3 dormitorios	Año de construcción : 1959
Categoría : Media	

OBTENCION
DATOS
INMUEBLE

Equipamiento e instalaciones del inmueble

Baño

- Con ventana (ducha)

Servicios

- Ascensor

Climatización

- Independiente

Instalaciones cercanas

- Jardines públicos (menos de 500 metros)
- Escuela (menos de 500 metros)
- Hospital (menos de 3 Km.)
- Farmacia (menos de 500 metros)
- Supermercado (menos de 500 metros)
- Restaurante (menos de 500 metros)
- Bar (menos de 500 metros)
- Centro comercial (menos de 1 Km.)
- Estación tren (menos de 1 Km.)
- Transporte público (menos de 500 metros)
- Autopista (menos de 3 Km.)
- Aparcamiento público (menos de 500 metros)
- Vías peatonales (menos de 500 metros)
- Carreteras de largo recorrido (menos de 3 Km.)

Cocina

- Reformada

Calefacción

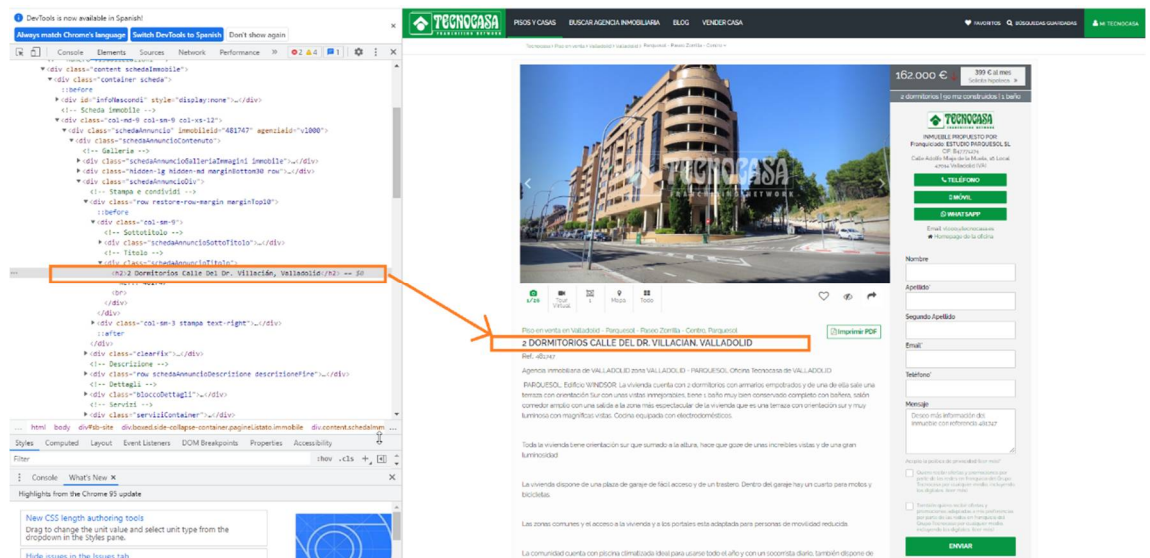
- Independiente (eléctrica)

Transmisión

- Construido - segunda y ulteriores transmisiones

The screenshot displays the Tecnocasa website interface for a property listing. The main content area shows the property details, including the price (108,000 €), location (Salamanca), and various features. A red box highlights the 'Características del inmueble' section. To the right, a browser developer tool is open, showing the HTML structure of the page. A red box in the developer tool highlights the 'Características del inmueble' section, with the text 'RESTO ELEMENTOS.' next to it. The developer tool also shows the 'Equipamiento e instalaciones del inmueble' section, which is highlighted with a red box. The overall layout is clean and professional, with a focus on providing detailed information about the property.

ciudad y dirección para poder identificar unívocamente cada uno de los registros.



6. Agradecimientos. 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 con los principios éticos y legales en el contexto del proyecto.

La idea primigenia para esta práctica era efectuar un spider que permitiese la extracción de datos del portal Idealista, el mayor del país y probablemente de Europa. El esqueleto de ambos scripts se ejecutó varias veces en dicho portal a fin de optimizarlo. El resultado fue un baneo de IP. Si examinamos las condiciones generales de idealista obtenemos lo siguiente:

¿Qué puedes y no puedes hacer en idealista?

Puedes navegar por la Web y Apps, registrarte para guardar búsquedas y favoritos, contactar con anunciantes, publicar inmuebles en venta o alquiler, así como contratar otros servicios adicionales.

“No puedes hacerles daño a terceros ni a idealista, hacer actos contrarios a la Ley, usar mecanismos automáticos para copiar o extraer nuestro contenido, hacer contactos fraudulentos ni utilizar las claves de acceso de otros sin su permiso.

Si eres un anunciante, te recomendamos que leas detalladamente nuestras normas de publicación”.

<https://www.idealista.com/ayuda/articulos/terminos-y-condiciones-generales-de-idealista/>

The screenshot shows the Idealista website's Terms and Conditions page. The browser address bar displays the URL: <https://www.idealista.com/ayuda/articulos/terminos-y-condiciones-generales-de-idealista/>. The page header includes the Idealista logo, a language selector set to 'Español', and navigation links: 'Home', 'Condiciones de uso y privacidad', 'Términos y Condiciones generales de idealista', and a search bar with the text 'Cómo usamos tus datos, términos y conc'. The main heading is 'Términos y Condiciones generales de idealista', with a subtext 'Última actualización: 20 de noviembre, 2020'. Below this is a section titled 'Versión resumida por si vas corto de tiempo' (Summary version in case you are short of time), followed by a paragraph explaining the goal of clarity and transparency. The page is divided into sections: '¿Quiénes somos?' (Who we are), '¿Qué servicios ofrecemos?' (What services we offer), and '¿Qué puedes y no puedes hacer en idealista?' (What you can and cannot do on Idealista). The last section, highlighted with a red border, states that users can navigate the website and use various services, but it explicitly prohibits copying content, making fraudulent contacts, or using automated access tools without permission. It also recommends reading the publication norms if the user is an advertiser.

Términos y Condiciones generales de idealista

Última actualización: 20 de noviembre, 2020

Versión resumida por si vas corto de tiempo

Queremos aportarte claridad y transparencia sobre idealista. Sabemos que los textos legales son importantes aunque farragosos, para hacértelo más sencillo, te contamos brevemente quienes somos, qué servicios te podemos ofrecer y qué puedes o no puedes hacer:

¿Quiénes somos?

idealista (Idealista, S.A.U. nuestro nombre legal) es la empresa que gestiona esta Web y Apps. Nuestra sede está en Madrid, en Plaza de las Cortes (sí, al lado de los famosos leones). Estamos inscritos en el Registro Mercantil de Madrid y tenemos el NIF A82505660.

¿Qué servicios ofrecemos?

Te facilitamos un espacio para que puedas publicar, o buscar, anuncios de venta o alquiler de inmuebles, sea un piso, una habitación en piso compartido o un garaje, por darte unos ejemplos. También te ofrecemos otros servicios relacionados con el sector inmobiliario, como valoraciones de inmuebles o el servicio de certificación energética, para darte una experiencia completa.

¿Qué puedes y no puedes hacer en idealista?

Puedes navegar por la Web y Apps, registrarte para guardar búsquedas y favoritos, contactar con anunciantes, publicar inmuebles en venta o alquiler, así como contratar otros servicios adicionales. No puedes hacerle daño a terceros ni a idealista, hacer actos contrarios a la Ley, usar mecanismos automáticos para copiar o extraer nuestro contenido, hacer contactos fraudulentos ni utilizar las claves de acceso de otros sin su permiso.

Si eres un anunciante, te recomendamos que leas detalladamente nuestras normas de publicación.

Ante esto se reciclo la mayoría del código para emplearlo en otro caso, preferentemente inmobiliario. Así aparece Tecnocasa como opción más válida.

En su portal no encontramos mención alguna al web scripting en la sección legal de su página web. Únicamente encontramos información sobre la relación franquiciador consumidor final a la hora de efectuar las transacciones y las obligaciones contractuales de cada parte, así como una explicación de las cargas fiscales que le corresponden a cada parte y un detalle del proceso.

<https://www.tecnocasa.es/legal/consumidor.html>

Agradecemos esto encarecidamente, primero por la cantidad y precisión de los datos que tiene su portal, por lo representativo que resulta su cartera en el panorama nacional, por permitirnos el acceso y porque no decirlo, en un momento en el que la fecha de entrega de esta PEC era acuciante, se erigía como única y mejor opción.

https://www.tecnocasa.es/legal/consumidor.html

NOCAS
REAL ESTATE NETWORK

PISOS Y CASAS BUSCAR AGENCIA INMOBILIARIA BLOG VENDER CASA FAVORITOS

INFORMACIÓN AL CONSUMIDOR

Las marcas Tecnocasa, Tecnorete y Kiron, son símbolos distintivos sin personalidad jurídica que identifican respectivamente redes en franquicia de intermediarios inmobiliarios y de intermediarios en la obtención de préstamos o créditos, cada uno de los cuales es una persona jurídica autónoma e independiente de su franquiciador así como del resto de los franquiciados y directamente responsable de los actos relacionados con el desarrollo de su actividad profesional y empresarial.

Los precios de venta/arrendamiento de los inmuebles que se publicitan no incluyen los honorarios de intermediación que deberá satisfacer el comprador/arrendatario, ni los tributos que pudieran gravar la compraventa/arrendamiento (ITPAJD o IVA), ni, en caso de haberlos, los notariales, los de gestoría y los registrales, que legalmente le corresponda satisfacer al comprador/arrendatario. Los precios proporcionados están sujetos a variación, para verificarlos le recomendamos que se ponga en contacto con la sociedad anunciante. La superficie de los inmuebles anunciados puede ser útil o construida y en algunos supuestos, aproximada.

De forma orientadora, los honorarios a satisfacer por el comprador corresponderán con el 3% (+ IVA al 21%) del precio de venta del inmueble, y por el arrendatario, el importe equivalente a una mensualidad de renta (+ IVA al 21%). Las sociedades franquiciadas disponen en su punto de venta de un Folleto de Tarifas en el que se detallan y concretan los honorarios que cada una de ellas perciben efectivamente tanto del comprador/arrendatario como del vendedor/arrendador por la prestación de sus servicios.

El derecho de elección de Notario le corresponde al comprador, sin que este pueda imponer un Notario, que, por su competencia territorial, carezca de conexión razonable con alguno de los elementos personales o reales del negocio.

Respectos a los gastos derivados de la formalización o perfección de la compraventa inmobiliaria, el art. 1455 del Código Civil dispone que los gastos de otorgamiento de las escrituras serán de cuenta del vendedor, mientras que los correspondientes a la primera copia y demás posteriores a la venta serán de cuenta del comprador, salvo que se pacte otra cosa atendiendo a la libertad de pacto establecida en el art. 1255 del Código Civil.

El art. 531.6 del Libro V del Código Civil de Cataluña, que dispone que los gastos de otorgamiento de la escritura y de la expedición de la primera copia y los demás gastos posteriores a transmisión corren a cargo de los adquirentes, salvo que una disposición o un pacto establezcan lo contrario.

Si la ley exigiere el otorgamiento de escritura u otra forma especial para hacer efectivos las obligaciones propias de un contrato, los contratantes podrán compelerse recíprocamente a llenar aquella forma desde que hubiere intervenido el consentimiento y demás requisitos necesarios para su validez (art. 1279 Código Civil). Deberán constar en documento público los actos y contratos que tengan por objeto la creación, transmisión, modificación o extinción de derechos reales sobre bienes inmuebles (art.1280.1 Código Civil).

Sobre los inmuebles que publicita, cada sociedad franquiciada tiene en su establecimiento a disposición de los consumidores un documento informativo abreviado y/o ficha informativa: nota explicativa del precio y de la forma de pago de los mismos, información sobre los periodos de validez de los anuncios de venta o alquiler, información sobre los periodos de duración de los contratos de intermediación inmobiliaria y los modelos de los mismos que utiliza en la prestación de sus servicios así como información sobre los procedimientos de los que dispone el consumidor para poner fin a los mismos y la preceptiva información sobre eficiencia energética, toda vez que esta haya sido facilitada al intermediario por la propiedad.

Se han realizado todos los esfuerzos para que la información proporcionada sobre los inmuebles sea completa, exacta y actualizada, no obstante, dicha información puede

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.

La inspiración para la realización de este spider nace de mi experiencia laboral en el mundo inmobiliario, y de la falta de datos existente en la estadística públicas de los datos por parte de la oferta. El INE recopila información a los promotores-constructores de precio de venta de sus propiedades. Muchas veces he tenido que realizar estas estadísticas, así como consignar en ellas buena parte de los estados contables de las empresas para las que trabajaba. Por esta razón me consta que la información consignada no es especialmente correcta ni detallada. Ya que el objetivo de esta es el cálculo de un precio medio del m2. Es evidente que esta valoración depende mucho de la oferta que tenga la empresa a la que se le demanda los datos. Es una cuestión de heterogeneidad de la cartera por tipología de producto, así como por ubicación de este. Toda esta información se obtiene en el caso de obra nueva, no recoge la de segunda mano ni la cartera de transformación, esto es, todos aquellos solares o productos que están en curso de cambio de uso administrativo. Así mismo la otra gran fuente de información del mercado inmobiliario se obtiene por la vía de notario y registradores públicos y basada en las operaciones que se han cruzado. Esta información también presenta deficiencias, en ocasiones no recoge el valor total por el que se efectúa la transmisión. No presenta la debida concreción y en ocasiones se mezclan valoraciones de

mercado con las que se emplean a efectos tributarios. Nuestro spider se concentra en medir las expectativas de precio en el momento actual por parte de la oferta.

En este entorno el spider permitiría, obtener todos los datos relevantes de una propiedad inmobiliaria.

Esta información serviría para realizar estadísticas de todo tipo, como tamaño medio de la vivienda en términos absolutos, como tamaños por número de habitaciones, si puedo incluir la antigüedad se podría calcular la edad media del parque de vivienda de segunda mano disponible, asimismo con el precio del m2 tanto absoluto como para cada tipo de casa podemos realizar tasaciones en función de la oferta de la zona, factor este empleado hasta por los propios arquitectos a la hora de justificar valoraciones para la concesión de hipotecas.

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar 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.**

"Reconocimiento-No Comercial-Compartir Igual CC BY-NC-SA

Esta licencia permite a otros modificar, adaptar y construir sobre su trabajo de forma no comercial, siempre que le otorguen crédito y licencian sus nuevas creaciones bajo los mismos términos".

seleccionaría esta licencia ya que es la más laxa de todas. Siguiendo la mentalidad de software de desarrollo libre, permitiría a otros desarrollar este trabajo siempre y cuando lo haga en esas condiciones y evidentemente sin pretensiones comerciales. He adquirido los datos de un tercero sin solicitarle un permiso expreso y se ha modificado para un uso didáctico lo lógico y más ético es que otro lo empleen con semejante propósito.

En Zenodo esta opción no esta disponible razón por la que selecciono Creative Commons Attribution 4.0 Internacional, por las mismas razones que las mencionadas para la anterior.

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

LINK GOOGLE-DRIVE:

<https://drive.google.com/file/d/1kc0f8VE3qZ7SxkUSDFwK3fSkaOX643Jj/view?usp=sharing>

LINK GIT-HUB

10. Dataset. Publicar el dataset obtenido (*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

<https://zenodo.org/record/5650725#.YYZM022ZOUk>

md5:62eda0b218886dd27065afca74749172 salamanca

md5:fe13092d564bda3cc39e5bf314e1630c malaga

(PRACTICA 1 TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS COPIA DE LA PRE-ENTREGA)-ANEXO

En primer lugar, quiero mencionar que la practica la voy a realizar solo, no he encontrado ningún compañero.

En esta práctica quería desarrollar un spider para extraer información sobre propiedades inmobiliarias. En primer lugar, lo intente en el mayor portal que es idealista, pero esa página tiene unas restricciones al web scraping muy importante y me han baneado creo que definitivamente. Todo lo que había realizado ya no sirve para nada. Eso mismo me ha pasado con otras webs deportivas, fabricantes de bicicletas donde he intentado conectarme, para efectuar un catálogo completo de sus productos con resultado no satisfactorio.

Probadas algunas más, retomé la idea de los inmuebles y al final di con la web de Tecnocasa.


La idea es muy sencilla, hacer scraping horizontal y vertical a un nivel.

Con el scraping horizontal vemos todas los inmuebles para el área designado, en este caso Salamanca mi ciudad, nos da un total de dos páginas, pero podría paginar muchas más.

Con el scraping vertical entramos dentro de cada anuncio, donde veremos todos los datos relativos al mismo:

Ciudad, provincia, calle, cp., precio, numero de dormitorios, m2 construidos.

Además, se incluirá el franquiciado que lo vende, su dirección completa.



PISOS Y CASAS

BUSCAR AGENCIA INMOBILIARIA

LOGO

VENDER CASA

Tecnocasa + Inmuebles en venta + Salamanca

Verta

Todas las tipologías

SALAMANCA (PROVINCIA)


Precio

Superficie

PISOS Y CASAS EN SALAMANCA PROVINCIA

Se han encontrado 21 inmuebles en venta en Salamanca provincia

Ordenar por




3 DORMITORIOS CALLE MIGUEL DE UNAMUNO, SALAMANCA 108.000 €

Piso en venta en Salamanca

3 Dormitorios | 79 m² construidos | 1 Baño

Agencia inmobiliaria de Salamanca, Camillo Norte. Oficina

Tecnocasa vende: Piso en Calle Miguel de Unamuno, cerca de la Estación de tren y perpendicular a María Auxiliadora El inmueble cuenta con setenta y nueve metros cuadrados construidos distribuidos en hall de...




3 DORMITORIOS CALLE VASCO DE GAMA, SALAMANCA 132.000 €

Piso en venta en Salamanca

3 Dormitorios | 93 m² construidos | 1 Baño

Agencia inmobiliaria de Salamanca, Camillo Sur: Oficina Tecnocasa vende: Piso en Calle Vasco de Gama, junto a Avenida María Auxiliadora y Avenida Portugal El inmueble cuenta con tres dormitorios, balcón, baño completa, cocina y salón-comedor. Se trata de una...




4 DORMITORIOS PASEO DE LA ESTACIÓN, SALAMANCA 129.000 €

Piso en venta en Salamanca


4 Dormitorios | 134 m² construidos

TECNOCASA Inmobiliaria Salamanca Camillo Sur - Estación vende en exclusiva amplio piso en pleno Paseo de la Estación, junto a la estación de tren y a cinco minutos del centro de Salamanca El inmueble cuenta con 112 metros cuadrados construidos distribuidos en hall...




DETALLES

CONTACTA




DETALLES

CONTACTA



DETALLES

CONTACTA



Piso en venta en Salamanca

3 DORMITORIOS CALLE MIGUEL DE UNAMUNO, SALAMANCA

Ref. 48792g

Agencia Inmobiliaria de Salamanca, Garrido Norte, Oficina Tecnocasa:

Vende: Piso en Calle Miguel de Unamuno, cerca de la Estación de tren y perpendicular a María Auxiliadora.

El inmueble cuenta con setenta y nueve metros cuadrados construidos distribidos en hallé de entrada, amplio salón-comedor, tres dormitorios, cuarto de baño completo y cocina completamente amueblada y equipada con posibilidad de instalar lavavajillas (dispone de toma de agua y desagüe) con solado a gresita.

La vivienda se encuentra en un gran espacio de conservación, cuenta con doble ventana, suelo cerámico y paredes recientemente pintadas. La calefacción eléctrica por acumuladores y el edificio dispone de la gama de gas natural.

Todos los dormitorios son exteriores además del salón, la vivienda hace chaffán Calle Miguel de Unamuno con Calle Gargabete por lo que disfruta de una espectacular luminosidad, hecho que se multiplica gracias a la altura que tiene la vivienda, una tercera planta.

El dormitorio principal dispone de un armario empotrado forrado por dentro de madera con cajoneras, además séde la cocina dispónese de una terraza cerrada que da al patio de Luces, ideal para guardar productos de limpieza y en este caso es el lugar donde se encuentra la lavadora de la vivienda.

La finca en la que se encuentra la vivienda ha sido completamente rehabilitada recientemente, mejorando su accesibilidad desde el portal, asessor completamente nuevo, escaleras modificadas, ventanas de aluminio climatí en las zonas

108.000 € al mes
Cocina completa

3 dormitorios | 79 m² construidos | 1 baño

TECNOCASA
INMOBILIARIA DE SALAMANCA

INMUEBLE PROPUESTO POR
FRANQUEADO ESTUDIO GARRIDO - SUR ESTAR
SL

C/F: Bzzyzzioo
Calle Oro 2
37004 Salamanca (SA)

TELÉFONO

MÓVIL

WHAT APP

Email: sanoo@tecnocasa.es
■ Homepage de la oficina

7. OBTENCION DATOS FRANQUICIADO.

La finca en la que se encuentra la vivienda ha sido completamente rehabilitada recientemente, mejorando su accesibilidad desde el portal, ascensor completamente nuevo, escaleras modificadas, ventanas de aluminio cimtal en las zonas comunes, fachada y tejado también actualizados. Los gastos de comunidad ascienden a 53 €/mes.

En Tecnocasa facilitamos el acceso a la financiación de nuestros clientes, a través de KIRON, la empresa de intermediación financiera del grupo Tecnocasa, con convenios nacionales con las principales entidades de España. Podemos llegar a conseguir hasta el 100% del valor de compraventa. Por eso, invitamos a todas las personas a las que podría interesarles esta vivienda, a contactar con nosotros para realizar un estudio financiero personalizado y concertar una visita sin ningún tipo de compromiso.

por parte de las entidades con franquicia del Grupo Tecnocasa por cualquier medio, incluyendo los digitales (por email)

ENVIAR

Características del inmueble

Provincia: Salamanca Ciudad: Salamanca
Dirección: Calle Miguel de Unamuno, 43 C.P.: 37004
Precio: 108.000 € Tipología: Piso
Superficie: 79 m2 construidos Dormitorios: 3
Subtipología: 3 dormitorios Año de construcción: 1959
Categoría: Media

OBTENCION
DATOS
INMUEBLE

Equipamiento e instalaciones del inmueble

Baño
- Con ventana (ducha)

Servicios
- Ascensor

Cocina
- Reformada

Calefacción
- Independiente (eléctrica)

Climatización
- Independiente

Instalaciones cercanas
- Jardines públicos (menos de 500 metros)
- Escuela (menos de 500 metros)
- Hospital (menos de 3 Km.)
- Farmacia (menos de 500 metros)
- Supermercado (menos de 500 metros)
- Restaurante (menos de 500 metros)
- Bar (menos de 500 metros)
- Centro comercial (menos de 1 Km.)
- Estación tren (menos de 1 Km.)
- Transporte público (menos de 500 metros)
- Autoista (menos de 3 Km.)
- Aparcamiento público (menos de 500 metros)
- Vías peatonales (menos de 500 metros)
- Carreteras de largo recorrido (menos de 3 Km.)

Transmisión
- Construido - segunda y ulteriores transmisiones

empresa de intermediación financiera del grupo Tecnocasa, con convenios nacionales con las principales entidades de España. Podemos llegar a conseguir hasta el 100% del valor de compraventa. Por eso, invitamos a todas las personas a las que podría interesarles esta vivienda, a contactar con nosotros para realizar un estudio financiero personalizado y concertar una visita sin ningún tipo de compromiso.

108.000 €
266 € al mes
Sólo hipoteca

3 dormitorios | 79 m2 construidos | 1 baño

Características del inmueble

Provincia: Salamanca Ciudad: Salamanca
Dirección: Calle Miguel de Unamuno, 43 C.P.: 37004
Precio: 108.000 € Tipología: Piso
Superficie: 79 m2 construidos Dormitorios: 3
Subtipología: 3 dormitorios Año de construcción: 1959
Categoría: Media

Equipamiento e instalaciones del inmueble

Baño
- Con ventana (ducha)

Servicios
- Ascensor

Cocina
- Reformada

Calefacción
- Independiente (eléctrica)

Climatización
- Independiente

Instalaciones cercanas
- Jardines públicos (menos de 500 metros)
- Escuela (menos de 500 metros)
- Hospital (menos de 3 Km.)
- Farmacia (menos de 500 metros)
- Supermercado (menos de 500 metros)
- Restaurante (menos de 500 metros)
- Bar (menos de 500 metros)
- Centro comercial (menos de 1 Km.)
- Estación tren (menos de 1 Km.)
- Transporte público (menos de 500 metros)
- Autoista (menos de 3 Km.)
- Aparcamiento público (menos de 500 metros)
- Vías peatonales (menos de 500 metros)
- Carreteras de largo recorrido (menos de 3 Km.)

Transmisión
- Construido - segunda y ulteriores transmisiones

CONTENEDOR INFORMACION FRANQUICADO.

empresa de intermediación financiera del grupo Tecnocasa, con convenios nacionales con las principales entidades de España. Podemos llegar a conseguir hasta el 100% del valor de compraventa. Por eso, invitamos a todas las personas a las que podría interesarles esta vivienda, a contactar con nosotros para realizar un estudio financiero personalizado y concertar una visita sin ningún tipo de compromiso.

108.000 €
266 € al mes
Sólo hipoteca

3 dormitorios | 79 m2 construidos | 1 baño

Características del inmueble

Provincia: Salamanca Ciudad: Salamanca
Dirección: Calle Miguel de Unamuno, 43 C.P.: 37004
Precio: 108.000 € Tipología: Piso
Superficie: 79 m2 construidos Dormitorios: 3
Subtipología: 3 dormitorios Año de construcción: 1959
Categoría: Media

Equipamiento e instalaciones del inmueble

Baño
- Con ventana (ducha)

Servicios
- Ascensor

Cocina
- Reformada

Calefacción
- Independiente (eléctrica)

Climatización
- Independiente

Instalaciones cercanas
- Jardines públicos (menos de 500 metros)
- Escuela (menos de 500 metros)
- Hospital (menos de 3 Km.)
- Farmacia (menos de 500 metros)
- Supermercado (menos de 500 metros)
- Restaurante (menos de 500 metros)
- Bar (menos de 500 metros)
- Centro comercial (menos de 1 Km.)
- Estación tren (menos de 1 Km.)
- Transporte público (menos de 500 metros)
- Autoista (menos de 3 Km.)
- Aparcamiento público (menos de 500 metros)
- Vías peatonales (menos de 500 metros)
- Carreteras de largo recorrido (menos de 3 Km.)

Transmisión
- Construido - segunda y ulteriores transmisiones

RESTO ELEMENTOS.

No he podido avanzar mucho más en esta práctica, por problemas de tiempo.

Al extraer información del contenedor de la propiedad, estoy haciendo algo mal porque obtengo un primer campo en blanco, que no soy capaz de eliminar directamente modificando la ruta, ni en la fase de preprocesado. Confío en solucionarlo en breve.

Se acompaña el archivo tecnocasa.py con el código desarrollado hasta el momento y debidamente comentado, como mejor esquema posible de todo lo desarrollado hasta el momento. Hay que ejecutarlo desde consola y en la última línea esta comentado la instrucción a ejecutar.

COPIA MENSAJE CON LAS INDICACIONES DE MEJORA.



Laia Subirats Maté

to me

Mon, Nov 1, 6:55 AM (5 days ago) ☆ ↶ ⋮

Buenos días David,

Creo que vas por buen camino en la práctica. A continuación te doy algunos comentarios de mejora:

- La página web escogida es correcta, pero no usas métodos más avanzados como selenium, prevención de web scraping, ni recorres la página web mediante clics, etc. mencionados en el tablón de la asignatura. Si quieres puedes hacer una extensión de la práctica usando estos métodos para practicar.
- Quizá puedes especificar qué versión de las librerías has usado, mira la función freeze https://pypi.org/en/stable/cli/robo_freeze/
- Es muy importante que revises el aspecto legal del web scraping (archivo robots, etc.)

Un saludo y ¡ánimo!

Laia