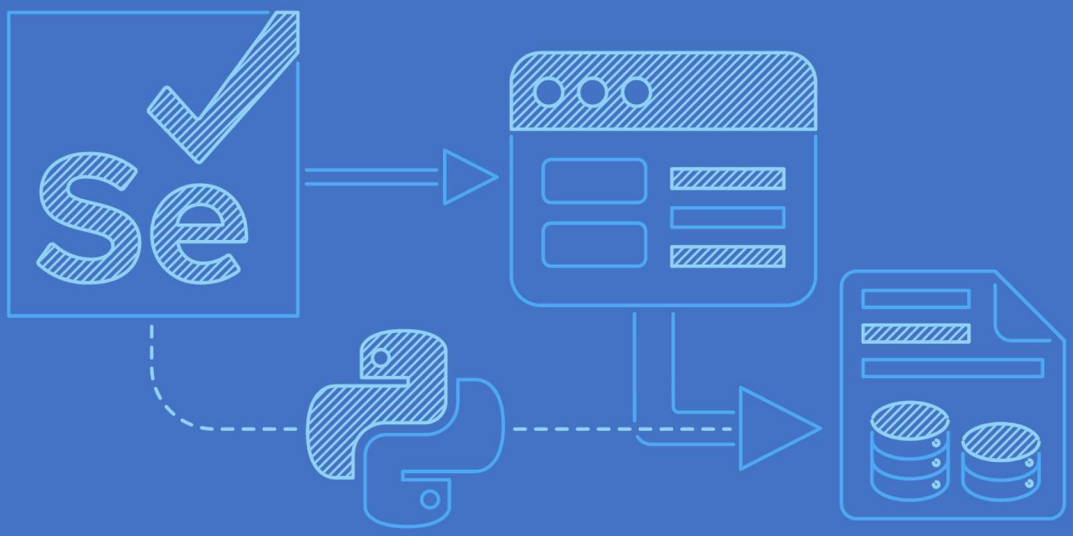
PRA - 1 ¿CÓMO PODEMOS CAPTURAR LOS DATOS DE LA WEB?



DIEGO SÁNCHEZ DE LA FUENTE - EDUARDO MORA GONZÁLEZ

Universitat Oberta de Catalunya

PRACTICA 1

Asignatura:

Tipología y ciclo de vida de los datos

Bloque:

PARA-1: ¿Cómo podemos capturar los datos de la web?

Apellidos: Sánchez de la Fuente

Nombre: Diego

Apellidos: Mora González

Nombre: Eduardo

Fecha: noviembre de 2022

Universitat Oberta de Catalunya

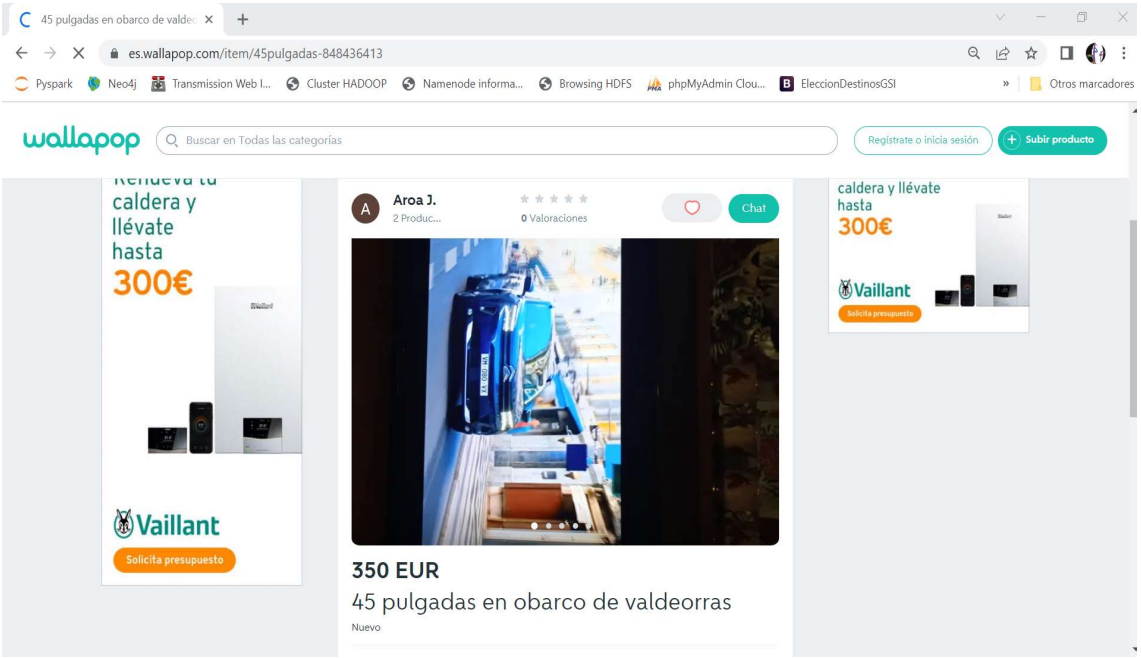
Contenido

1. INTRODUCCIÓN

En el presente documento, se detallan los pasos a seguir para la obtención de una *datasets* procedente de la tienda de segunda mano online de *Wallapop*, en él se describe el código necesario para rastrear las distintas páginas del portal, seleccionando una tipología de ellas (las que contienen la palabra "***item***" en su URL).

1.1. EJEMPLO ANUNCIO VISITADO

Un posible ejemplo de lo que se pretende se puede ver a continuación, en donde visitando una publicación como la siguiente se pretende capturar toda la información de esta:



https://es.wallapop.com/item/45pulgadas-848436413

(visitado el día 12/11/2022 puede que en el momento de lectura de este documento el anuncio haya sido eliminado, captura del anuncio)

1.2. JUSTIFICACIÓN Y OBJETIVOS

Teniendo en cuenta una serie de aspectos incluidos en el temario de la asignatura, nuestro código concluye con la creación de un *Dataset Pandas*, y su exportación a formato *CSV*, en donde se incluyen características (Precio, Producto, Fecha de Publicación, Usuario, etc) de los anuncios publicados en Wallapop.

Los objetivos que se esperan conseguir con la práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes cuyo tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios).
- Actuar según los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2. DESCRIPCIÓN DE LAS TAREAS Y RESPUESTAS

En los siguientes subapartados se darán respuestas a todas las preguntas y cuestiones que se platean en el enunciado:

2.1. CONTEXTO

"Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web"

El contexto en el que se recolecta la información para este trabajo es el de la compraventa de segunda mano. Se elige el sitio de Wallapop, puesto que es una de los más populares, plataforma de compraventa de productos de segunda mano.

La idea es analizar el contenido de todo el portal, para establecer estadísticas de productos más vendidos, productos de los que se obtiene mayor rendimiento económico, posibilidades de nichos de negocio, etc.

En la práctica se recolecta una muestra de unos 1800 artículos puestos a la venta en Wallapop el día 12/11/2022, que se podría utilizar como inicio del análisis anteriormente comentado.

2.2. TÍTULO DEL DATASET

"Título. Definir un título que sea descriptivo para el Dataset."

El título que le damos al Dataset es: *"Base de Datos con Información básica de 1842 artículos de Wallapop"*

2.3. DESCRIPCIÓN DEL DATASET

"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."

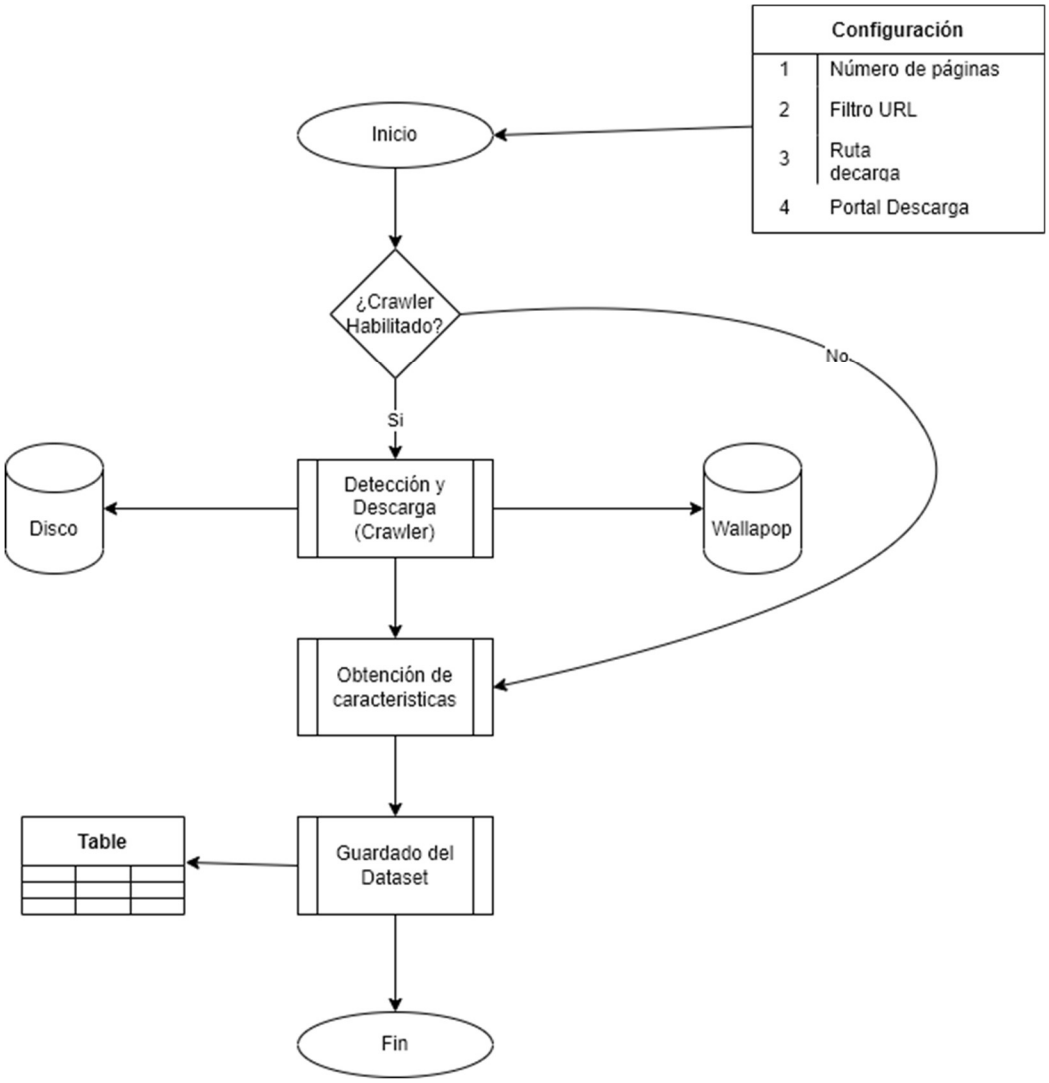
El dataset extraído de la plataforma Wallapop, presenta una recopilación en formato CSV con 9 características principales (aunque estas características se podrán ir ampliando en futuras revisiones). Las características son:

- **Producto**
- **Usuario**.
- **Detalle del Producto**
- **Precio**
- **Ubicación**
- **Fecha**
- **Visitas**
- **Hash**
- **URL**

2.4. REPRESENTACIÓN GRÁFICA

"Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido."

A continuación, se presenta un diagrama de flujo con las tareas para la generación del dataset:



2.5. CONTENIDO

El dataset se ha generado a partir del *scrapping* de la URLs filtradas previamente por el término "***item***". Las características principales (mencionadas anteriormente) son:

- **Producto** → Título del producto puesto a la venta por el usuario de la plataforma.
- **Usuario** → Nombre del usuario del producto en la plataforma.
- **Detalle del Producto** → Descripción del producto que se encuentra dentro del anuncio.
- **Ubicación** → Localidad desde donde se pone en venta el producto, si no se ha podido *scrappear* este contenido, aparecerá "Ninguno".

- **Fecha** → Fecha de publicación en la plataforma del producto, si no se ha podido scrappear este contenido aparecerá “Ninguno”.
- **Visitas** → Visitas registradas del producto, si no se puede scrappear dicho contenido aparecerá 0.
- **Hash** → Etiquetas de productos que Wallapop detecta como relacionados al que se está viendo.
- **URL** → URL a la página de Wallapop con el producto.

La vigencia del dataset es el de la propia fecha de captura, ya que los contenidos de Wallapop son altamente volátiles, llegando a cambiar incluso en el mismo día por lo que se recomienda para la implementación de este proyecto en producción, como mínimo una ejecución del *crawler* diariamente, pudiendo habilitar la caché de resultados, o la descarga incremental, es decir, detectar y descargar un contenido sólo si es nuevo o ha cambiado, con respecto a la última ejecución, para minimizar y optimizar recursos.

2.6. PROPIETARIO

“Propietario. 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. Justificar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto.”

Los datos publicados por Wallapop son de carácter público, ya que el principal cometido de la empresa es publicitar el contenido que venden sus usuarios, como si de una tienda física se tratase, los escaparates pueden ser vistos, fotografiados, publicados en plataformas de terceros siempre que se cite la fuente original, y no se modifique su contenido.

2.7. INSPIRACIÓN

“Explicar 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“

En los últimos tiempos la compraventa de productos de segunda mano se ha puesto de moda, ya que es una buena práctica desde el punto de vista de la ecología, se reutilizan los artículos dándoles una segunda vida, en vez de llenar de productos que aun sirven los vertederos, por todo esto se propone un analítica de los productos más vendidos, y los más rentables dentro de la plataforma Wallapop.

2.8. LICENCIA

"Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse"

Por lo que la licencia de uso de los datos, que más se adapta para este fin sería:

Creative Commons (CC): *"Attribution Non-Commercial Share Alike / Atribución-NoComercial-CompartirIgual (CC BY-NC-SA)"*

Fuente: Wikipedia

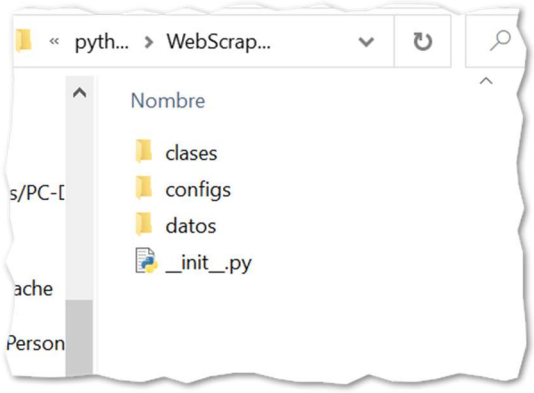
Esta licencia permite el uso de los datos siempre que no se modifiquen y su compartición, siempre que no se obtengan beneficios a costa o a través de dichos datos.

2.9. CÓDIGO

"Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R."

El código fuente se encuentra en:
https://github.com/diegosanfuen/ciclodevidadeldato/source

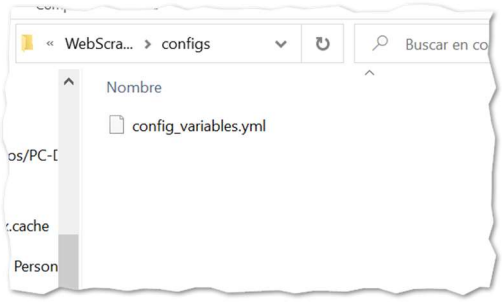
El proyecto se divide inicialmente entre 3 carpetas y un script de inicialización, junto con el fichero de requirements.txt:



El script __init__.py, será el que deberemos de ejecutar (previa configuración e instalación de los requerimientos del entorno *conda*) de la siguiente forma:

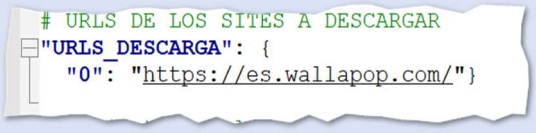
python __init.py

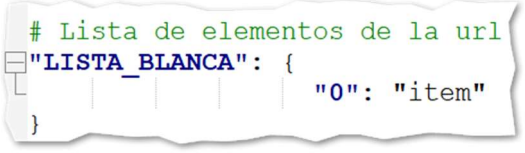
En la carpeta de */config* tenemos el siguiente fichero. *yml* en donde se podrá editar las distintas variables para cambiar la configuración:




```
# Variables de configuracion del proyecto

# Rutas necesarias para el Crawlwler
"RUTA_SITE_DESCARGA": {
                    "0":"wallapop"}
# RUTA RELATIVA A LA RUTA DE ESCRITURA DE LOS DATOS PARA DESCARGAR EL PROYECTO
# Ruta Base de las decargas del Crawler
"RUTA_BASE_DESCARGAS_CRAWLER": "datos/crawler_wallapop"
# "RUTA_BASE_DESCARGAS_CRAWLER": "datos/crawler_wallapop"

# URLS necesarias para el Crawler
# URLS DE LOS SITES A DESCARGAR
"URLS_DESCARGA": {
  "0": "https://es.wallapop.com/"}

# Parámetros del navegador (Crawler)
# USER AGENT PARA LOS DETECTORES DE ROBOTS
"USER_AGENT": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/106.0.0.0
Safari/537.36"

# TENER EN CUENTA ROBOTS RESTRICCIONES
"ENABLE_ROBOTS": "1" # 0 NO LOS TIENE EN CUENTA, 1 PARA TENERLOS EN CUENTA

# Lista de elementos de la url a descargar
"LISTA_BLANCA": {
                  "0": "item"
}

# 0 Para deshabilitar el crawler y 1 para habilitarlo
"ENABLE_CRAWLER": "1"
```

Los parámetros que se puede configurar en el fichero anteriormente mencionado son los siguientes,

PARÁMETRO	SIGNIFICADO
RUTA_SITE_DESCARGA	Es el nombre del directorio dentro del parámetro: RUTA_BASE_DESCARGAS_CRAWLER, (siguiente parámetro), donde se almacenarán de manera temporal todas las páginas descargadas.
RUTA_BASE_DESCARGAS_CRAWLER	Ruta a partir de la que cuelga la carpeta configurada anteriormente en el parámetro: RUTA_SITE_DESCARGA. ***Ejemplo:*** *RUTA_BASE_DESCARGAS_CRAWLER= "datos/crawler_wallapop"* *RUTA_SITE_DESCARGA= {"0":"wallapop"}* Las páginas descargadas de manera temporal serán guardadas en: *". \datos\crawler_wallapop\wallapop"*
URLS_DESCARGA	URL base del porta que queremos scrappear, en este caso Wallapop.  # URLS DE LOS SITES A DESCARGAR "URLS_DESCARGA": { "0": "https://es.wallapop.com/"}

LISTA_BLANCA	Listado con los filtros por URL de las páginas que queremos scrappear en este caso sólo se *scrappearan* las que contengan el término "***ítem***".
	
	 
	# Lista de elementos de la url
	"LISTA_BLANCA": {
	"0": "item"
	}

ENABLE_CRAWLER	Habilita o Deshabilita la ejecución del Crawler (que es el software que navega por las páginas y descarga las que nos interesan procesar, también llamado araña o spider).
NUMERO_PAG_DESCARGAS	Es un contador de páginas descargadas, como la web de Wallapop es muy extensa, y el algoritmo del crawler es muy exhaustivo, si no ponemos condición de parada, la descarga podría durar horas, este parámetro hace que cuando se llegue al número de páginas fijado, el algoritmo del crawler pare y continue con la fase de extracción de características.
MEJORA	Como mejora se propone la parametrización de la parte de extracción de características.

2.10. DIFICULTADES Y VISION FUTURA DEL PROYECTO

Entre las dificultades encontradas en la realización de esta práctica y del código destacamos dos:

- La enorme variabilidad del contenido, prácticamente el total de los datos descargados varían de un día para otro.
- Los *hacks* o trucos que existen dentro del código fuente de Wallapop, ya que una parte del código fuente de cada página se genera de manera dinámica. También existen partes del código generadas con *React*, o con otras técnicas como las siguientes que dificultan el procesamiento (*scrapp*) de las páginas.

Si queremos dar una vida futura al proyecto podemos plantearnos los siguientes puntos:

- Vemos que existe parte de la página que con la librería request, no se muestra, como por ejemplo el código en JavaScript, para evita esto se puede utilizar el navegador embebido de Selenium (WebDriver), que renderiza la página y permite ejecutar acciones de usuario de manera programada.
- En una segunda revisión del proyecto podría implementarse el WebDriver de Selenium.

- Por otro lado, estos portales, disponen de gestores de contenidos, que pueden cambiar totalmente la estructura interna de la página, lo que obliga a replantear el *Webscrapper*.

2.11. PUBLICAR EL DATASET EN ZENODO

"Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción."

El dataset se publica en Zenodo de manera pública en el siguiente enlace:

https://zenodo.org/record/7316004#.Y3EwDHZKg2w

```
El código DOI: 10.5281/zenodo.7316004
```

El dataset se publica en:

https://github.com/diegosanfuen/ciclodevidadeldato/tree/main/dataset

3. CONTRIBUCIONES

CONTRIBUCIONES	FIRMA
Investigación previa	DSF - EMG
Redacción de las respuestas	DSF - EMG
Desarrollo del código	DSF - EMG
Participación en el vídeo	DSF - EMG

4. CONCLUSIONES

En la práctica hemos desarrollado un algoritmo capaz de extraer información del portal Wallapop y almacenarla de manera estructurada en un fichero .csv, para su posterior análisis.

También hemos visto las principales dificultades que ataña el web Scraping, como la de mantener actualizados los datos, la variabilidad y las técnicas anti Scraping de las que disponen los portales web.