

Oscar Rojo Martín, Álvaro Rodríguez Pardo

zumaia@uoc.edu, alvarorp22@uoc.edu

m2.851 - Tipología y ciclo de vida de los datos

PRA1

Índice

[Puntos para desarrollar 1](#_Toc68439948)

[1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. 1](#_Toc68439949)

[1.1. ***Robots.txt*** 1](#_Toc68439950)

[2. Definir un títulopara el dataset. Elegir un título que sea descriptivo. 1](#_Toc68439951)

[3. Descripcióndel dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido). 1](#_Toc68439952)

[4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente. 1](#_Toc68439953)

[5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. 1](#_Toc68439954)

[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. 2](#_Toc68439955)

[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. 2](#_Toc68439956)

[8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: 2](#_Toc68439957)

[9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. 2](#_Toc68439958)

[10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción. 2](#_Toc68439959)

# Puntos para desarrollar

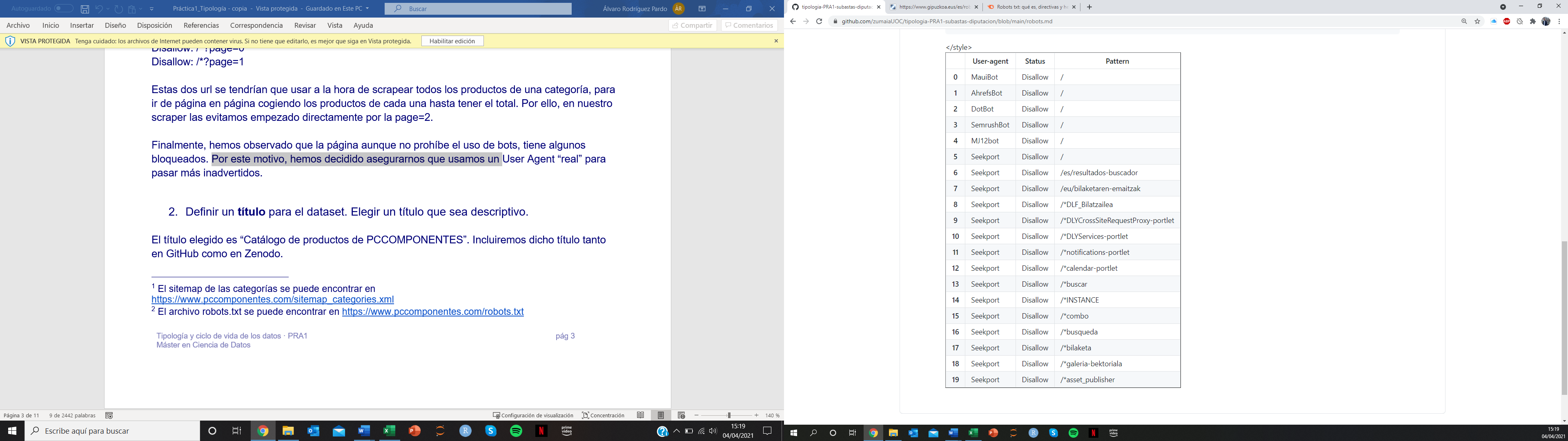
## Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

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. Además, hay una columna llamada “Detalles” en la cual aparece un enlace con información sobre cada una de las subastas. Esta información también se recogerá para disponer de una tabla con todo lujo de detalles y se proporcionarán dichos links desde donde ha sido extraída dicha información.

* 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:



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 problemas con los bots bloqueados.

## Definir un títulopara el dataset. Elegir un título que sea descriptivo.

## Descripcióndel dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

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

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

## 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.

## 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.

## 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*

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

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