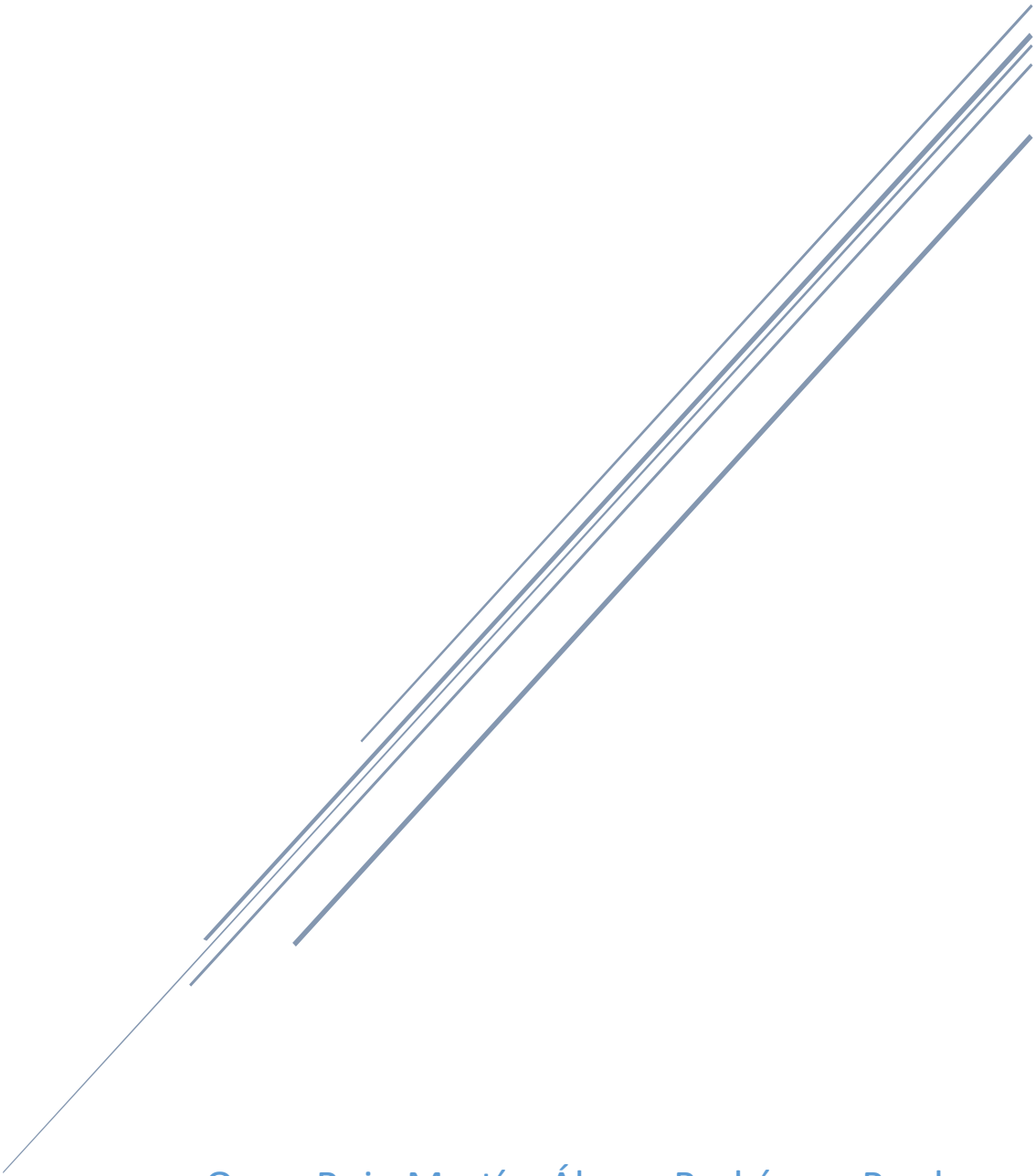
Universitat Oberta de Catalunya

uoc.edu

M2.851 - TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRA1



Oscar Rojo Martín, Álvaro Rodríguez Pardo

Índice

Tablas

Ilustraciones

Puntos para desarrollar

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

Para la realización de web scraping se ha elegido la página de la Diputación Foral de Gipuzkoa (https://www.gipuzkoa.eus/es), la cual es es el órgano de gobierno del territorio histórico de Gipuzkoa, País Vasco. La información que presenta este sitio web viene estructurada en forma de categorías definidas en un archivo XML, el cual puede encontrarse en la siguiente página: https://www.gipuzkoa.eus/sitemap.xml. Dentro de la página principal, se han escogido dos páginas web que contienen los datos de las subastas de propiedades que están pendientes (https://www.gipuzkoa.eus/es/web/ogasuna/subastas) y los datos de las subastas que ya han sido celebradas (https://www.gipuzkoa.eus/es/web/ogasuna/subastas/celebradas). Dichos datos se encuentran estructurados en dos tablas, las cuales serán las que se obtendrán a partir del scraper y serán guardadas en un archivo CSV.

1.1. robots.txt

Antes de realizar web scraping sobre la página se ha procedido a estudiar el archivo robots.txt (https://www.gipuzkoa.eus/es/robots.txt) para comprobar las limitaciones que podemos tener a la hora de extraer los datos. En este sentido, un archivo robots.txt indica a los rastreadores de los buscadores qué páginas o archivos del sitio se pueden solicitar y cuáles no. Principalmente, se utiliza para evitar que las solicitudes que recibe el sitio lo sobrecarguen; no es un mecanismo para impedir que una página web aparezca en Google. Si lo que se busca es esto último, se debe usar directivas *noindex* o proteger esas páginas con contraseña.
Pasamos a ver los bots que la página tiene bloqueados:

	User-agent	Status	Pattern
0	MauiBot	Disallow	/
1	AhrefsBot	Disallow	/
2	DotBot	Disallow	/
3	SemrushBot	Disallow	/
4	MJ12bot	Disallow	/
5	Seekport	Disallow	/
6	Seekport	Disallow	/es/resultados-buscador
7	Seekport	Disallow	/eu/bilaketaren-emaitzak
8	Seekport	Disallow	/*DLF_Bilatzailea
9	Seekport	Disallow	/*DLYCrossSiteRequestProxy-portlet
10	Seekport	Disallow	/*DLYServices-portlet
11	Seekport	Disallow	/*notifications-portlet
12	Seekport	Disallow	/*calendar-portlet
13	Seekport	Disallow	/*buscar
14	Seekport	Disallow	/*INSTANCE
15	Seekport	Disallow	/*combo
16	Seekport	Disallow	/*busqueda
17	Seekport	Disallow	/*bilaketa
18	Seekport	Disallow	/*galeria-bektoriala
19	Seekport	Disallow	/*asset_publisher

Tabla 1: Dataframe con los bots bloqueados por la página web

Aunque la página no prohíbe el uso de bots, nos hemos asegurado de que usamos un *User Agent* real para pasar más inadvertidos y no tener ningún tipo de problema con estos bots bloqueados.

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

En este trabajo se han creado dos conjuntos de datos:

- Subastas Pendientes (subastas_pendientes.csv), donde se encuentra la tabla con las subastas que aún tienen que realizarse.
- Subastas Resueltas (subastas_resueltas.csv), donde está la tabla que recoge los datos de las subastas que ya han sido celebradas.

Ambos datasets están incluidos en la carpeta *data*, la cual está dentro de la carpeta code, del repositorio de GitHub.

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).

Como se introdujo en el contexto, los dos conjuntos de datos extraídos hacen referencia a las subastas pendientes y a las ya celebradas por la Diputación Foral de Gipuzkoa, cuyo Departamento de Hacienda y Finanzas realiza subastas de bienes muebles e inmuebles. Ambos datasets contienen una tabla con el mismo número de variables, que son nueve (url, número, tipo, lugar, descripción, importe, fecha, procedimiento y situación), las cuales veremos con más detenimiento en el punto 5. Si uno se fija en la página web, verá cómo la tabla de subastas tiene menos variables de las que se han recogido en el conjunto de datos obtenido, además de tener una llamada “Detalles” en la cual aparece un enlace con información sobre cada una de las subastas. Si se *clicka* en dicho enlace se podrá ver con todo lujo de detalles la subasta, y es precisamente en esta página desde donde se han obtenido variables cómo la descripción, el procedimiento de enajenación y la fecha, esta última estando disponible en la tabla principal, pero viéndose con más detenimiento en la web de detalles.
Como matiz final hay que añadir que, si todas las subastas han sido celebradas, la página web desde donde se puede ver la tabla con las subastas pendientes no tendrá datos, pues todas se han resuelto hasta el momento.

Subastas

El Departamento de Hacienda y Finanzas de la Diputación Foral de Gipuzkoa, muestra en el siguiente listado las subastas celebradas.

Subasta	Tipo de bien	Municipio	Importe	Fecha	Situación	Detalles
228	Garaje	Donostia	13.140,00 €	26/03/2021	Adjudicada	+info
227	Garaje	Donostia	16.060,00 €	26/03/2021	Adjudicada	+info
225	Garaje	Donostia	19.150,00 €	26/03/2021	Adjudicada	+info
224	Garaje	Donostia	19.150,00 €	26/03/2021	Adjudicada	+info
230	Rústica	Ondarroa	23.991,00 €	24/03/2021	Anulada	+info
229	Rústica	Ondarroa	10.009,00 €	24/03/2021	Anulada	+info
226	Garaje	Donostia	16.020,00 €	24/03/2021	Adjudicada	+info
223	Local	Donostia	72.626,40 €	24/03/2021	Suspendida por pago	+info
222	Local	Donostia	25.938,00 €	24/03/2021	Suspendida por pago	+info
221	Pabellón industrial	Azpeitia (Gipuzkoa)	751.340,00 €	27/11/2020	Desierta	+info
220	Vivienda	Lasarte - Oria (Gipuzkoa)	101.768,59 €	25/11/2020	Suspendida por pago	+info
219	Trastero	Beasain (Gipuzkoa)	6.165,76 €	25/11/2020	Adjudicada	+info
218	Garaje	Hondarribia (Gipuzkoa)	30.105,05 €	25/11/2020	Adjudicada	+info
217	Trastero	Beasain	2.270,00 €	06/11/2020	Desierta	+info
216	Trastero	Beasain	5.260,00 €	06/11/2020	Desierta	+info
215	Trastero	Beasain	3.060,00 €	06/11/2020	Desierta	+info
214	Trastero	Beasain	4.880,00 €	06/11/2020	Desierta	+info
213	Terreno	Estepona	257.740,00 €	06/11/2020	Adjudicada	+info

Ilustración 1: Tabla de subastas desde donde se han recogido los datos

Detalle de la subasta

Nombre: 228

Tipo de bien: Garaje

Municipio: Donostia

Descripción: Plaza de aparcamiento nº 19 en la planta de semisotano del conjunto residencial integrado por tres casas señaladas con los números cinco, siete y nueve de la calle Eguzki Eder. 13 m2. Trastero anejo número diecinueve, 7,79 m2

Importe inicial: 13.140,00 €

Fecha de la subasta: Primera licitación, 24/03/2021, y segunda licitación: 26/03/2021

Procedimiento de enajenacion: Subasta: Primera Licitación, procedimiento de presentación ofertas en sobre cerrado; y, segunda licitación mediante sucesivas pujas de viva voz.

Situación: Adjudicada

Ilustración 2: Ejemplo de la información que aparece en los enlaces que contiene la columna "Detalles"

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

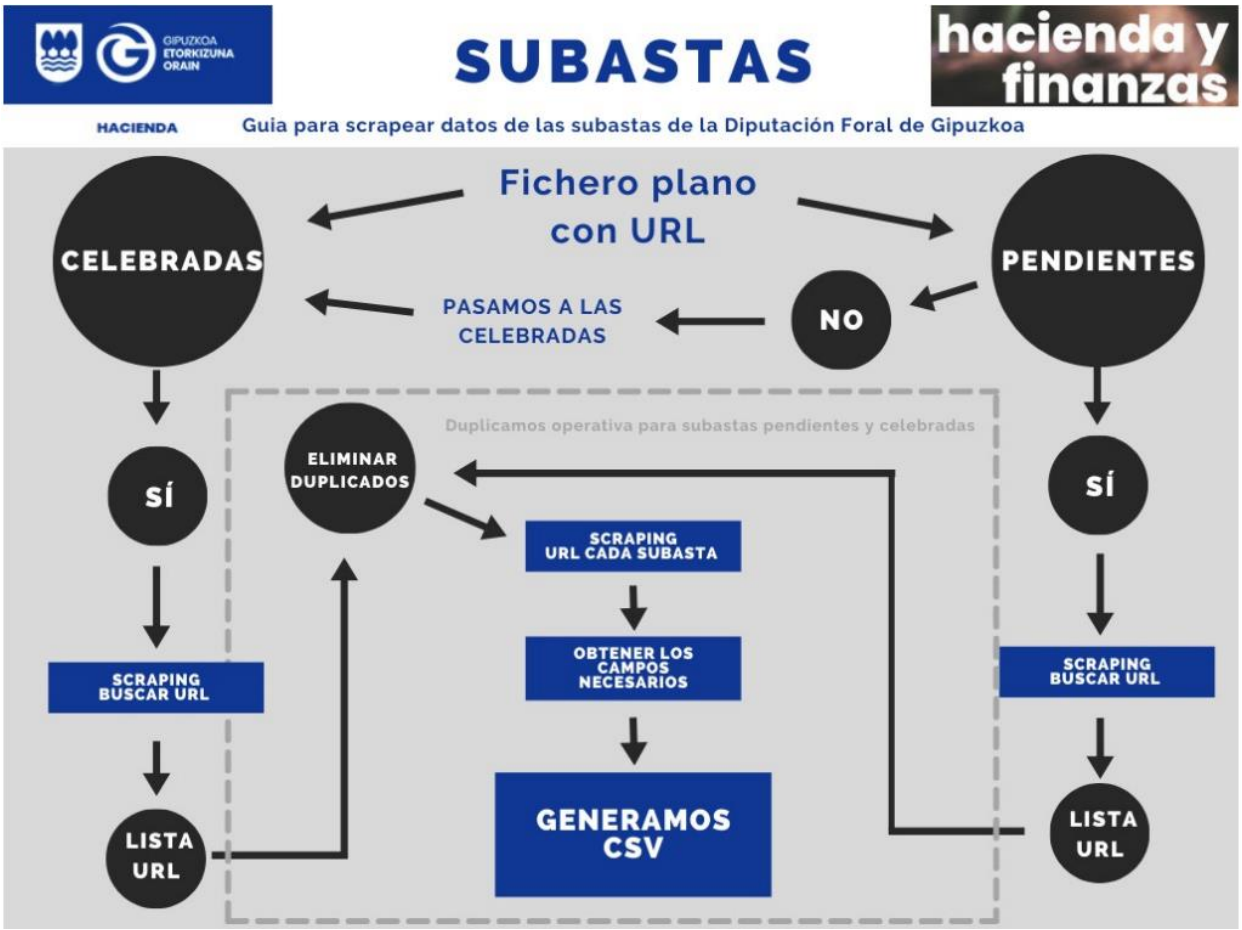


Ilustración 3: Diagrama de flujo sobre el proyecto

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

Pasamos a ver una tabla con los campos que tiene cada uno de los dos conjuntos de datos obtenidos:

Campo	Tipo	Descripción
url	String	URL desde donde se pueden ver los detalles de cada subasta.
numero	Integer	Número de la subasta.
tipo	String	Tipo de bien que se está subastando.
lugar	String	Municipio donde se encuentra el bien subastado.
descripcion	String	Descripción del bien subastado.
importe	String	Importe, en euros, del bien subastado.
fecha	String	Fecha en la que se produjo la subasta.
procedimiento	String	Procedimiento de enajenación o de transferencia del bien subastado.
situacion	String	Situación en la que se encuentra la subasta.

Tabla 2: Campos que incluyen los datasets

El período de tiempo de los datos de subastas que contiene la página web abarca desde el año 2014 hasta el presente año 2021. Pasamos a continuación a ver cómo se han recogido dichos datos para la obtención de los dos archivos CSV, cuyo procedimiento de recopilación ha tenido el siguiente orden:

1) Se obtiene la URL raíz de la diputación foral (contendida en un archivo plano).
2) Se sustituye la URL por la raíz necesaria.
3) Se recorre la raíz inicial en busca de la URL "hijos".
4) Se recopilan las diferentes URLs en una lista que se ha de limpiar de datos innecesarios y elementos duplicados.
5) Se recorre la lista de hijos en busca de nuevas URLs.
6) Se recopilan las nuevas URLs y se realiza una vez más una limpieza de la nueva lista.
7) De esta última lista de URLs donde se detallan cada uno de los productos, se realiza el scraping.
8) Finalmente, se convierte el diccionario en un dataframe y este en un archivo CSV.

Veamos a continuación dos imágenes con los datasets obtenidos en formato CSV:

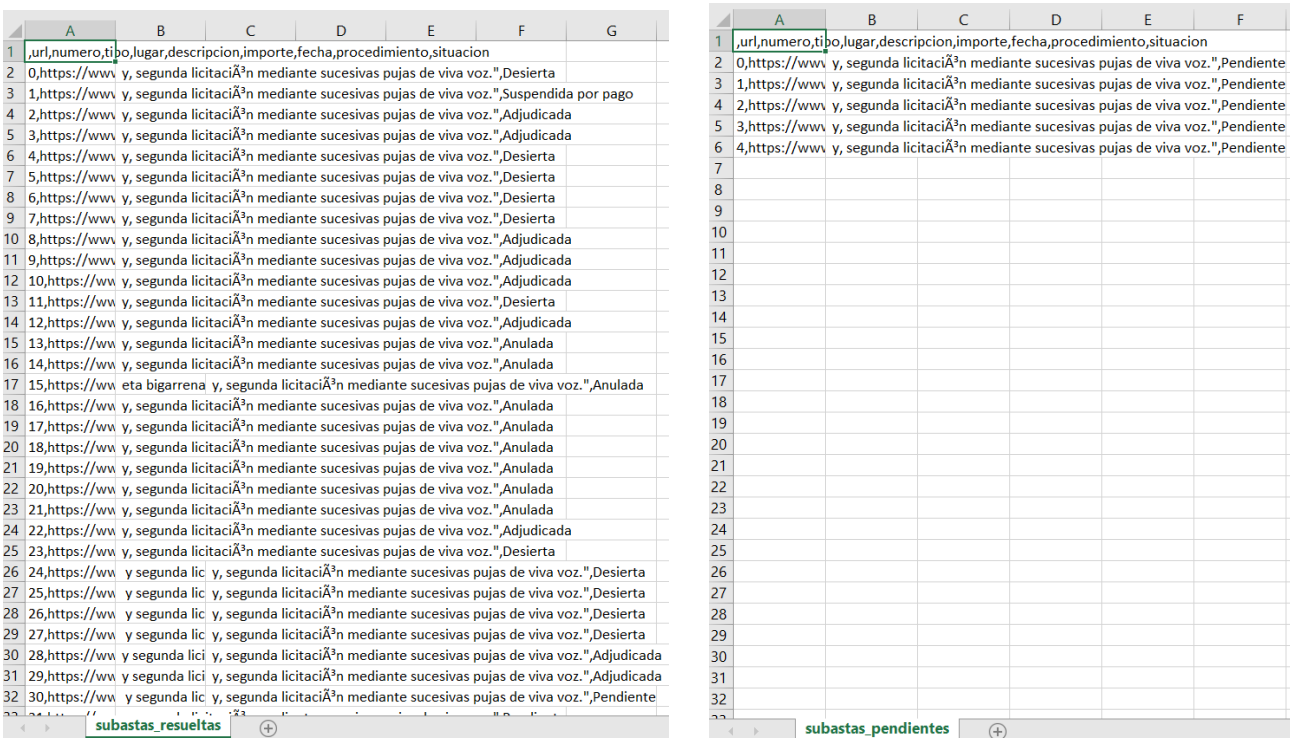


Ilustración 4: A la izquierda, las subastas celebradas y a la derecha, las subastas pendientes

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.

El propietario del conjunto de datos es la Diputación Foral de Gipuzkoa (https://www.gipuzkoa.eus/es/), concretamente el Departamento de Hacienda y Finanzas (Ogasuna), así que los agradecimientos van dirigidos a esta institución por proporcionar unos datos tan interesantes de forma pública. En cuanto a investigaciones o análisis anteriores, no se ha encontrado ninguno que haga referencia a los datos escogidos.

7. Inspiración. Explique 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 presente proyecto se inspira en la primera publicación que realizó la Diputación Foral de Gipuzkoa en los portales inmobiliarios de internet como "Idealista" y "Fotocasa", donde se informaba de las subastas. Es un conjunto de datos interesante porque contiene las subastas realizadas en distintos municipios del País Vasco con todo lujo de detalles y podrá responder preguntas del tipo:

- ¿Cuál ha sido el importe máximo y mínimo al que se ha subastado un bien?
- ¿Qué tipo de bien es el más subastado en el País Vasco? ¿Y el menos?
- ¿En qué municipios se realizan más y menos subastas?

- ¿En qué fechas se han realizado más y menos subastas? ¿El por qué de esto se puede explicar por la situación económica de España en esos años? Obviamente, esta última pregunta requeriría de un poco de investigación propia ya que el conjunto de datos en sí mismo no puede explicar la evolución económica del país.

Si observamos el dataset con detenimiento, se puede comprobar que muchas de las subastas que hay en este han quedado desiertas. Habría que valorar si el motivo fue el precio, las condiciones del inmueble o la falta de publicidad del evento.

Finalmente, no se realizarán comparaciones con los análisis anteriores ya que, como se explicó en el apartado 6, no se ha encontrado ningún otro estudio similar.

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*

Se ha elegido la licencia CC BY-NC-SA 4.0 ya que esta no permite un uso comercial de la obra original ni de las posibles obras derivadas, pues consideramos que al ser un proyecto de carácter académico no se debe ser utilizado con fines comerciales. Además, la distribución de estas obras derivadas se debe hacer con una licencia igual a la que regula la obra original, con lo cual, nos aseguramos de que se nos dé el crédito por el conjunto de datos obtenido.
Dicha licencia se puede leer detenidamente en el repositorio que contiene el proyecto realizado (https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/LICENSE.md)

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

El código para realizar el web scraping se ha escrito en lenguaje Python y se encuentra en la carpeta *code* con el nombre de *diputación.py* (https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/code/diputacion.py). Además, también se presentan otros dos códigos, uno para estudiar el archivo robots.txt, en donde se define una función genérica para obtener un dataframe con aquellos bots que están bloqueados por una página web, el cual también se puede encontrar en la carpeta *data* con el nombre de *utils.py* (https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/code/diputacion.py), y otro realizado para aplicar dicha función

a este proyecto en específico, alojado en la misma carpeta y llamado *robots.py* (https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/code/robots.py).

Para la correcta ejecución del código de web scraping se recomienda lo siguiente:

1) Generar una carpeta:

```
$ mkdir -directorio
```

2) Generar un entorno virtual. Cómo hacerlo:

En Linux:

```
$ python3 -m venv /path/to/new/virtual/environment
```

En Windows:

```
c:\>c:\Python35\python -m venv c:\path\to\myenv
```

3) Instalar los módulos necesarios detallados en el fichero *requirements.txt*:

```
$ pip install requirements.txt
```

4) Finalmente ejecutar el scraping sobre las subastas de la Diputación:

```
$ python code/diputacion.py
```

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Se ha publicado el conjunto de datos en formato CSV en Zenodo donde ha añadido una pequeña descripción del proyecto y también el enlace al repositorio de GitHub para quien desee ampliar información sobre cómo se han recogido los datos. El enlace es el siguiente: https://doi.org/10.5281/zenodo.4662752. El DOI obtenido se ha añadido al final del README.md (https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/README.md) que se encuentra en el repositorio.

Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	ORM, ARP
Redacción de las respuestas	ORM, ARP
Desarrollo código	ORM, ARP