dpreview.com scraping

Scraper de datos sobre cámaras digitales de dpreview.com

Práctica 1 de la asignatura *"Tipología y ciclo de vida de los datos"* del Máster en Ciencia de Datos de la UOC

Autores

- Pablo Benito
- Miquel Rived

Contexto

Esta práctica se ha realizado bajo el contexto de la asignatura Tipología y ciclo de vida de los datos, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya. En ella, se aplican técnicas de web scraping mediante el lenguaje de programación Python para extraer así datos de la web dpreview.com y generar un dataset.

Digital Photography Review (dpreview.com) es un sitio web sobre cámaras digitales y fotografía digital en el que se pueden encontrar análisis de cámaras digitales, guías de compra, opiniones de usuarios y foros muy activos. Es uno de los 1.500 sitios web más visitados en Internet, además de ser actualmente el sitio de fotografía difital con mayor audiencia.

Además de lo comentado, lo que ha hecho decantarse por esta dirección para realizar el web scrapping es su amplia base de datos con información sobre cámaras dígitales.

Definir un título para el dataset

Características y evaluación de cámaras fotográficas digitales

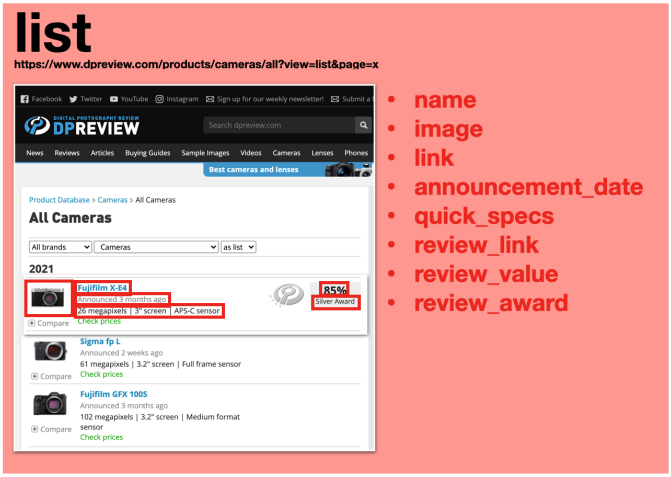
Descripción del dataset

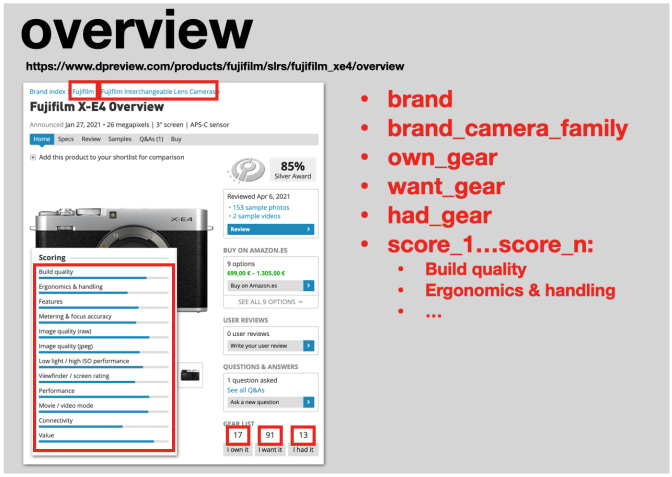
El `dataset` obtenido mediante el scraper contiene los datos de la base de datos de cámaras digitales recopiladas por dpreview.com.

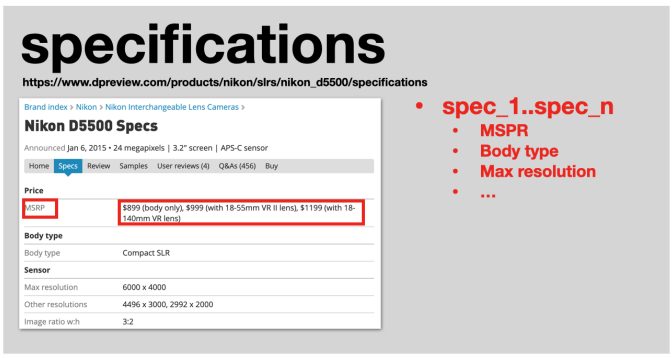
El dataset contiene tanto características técnicas de las propias cámaras, así como la valoración otorgada por los expertos de la propia página y de los usuarios de su comunidad.

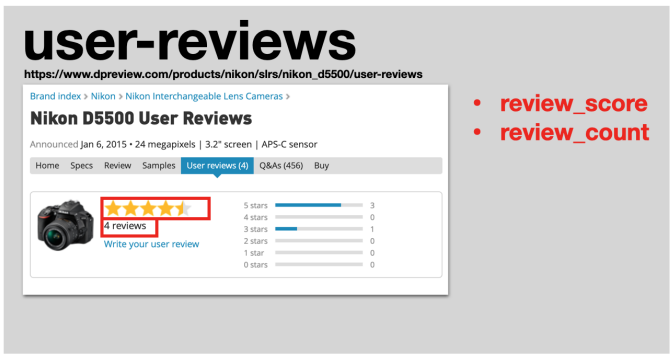
Representación gráfica

Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido









Contenido

Se podrían separar los campos extraídos en cuatro áreas, la primera (list) con breves características de la cámara, la segunda (overview) en la que se profundiza en la valoración dada por los expertos y se extraen otras características como la marca, la tercera (specifications) en la que se extraen la mayor cantidad de especificaciones de la cámara, y la última (user-review) de la que se extraen las valoraciones otorgadas por los usuarios.

Referente a lo que se extrae de la pantalla list, extraemos el nombre y una imagen de la cámara, la fecha del anuncio, especificaciones rápidas, el link de las reviews, así como el valor de las valoraciones de los especialistas.

En la pantalla overview se extrae la marca de la cámara y su familia, las personas que la tienen, la han tenido o la querrían tener, además una valoración del 0 al 100 referente a los siguientes aspectos: calidad de construcción, ergonomía y manejo, características, precisión de medición y enfoque, calidad de imagen (raw), calidad de imagen (jpeg), rendimiento con poca luz, valoración del visor, modo de vídeo, conectividad y el valor, además de la media de la valoración de los usuarios.

En cuanto a las características de las cámaras se ha extraído una gran cantidad de campos, destacando el precio, los píxeles del sensor, la máxima resolución, el tamaño de la pantalla, el tipo de cuerpo o el MSPR entre muchos otros.

Por lo que se refiere a la pantalla de user-reviews, se ha extraído información de la valoración media de los usuarios y cuantas valoraciones se han hecho en cada cámara.

El `dataset` consta en total con 125 atributos de todo tipos: numéricos, factores, texto. Dado que la información recopilada se corresponde con cámaras

Agradecimientos

dpreview.com es el sitio web de referencia en lo que a cámaras fotográficas digitales se refiere.

Lleva activo desde 1999 y cuenta con una gran comunidad de usuarios muy activos, sus reviews destacan por su calidad, incorporando muestras fotográficas de un gran número de cámaras digitales. Además, dpreview pertenece al grupo IMDB, famoso por su base de datos de valoraciones de películas y actores.

En el análisis hecho por fongfan999 se estudiaron las reseñas en Amazon de las cámaras publicadas en dpreview.com.

Por otro lado, en el análisis realizado por nmounika se analizaron las especificaciones de distintas cámaras. En su aproximación, busca detalles concretos (p.ej. Body type), dentro de las especificaciones. Nuestro desarrollo en ese sentido es más genérico, adaptándose tanto a las especificaciones con más antiguedad, como a las más modernas (p.ej. GPS).

Inspiración

Lo más interesante del conjunto de datos extraído es la gran cantidad de especificaciones diferentes que se encuentran, así como el gran abanico de cámaras digitales que lo abarcan.

En primer lugar se quiere analizar que características técnicas afectan más en el aumento de precio de una cámara digital.

Por otro lado, se tratará de determinar qué cámaras son las más valoradas por los usuarios o los expertos, por lo que se pretenderá analizar las marcas más valoradas, si el precio influye en la valoración final, o que tipo de especificaciones son las que buscan los usuarios en una cámara digital para realizar una valoración alta.

Además de lo comentado con anterioridad, se pretenden responder preguntas como las siguientes:

- ¿Cuál es la cámara mejor valorada por los usuarios?
- ¿Cuál es la cámara más cara y más ergónomica?
- ¿Qué cámara es capaz de disparar más fotografías en modo ráfaga?
- ¿Cuál es la cámara con GPS más ligera y mayor autonomía de batería?

Estas preguntas pueden variar a lo largo de las prácticas, ya que disponemos de una gran variedad de cámaras y campos para analizarlas que seguro que al visualizar con mayor detenimiento nos hacen hacernos nuevas preguntas.

La diferencia principal con los otros estudios encontrados de dpreview.com es la gran variedad de campos que hemos seleccionado para analizar las cámaras.

Licencia

La licencia escogida para la publicación del dataset es **Released Under CC BY-SA 4.0** ya que por los siguientes motivos relacionados con sus cláusulas se considera la más idónea:

- En primer lugar, el hecho de tener que proveer el nombre del creador del conjunto de datos junto con los cambios realizados hace que se valore el trabajo de dpreview.com a la par que se exponen las aportaciones realizadas por nosotros en la extracción.
- Al permitirse su uso comercial hace que se puedan realizar trabajos a partir del dataset que nos reporten cierto reconocimiento.
- Toda contribución realizada a posteriori deberá distribuirse bajo la misma licencia, por lo que todo trabajo realizado sobre el que se está haciendo deberá seguir distribuyéndose bajo los términos planteados.

Código

El scraper se ha desarrollado en Python utilizando las librerías `requests` y `beautifulsoup`.

En el *script* en primer lugar se obtiene el listado de todas las cámaras disponibles, recorriendo todas las páginas disponibles. A continuación, a continuación se obtienen los datos de las páginas `overview`, `specifications` y `user-reviews`

El *script* sigue las mejores prácticas recomendadas:

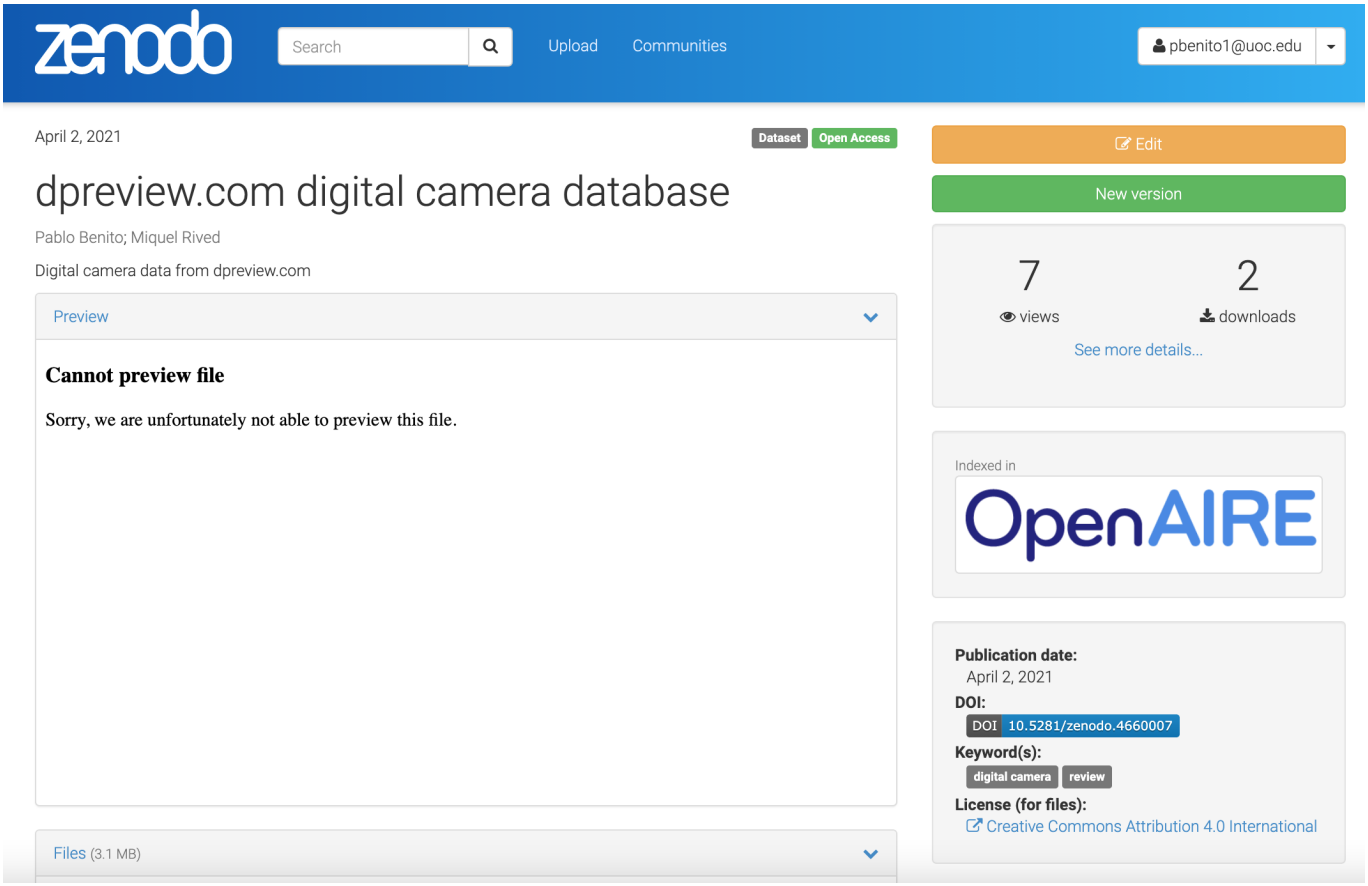
- Analiza el fichero `robots.txt`, evitando realizar peticiones a aquellas URLs que el administrador del sitio web ha indicado expresamente como `Disallow`
- Realiza las peticiones a dpreview.com utilizando el `user-agent` correspondiente a los navegadores más populares.
- Realiza una pausa de 500ms entre peticiones para evitar saturar con nuestras peticiones al servidor web.

El *script* se puede consultar aquí

Dataset

El dataset se encuentra publicado en Zenodo en el siguiente repositorio:

A continuación se muestra una captura del registro del dataset en Zenodo:



Contribuciones al proyecto

La siguiente tabla resume la contribución de los autores a los diferentes apartados del proyecto:

Contribuciones	Firma
Investigación previa	PB, MR
Redacción de las respuestas	PB, MR
Desarrollo código	PB, MR