

M2.851

Tipología y ciclo de vida de los datos

PRA1

Alumnos: Víctor Colomé y Carlos Marcos

Contenido

Puntos a desarrollar en la práctica	3
1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.	3
2. Definir un título para el dataset. Elegir un título que sea descriptivo.	3
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).	4
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente	4
5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.	6
6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).	8
7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.	8
8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:	9
9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.....	9
10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.	10
11. Entrega. Presentar el trabajo con el DOI del dataset en Github.....	10
Checklist de contribuciones del equipo	11

Puntos a desarrollar en la práctica

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

El sitio web elegido para practicar el web scraping es www.pccomponentes.es, una conocida web española de venta online de electrónica de consumo. Como tal, el sitio web está muy orientado a la venta y proporciona un completo catálogo de sus productos agrupados por distintas categorías (smartphones, portátiles, etc).

La información que presenta el sitio web tiene una cierta estructura en la forma de categorías definidas en el fichero sitemap_categories.xml¹, lo cual puede ayudar a diseñar de manera más efectiva el scraping. Pero una vez hecho esto, debe descargarse cada categoría y artículo de la misma por separado, lo cual sí entraña cierta dificultad adicional. Es decir, que a juicio de los alumnos, el sitio web elegido presenta un equilibrio entre idoneidad y dificultad adecuado para la realización de la práctica.

1.1. *Robots.txt*

Antes de elegir la página de pccomponentes para desarrollar la práctica, observamos su archivo robots.txt² para comprobar si había alguna limitación o recomendación.

Lo primero que encontramos fue la url para los diferentes sitemaps, incluido el de las categorías. De entre todas las url protegidas o prohibidas (por problemas técnicos) encontramos dos url que afectan directamente al scraper:

Disallow: /*?page=0

Disallow: /*?page=1

Estas dos url se tendrían que usar a la hora de scrapear todos los productos de una categoría, para ir de página en página cogiendo los productos de cada una hasta tener el total. Por ello, en nuestro scraper las evitamos empezado directamente por la page=2.

Finalmente, hemos observado que la página aunque no prohíbe el uso de bots, tiene algunos bloqueados. Por este motivo, hemos decidido asegurarnos que usamos un User Agent “real” para pasar más inadvertidos.

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

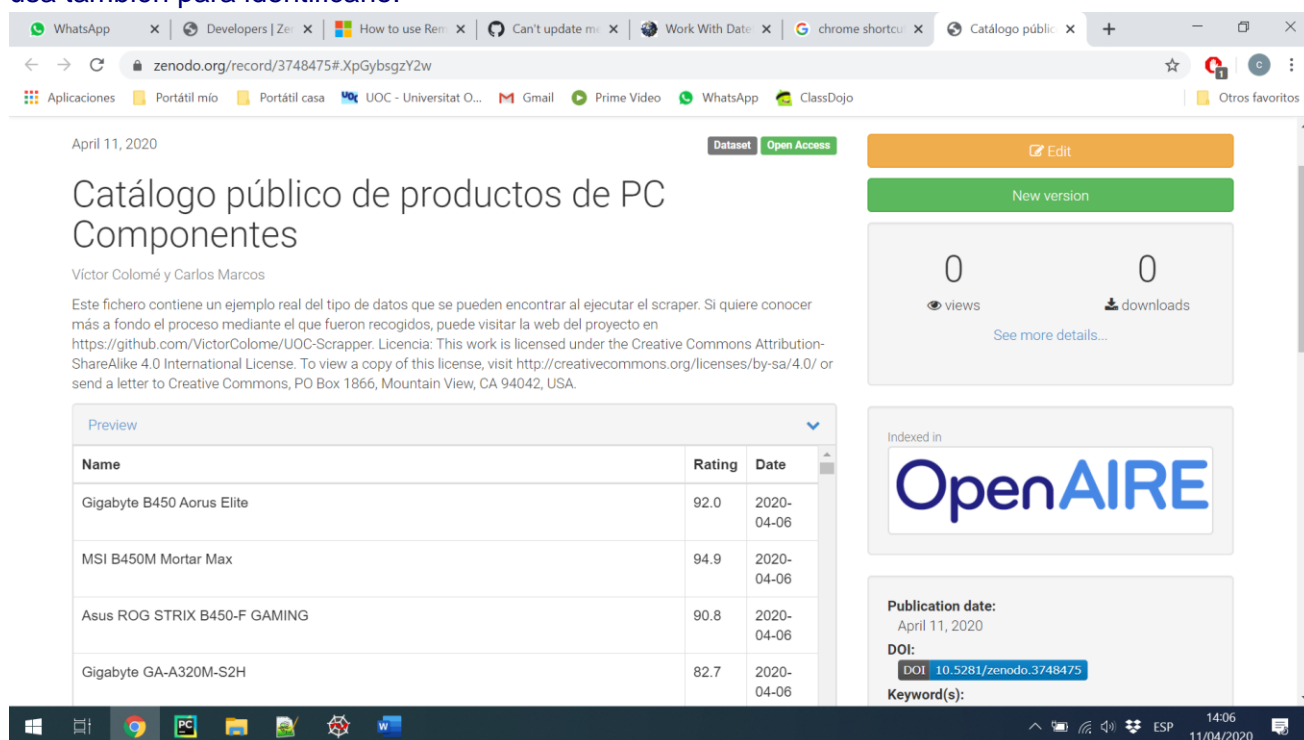
El título elegido es “Catálogo de productos de PCCOMPONENTES”. Incluiremos dicho título tanto en GitHub como en Zenodo.

¹ El sitemap de las categorías se puede encontrar en https://www.pccomponentes.com/sitemap_categories.xml

² El archivo robots.txt se puede encontrar en <https://www.pccomponentes.com/robots.txt>

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).

El juego de datos contiene una representación temporal de las características principales de los artículos de una de las múltiples categorías del sitio web que ha sido procesado. En concreto, nos centraremos en aquellas características que presenten mayor utilidad de cara a un análisis estadístico. Éstas son, principalmente, el precio del artículo y la puntuación o rating que le asignan los usuarios de la página web. El juego de datos consistirá entonces en una matriz tridimensional donde una de las dimensiones será el artículo perteneciente a la categoría, otra será el precio o puntuación y la tercera será la temporal (el día en que se han recolectado dichos datos). Los datos se presentan en Zenodo en forma de tabla bidimensional, donde una columna adicional indica la fecha de descarga de la web y la primera columna indica el nombre del artículo, el cual se usa también para identificarlo.



April 11, 2020

Catálogo público de productos de PC Componentes

Victor Colomé y Carlos Marcos

Este fichero contiene un ejemplo real del tipo de datos que se pueden encontrar al ejecutar el scraper. Si quiere conocer más a fondo el proceso mediante el que fueron recogidos, puede visitar la web del proyecto en <https://github.com/VictorColome/UOC-Scraper>. Licencia: This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Name	Rating	Date
Gigabyte B450 Aorus Elite	92.0	2020-04-06
MSI B450M Mortar Max	94.9	2020-04-06
Asus ROG STRIX B450-F GAMING	90.8	2020-04-06
Gigabyte GA-A320M-S2H	82.7	2020-04-06

Indexed in OpenAIRE

Publication date: April 11, 2020

DOI: [10.5281/zenodo.3748475](https://doi.org/10.5281/zenodo.3748475)

Keyword(s):

Ilustración 1: Captura de la web de Zenodo donde aparece el juego de datos de la práctica

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

La parte de visualización se puede lanzar como proceso diferenciado del scraper y de la publicación. Se ha dividido dicha representación en dos partes:

Estudio agrupado de las características

En esta primera parte, se dibuja un diagrama de cajas representando en él todos los artículos de una categoría elegida junto con sus características principales:

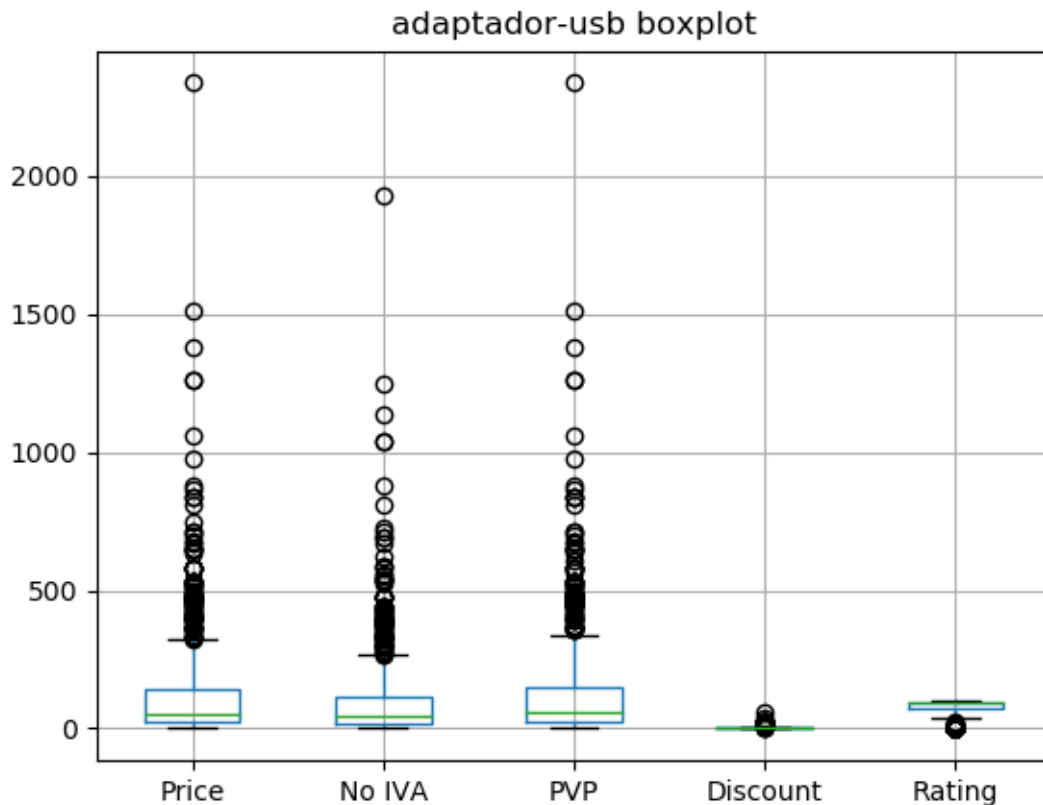


Ilustración 2: Diagrama de cajas con la distribución de cada característica para esa categoría

En este gráfico puede observarse de un vistazo la distribución de los valores que toma cada una de las características para los artículos de la categoría “adaptador-usb”. De esta forma, se puede observar que el rating y los descuentos son características con una distribución de los valores bastante agrupada, mientras que el Precio y el IVA muestran una distribución claramente sesgada hacia valores bajos con unos cuantos outliers en valores altos, es decir, que en esta categoría tenemos muchos artículos con precios bajos y unos cuantos con valores por encima de 1000€, aunque no es lo normal. Si asumimos que por ejemplo, nuestra población objetivo para una web de estas características es gente con ingresos reducidos, podríamos tratar de reducir nuestra oferta de aquellos artículos que aparezcan aquí como outliers, y centrarnos más en electrónica de consumo con precios más reducidos.

Estudio temporal

Lo que se hace aquí es tomar como entrada los ficheros CSV que se le especifiquen (normalmente del directorio donde el scraper los ha descargado previamente: `sample\csv`) y generar a partir de ellos un gráfico de líneas representando los artículos en diferentes colores y su evolución temporal:

en el eje de abcisas se representa el tiempo y en el eje de ordenadas, la característica que se quiera estudiar, como el precio o el rating de cada artículo.

La imagen siguiente muestra un ejemplo del gráfico generado a partir de los datos. Dentro del código se han filtrado los artículos para mostrar solamente aquellos que tienen alguna variación en el valor de la variable objetivo, entendiendo que los que no lo hacen son menos interesantes y por tanto pueden ser eliminados de esta visualización.

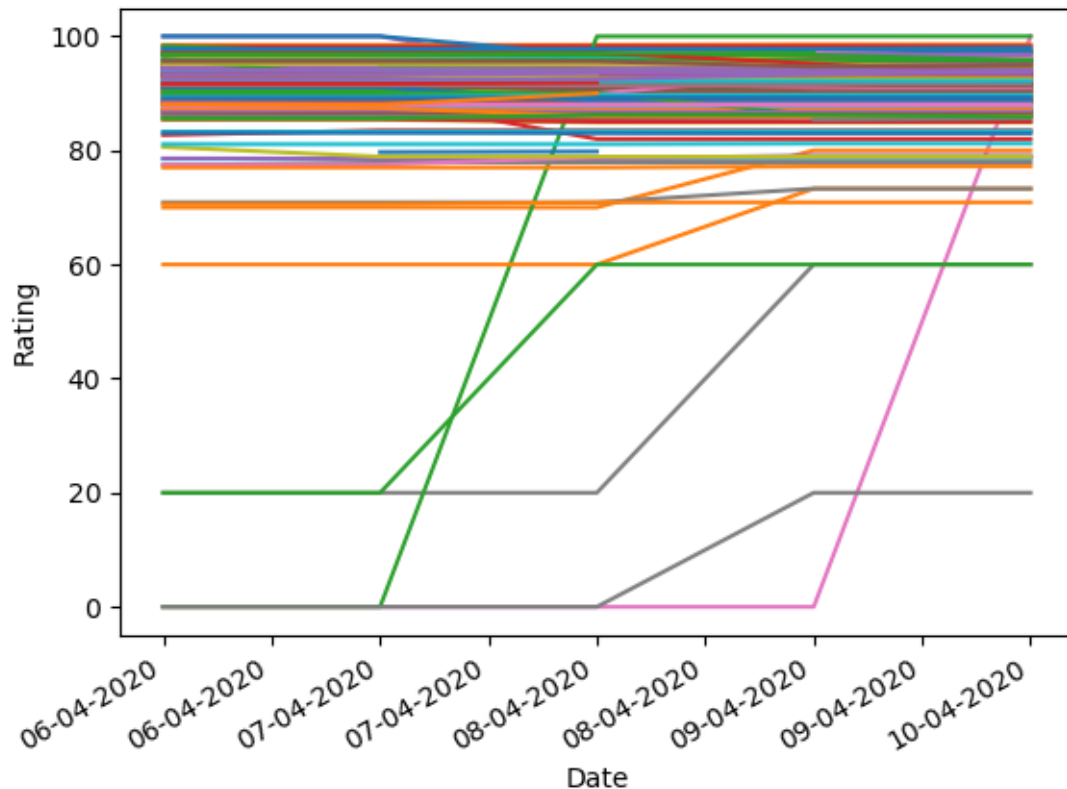


Ilustración 3: Comparativa del rating para todos los artículos de la categoría objetivo

En el gráfico anterior se puede apreciar cómo la puntuación de varios artículos se eleva en los últimos días, quizá debido al periodo vacacional de Semana Santa. En una serie temporal más larga podrían establecerse tendencias a partir de la representación e incluso tratar de hacer predicciones de cómo se van a comportar los indicadores de precio o rating en función del momento temporal que sea.

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

El dataset está formado por un conjunto de csv, donde cada uno representa la lista de productos / artículos de una categoría. El nombre de los archivos csv contiene el nombre de la categoría y la

fecha de ejecución de la forma: {categoria}_articles_attributes_{fecha}, por lo que aprovechando la marca temporal se pueden utilizar a posteriori para realizar estudios históricos; como por ejemplo, de la evolución de precios.

Campo	Tipo de dato	Descripción
Name	String	Nombre.
price	Float	Precio.
pvp	Float	Precio P.V.P.
discount	Float	Porcentaje de descuento.
no_iva	Float	Precio sin iva.
rating	Float	Valoración sobre 100.
features	Feature	Este campo es un tipo de datos complejo que contiene a su vez varios datos:
features.characteristics	Array of Strings	Lista con las características principales
features.specifications	Array of Strings	Lista con las especificaciones técnicas
features.manufacturer_url	String	URL a la web del fabricante del artículo

Para recogerlo, nuestro scraper se conecta a la web y va recorriendo todo el árbol de categorías (y dentro de éstas, de sus artículos), parseando y descargando la información descrita en el punto anterior.

El scraper guarda cada vez que se ejecuta un CSV con la información de las categorías y artículos existentes en ese momento en la web. El nombre del CSV incluye la marca temporal de cuando se recogió, por lo que se pueden utilizar a posteriori para realizar estudios de evolución de precios, por ejemplo.

Para publicar el dataset en Zenodo se ha decidido utilizar únicamente los campos principales (Name, price, pvp, discount, no_iva, rating). La razón de esto es doble: por un lado los campos de features y sus derivados están compuestos íntegramente por texto libre descriptivo, con lo que no tienen demasiado interés para realizar un análisis estadístico, a menos que se utilizaran herramientas de procesamiento del lenguaje natural, lo cual está fuera del alcance de esta práctica. Por otro lado, estos campos ocupan bastante espacio (hasta 20 veces más que los otros), y podría dar problemas de rendimiento o almacenaje al subirlos a Zenodo.

Se muestra a continuación una captura de la carpeta donde se han ido almacenando los resultados del scraper:

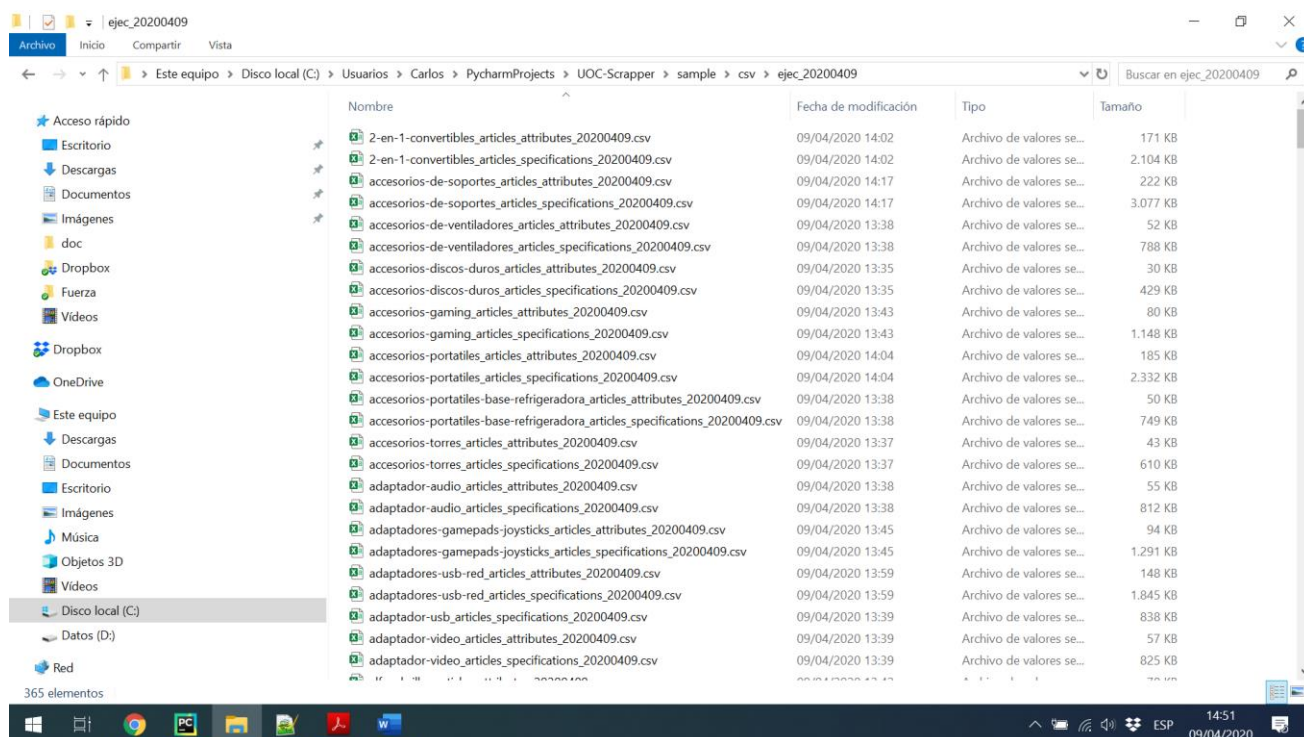


Ilustración 4: Captura de pantalla de uno de los directorios donde se almacenan los resultados del proceso de scrapping

También debido al alto volumen de datos obtenido, para la prueba de concepto se ha decidido restringir los resultados obtenidos tanto en los gráficos mostrados en este documento como en los ficheros subidos a Zenodo a una categoría representativa. La categoría elegida ha sido “adaptador-usb”.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El propietario del conjunto de datos es la empresa “PC COMPONENTES Y MULTIMEDIA SL”. Agradecemos a dicha empresa el poner a disposición de los usuarios dichos datos de manera pública. No se han encontrado análisis anteriores enfocados a esta web.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Nos llamó la atención este conjunto de datos porque presentaba el reto de poder parsear una página web comercial muy conocida, con bastante tráfico y una gran variedad de productos. Respecto a los snapshots de un momento concreto, nuestra idea es mostrar la distribución de las variables significativas (precio, descuento, puntuación del artículo) en ese instante, lo cual puede

ser útil para ver por ejemplo cuáles son los artículos más valorados o con mejores descuentos ese día.

A partir de los datos históricos, se podría hacer un estudio de la evolución de precios (y otras variables) y cómo dicha variable se comporta según determinado marco temporal: por ejemplo, si los precios aumentan el fin de semana o hay ciertas épocas (Navidades, Black Friday, etc) en las que los precios varían significativamente con respecto al resto del año.

Algunos ejemplos de preguntas que se podrían contestar con este análisis son:

- ¿Cuál es la distribución de las variables significativas (precio, rating, ...) para los artículos de la misma categoría? → ¿Qué artículos son outliers de estas distribuciones?
- ¿Qué artículos sufren variaciones bruscas en el rating o precio durante el periodo de tiempo?
- ¿Qué artículos de determinada categoría presentan una evolución más significativa en el periodo temporal de estudio?
- Si se contasen con series temporales más amplias: ¿existen ciertos momentos en el año en que las variaciones en precio sufren cambios que se repiten a lo largo de los años (ejemplo, Black Friday, Navidades, vacaciones de verano, ...)

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*

Considerando que nuestro proyecto es meramente académico, consideramos que debe ser público para que lo pueda usar cualquier persona con cualquier fin, excepto con fines comerciales. De esta manera, nos hemos decantado por la licencia [BY-NC-SA](#) ya que cumple con todas nuestras premisas y, además, se nos da crédito por el dataset y cualquier cambio en el mismo ha de ir con la misma licencia.

Se ha añadido el texto pertinente de la licencia al dataset de Zenodo

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

El código se ha realizado íntegramente en Python y la sincronización del trabajo entre los integrantes se ha realizado mediante Git y GitHub. El proyecto de Python se divide en 3 partes bien diferenciadas, correspondiente cada una con una carpeta del proyecto:

1. Publishing. Esta parte es la encargada de publicar el juego de datos en Zenodo

2. **Sample.** En esta parte es donde se realiza todo el scraping del sitio web
3. **Visualization.** Esta parte contiene las clases para generar los gráficos (ver Ilustración 2 e Ilustración 3)

Adicionalmente se tiene la carpeta doc, que contiene este mismo documento, la carpeta test, con las clases que permiten probar el núcleo del scraper, y el fichero README.md, que contiene una página de bienvenida al proyecto en GitHub en formato markdown y donde se describe el código con más detalle, así como los comandos necesarios para lanzarlo desde modo consola.

La URL al GitHub donde se ha desarrollado el scraper es <https://github.com/VictorColome/UOC-Scraper>

10. **Dataset.** Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

Tal y como recomienda Zenodo³, se proporciona el enlace a la versión específica de los datos. En caso de que el scraper se utilizara de manera intensiva, se debería añadir una parte de versionado al juego de datos, para poder realizar actualizaciones periódicas del mismo dataset en vez de subir un nuevo fichero.

En Zenodo se ha añadido una pequeña descripción que incluye el enlace al proyecto GitHub para quien desee ampliar información sobre cómo se han recogido los datos.

Puede verse el Dataset en <https://doi.org/10.5281/zenodo.3748520>

11. **Entrega.** Presentar el trabajo con el DOI del dataset en Github

En el fichero README.md de GitHub se ha añadido el enlace al dataset de Zenodo

³ <https://help.zenodo.org/#versioning>

Checklist de contribuciones del equipo

CONTRIBUCIONES	FIRMA
Investigación previa	VC, CM
Redacción de las respuestas	VC, CM
Desarrollo código	VC, CM