Web Scraping con Python 3.6

Práctica 1, Web Scraping

Miembros del equipo

- Morales Moreira Julio
- Mejía Borja Alexiss

Noviembre de 2022

Versión 1.0

Contenido

Práctica 1, Web Scraping

Descripción

El siguiente código en Python 3.6 tiene por objetivo aplicar las técnicas de Web Scraping para extraer datos de la web de la Organización Internacional de Fabricantes de Vehículos Motorizados OICA por sus siglas en inglés y generar como resultado un archivo de datos en formato CSV de las estadísticas de producción histórica de vehículos (desde el año 2000 hasta el 2021)

Esta práctica pertenece a la asignatura Tipología y ciclo de vida de los datos, correspondiente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya.

1. Contexto

Uno de los hechos más relevantes cuando se requiere analizar de datasets históricos es el proceso de descarga de cada uno de los años que se desean analizar, esta actividad puede tomar un lapso de tiempo considerable y más tratándose de varios periodos de tiempo, tradicionalmente es un proceso que consiste en visitar manualmente cada uno de los repositorios para recoger y descargar los datos que se necesiten. En este contexto donde los datos como la producción de vehículos de distintos países se conforman en datos importantes para analizar tendencias de crecimiento, o contrastes entre los distintos países, el uso de las herramientas que permitan acelerar la recopilación y posterior tratamiento de los mismos son indispensables.

La presente práctica tiene como objeto desarrollar un Script en el lenguaje de programación Python que permita la extracción anual de los datos correspondientes a la producción de vehículos, a partir de los datos publicados en la página web de la OICA https://www.oica.net/production-statistics/.

Se optado por el uso de lenguaje Python ya que tiene todas las herramientas para el manejo y procesamiento de datos, así como para llevar a cabo el proceso de escarbar dentro de páginas web, ya que cuenta con librerías y funciones que facilitan no solo este tipo de acciones si no también el posterior análisis de los datos obtenidos.

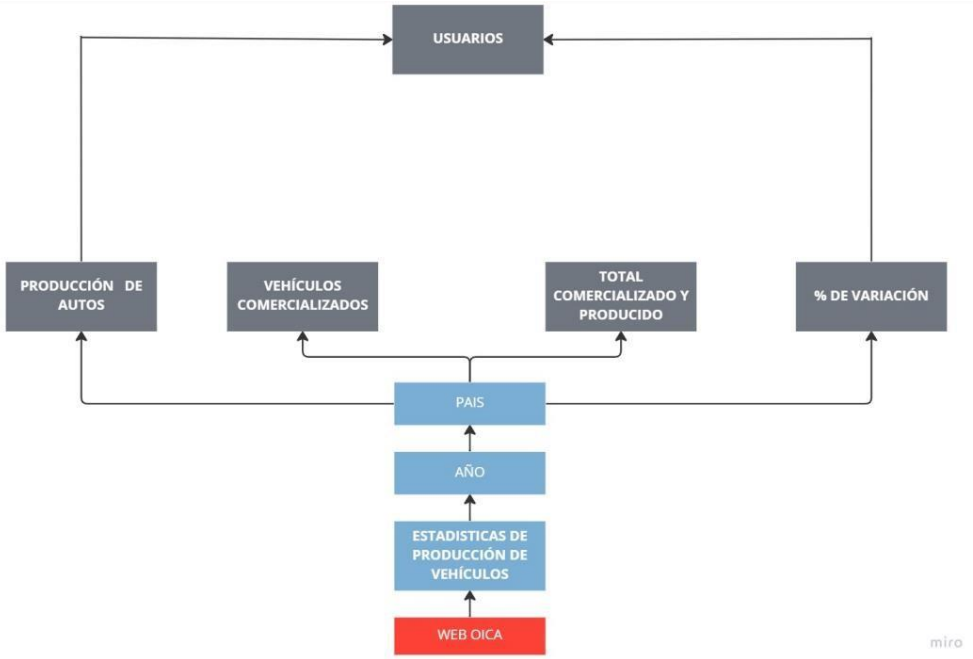
2. Título

"Estadísticas históricas de Producción de Vehículos en 40 países que reportan información en la OICA".

3. Descripción de dataset

El dataset se encuentra conformado por un conjunto de archivos que comprende un periodo de tiempo de 22 años y está constituido por las estadísticas relacionadas con, la producción de vehículos de un total de 40 países que han reportado estos datos a la Organización Internacional de Fabricantes de Vehículos Motorizados. Los datasets obtenidos son estáticos y anualizados, es decir que los valores presentados tienen un corte a final de cada año y no se los vuelve a actualizar hacia atrás del último periodo de tiempo; por lo que se tiene un conjunto de datos hasta el 2021, en vista de que el año 2022 aún se encuentra en curso.

4. Representación gráfica



5. Contenido

El dataset se encuentra conformado por un conjunto de archivos que comprende un periodo de tiempo de 22 años y está constituido por las estadísticas relacionadas con, la producción de vehículos cantidad de autos producidos, los vehículos comercializados, producción total y una variación porcentual de la producción, todo esto para alrededor de 40 países que reportan esta información, es importante mencionar que cada dataset se encuentra en diferentes menús, es decir que para cada año se tiene un dataset individualizado, adicionalmente es importante señalar que no se encuentra el mismo número de países en cada uno de los años seleccionados. Para cada conjunto de datos se recogen los siguientes antecedentes:

Country/Region: País o Región de producción de autos

Cars: Son vehículos de motor con al menos cuatro ruedas, utilizados para el transporte de pasajeros, y que comprenda no más de ocho asientos además del asiento del conductor.

Commercial Vehicle: Incluyen vehículos comerciales ligeros (turismos y vehículos comerciales ligeros), camiones pesados (Son vehículos destinados al transporte de mercancías. La masa máxima autorizada es por

encima del límite, que oscila entre 3,5 y 7 toneladas, de los vehículos comerciales ligeros. Incluyen tractor vehículos diseñados para arrastrar semirremolques), autocares y autobuses (se utilizan para el transporte de pasajeros, comprendiendo más de ocho asientos además del asiento del conductor, y que tengan una masa máxima por encima del límite, que oscila entre de 3,5 a 7 toneladas, de vehículos comerciales ligeros)

Total: Total de autos producidos.

Change: Diferencia porcentual de producción del año en curso respecto al año anterior.

6. Propietario

El propietario del dataset es Organización Internacional de Fabricantes de Vehículos Motorizados, que es una organización fundada en 1919 y que tiene por misión defender los intereses de los fabricantes, ensambladores e importadores de vehículos agrupados en su federación. La organización lleva a cabo actividades en los campos de asuntos técnicos, comunicación y estadísticas de la industria automotriz.

Como se puede notar es una organización con un amplio periodo de vigencia y de desarrollo de sus actividades, cuyos datos permiten realizar el seguimiento y el análisis de una de las industrias que genera más fuentes de trabajo directos e indirectos y que dinamiza en gran medida la economía de un país, tal es el caso que esta industria aporta más de 430 000 millones de euros a los ingresos de los gobiernos de veintiséis países.[1]

7. Inspiración

La recopilación de producción de autos tanto de pasajeros como de uso comercial se puede utilizar para realizar predicción y participaciones de los diferentes países. Por otro lado, este input de información se puede utilizar

[1] OICA (https://www.oica.net/)

en modelos de Minería de Datos orientados a extraer características al cruzar con otra información, tal como los niveles de contaminación en la producción de autos, la participación de la industria automotriz en el PIB de los países o el impacto de la industria automotriz en variables macro y microeconómicas como la generación de empleo o la apreciación/depreciación de las monedas locales.

Todos estos elementos permiten una amplia gama de temáticas que se pueden abordar con el dataset seleccionado para el raspado, sobre todo al tratarse de datos de diferentes países donde incluso se puede orientar los futuros resultados para evaluar las tendencias en las ventas y producción de vehículos y de esta manera tener un análisis incluso de los aportes económicos que tiene esta industria en la economía de cada uno de los países registrados en el dataset.

8. Licencia

La licencia que se utiliza para la publicación del conjunto de datos generado se define como ***Released Under CC BY-NC-SA 4.0 License,*** en base a:

BY: La atribución se debe citar de forma explícita al autor, comentando los cambios generados de manera de reconocer el trabajo colaborativo.

NC: No Commercial, es decir, se define un uso en base a un trabajo exploratorio universitario sin fines de lucro para proyectos de uso no comerciales

SA: Es posible compartir y utilizar el conjunto generado, siempre y cuando se publique con la misma licencia original.

9. Código

El código fue desarrollado en Python 3 y se encuentra conformado por 3 secciones.

La primera que contiene las librerías que fueron usadas para llevar a cabo el raspado de la web (requests, BeautifulSoup y pandas).

La segunda sección que contiene un recurso de un diccionario que permite agrupar y completar las tablas para los 22 años que contiene el dataset completo.

El tercer punto contiene el cuerpo del código que permite realizar el web scraping, el cual contiene por un lado el uso de la librería Requests la cual permite enviar una solicitud para obtener el HTTP de la página web; por otro lado se ha incorporado el analizador para extraer datos de HTML mediante la librería BeautifulSoup, concretamente para esta práctica se ha utilizado el analizador "`lxml`".

Adicionalmente, dentro de la tercera sección se podrá encontrar el proceso para obtener la información de la etiqueta de las tablas por medio de la creación de un id de cada tabla con el fin de poder realizar las iteraciones de cada una de las 22 tablas; se ha incorporado un controlador para verificar la correcta descarga de los años.

Finalmente, la cuarta sección contiene el proceso para exportar el dataset a un formato CSV, tal y como se requiere en el enunciado de la práctica.

El principal problema que se ha encontrado en el raspado de la página web de la OICA, es la disparidad en los nombres de algunas variables (columnas) de ciertos años del dataset y los diferentes periodos de tiempos sobre los cuales se desea realizar la descarga, para ello se ha utilizado un recurso de un diccionario de años que se va completando y diferenciando según los id de las tablas, esto posibilita tener un dataset histórico (22 años) en un solo dataframe.

Link Video

https://drive.google.com/drive/folders/1hucr_EOSYPiRm6_xLVLnQrBdSqdHEVVo?usp=share_link

Link DOI

10.5281/zenodo.7338483

Recursos

1. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
2. Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.
3. International Organization of Motor Vehicle Manufacturers

Contribuciones	Firma
Investigación previa	AM ,JL
Redacción de las respuestas	AM ,JL
Desarrollo del código	AM ,JL
Participación en el vídeo	AM ,JL