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](#_Toc37334668)

[1. **Contexto**. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. 3](#_Toc37334669)

[2. Definir un **título** para el dataset. Elegir un título que sea descriptivo. 3](#_Toc37334670)

[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](#_Toc37334671)

[4. ***Representación gráfica***. Presentar una imagen o esquema que identifique el dataset visualmente 4](#_Toc37334672)

[5. ***Contenido***. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. 4](#_Toc37334673)

[6. **Agradecimientos**. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay). 5](#_Toc37334674)

[7. **Inspiración**. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. 5](#_Toc37334675)

[8. **Licencia**. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: 5](#_Toc37334676)

[9. **Código**. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. 6](#_Toc37334677)

[10. **Dataset**. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción. 6](#_Toc37334678)

[11. **Entrega**. Presentar el trabajo con el DOI del dataset en Github 6](#_Toc37334679)

[Checklist de contribuciones del equipo 7](#_Toc37334680)

# Puntos a desarrollar en la práctica

## **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](http://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.xml[[1]](#footnote-1), 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.txt[[2]](#footnote-2) 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.

## 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.

## ***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 conjunto de datos elegido se presentará en ficheros CSV. Esta parte aún no está definida del todo, pero esperamos poder presentar un CSV que contenga todos los artículos clasificados por categorías y que contenga en cada línea al menos el nombre del artículo y sus atributos principales. Estamos aún definiendo cómo almacenar y presentar también las características y especificaciones.

## ***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:

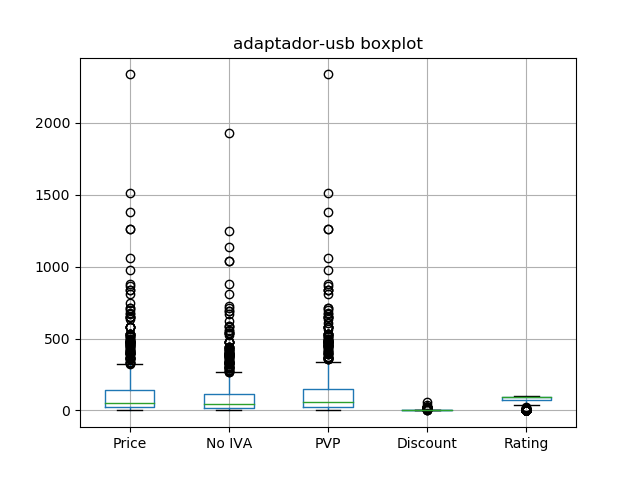


Ilustración : Diagrama de cajas comparativo de artículos y características

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 muestras 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.

### 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.

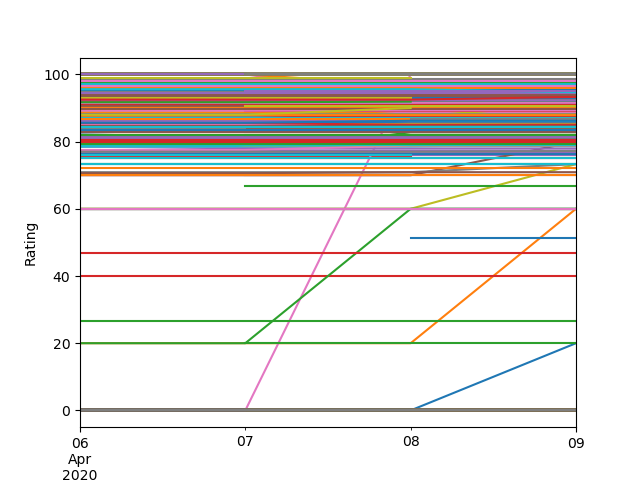


Ilustración : Comparativa del rating para todos los artículos

INSERTAR AQUÍ LA ÚLTIMA IMAGEN DE EJEMPLO

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.

## ***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:

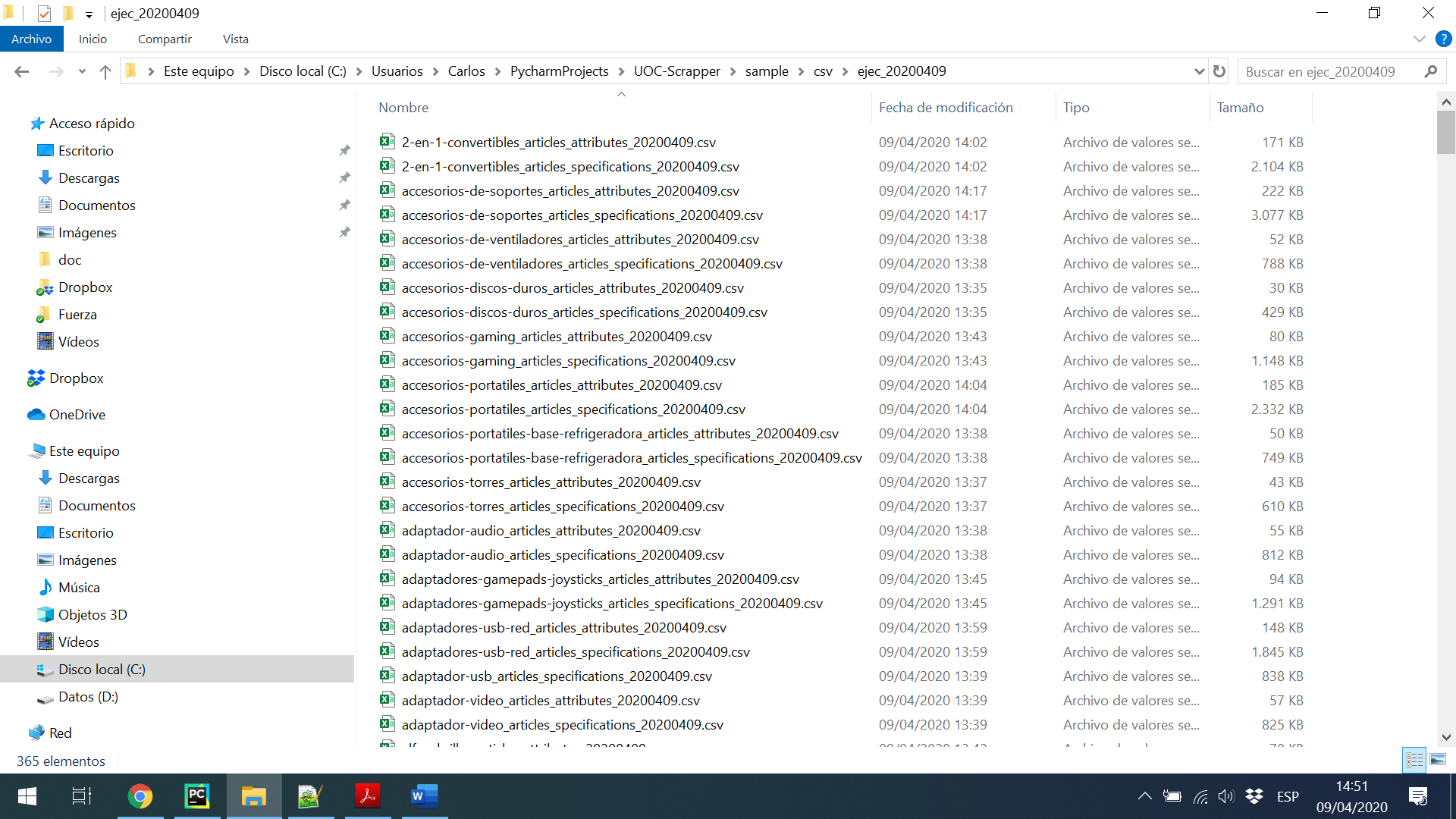


Ilustración : Captura de pantalla de uno de los directorios donde se almacenan los resultados del proceso de scrapping

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”.

## **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”

## **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 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

## **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](https://creativecommons.org/licenses/by-sa/4.0/) 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.

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

En estos momentos el código se encuentra terminado aproximadamente al 40%. Faltaría entre otras cosas establecer la frecuencia de recogida de datos y toda la parte de preparación y análisis de datos, presentación y publicación de los mismos en Zenodo y la documentación, es decir, la respuesta completa a todas las preguntas de este mismo documento.

El GitHub donde se está desarrollando el scraper es <https://github.com/VictorColome/UOC-Scrapper>

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

Falta poner aquí el enlace al último dataset que subamos

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

**Falta poner aquí el DOI del último dataset que subamos**

# Checklist de contribuciones del equipo

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

1. El sitemap de las categorías se puede encontrar en <https://www.pccomponentes.com/sitemap_categories.xml> [↑](#footnote-ref-1)
2. El archivo robots.txt se puede encontrar en <https://www.pccomponentes.com/robots.txt> [↑](#footnote-ref-2)