M2.851 – Tipología y ciclo de vida de los datos

PRÁCTICA DE WEB SCRAPING

Elena Ruíz Martínez

Alberto Bayón Valtierra

Tabla de contenido

1. Contexto.

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

El contexto en el que se ha desarrollado esta actividad es la asignatura Tipología y ciclo de la vida de datos del master de ciencia de datos de la Universidad Oberta de Cataluña que se realiza en el segundo semestre del curso 2018-2019.

La práctica elaborada consiste en un trabajo de web-scraping donde hay que utilizar la información de una página web para generar un conjunto de datos. El sitio web elegido (*http://coches.idae.es/base-datos*) proporciona una base de datos estructurada con información de coches pública lo que hace que el conjunto sea de gran interés para realizar esta práctica.

Es necesario encuadrar las técnicas de web-scrapping dentro de lo que se conoce como la sociedad de la información basada en el uso intensivo de las tecnologías de comunicación donde se genera gran cantidad de datos. Internet es parte de esa sociedad y pone a disposición de millones de usuarios un volumen de datos inimaginable hace años. La ventaja de las técnicas web-scrapping es que ofrece la posibilidad de extraer y descargar información que posteriormente se podrá utilizar con otros fines particulares siempre y cuando el propietario permita esa posibilidad. Además el uso de esta tecnología va unido a técnicas como Big Data y Analytics que permiten dar valor a los datos recogidos en la web

Desde el punto de vista comercial, la extracción automática de datos está presente en muchas empresas ya que es posible acceder a información de la competencia casi en tiempo real, de modo que se existe la posibilidad de analizar el crecimiento frente a los competidores utilizando esos datos de manera específica.

El sitio web que se ha seleccionado en esta práctica ofrece información precisa y comparable sobre el consumo de combustible y emisiones de CO2 de cualquier turismo que se está comercializando. La posibilidad de analizar los datos una vez extraídos puede ayudar a potenciales compradores a decidirse por vehículos que consuman menos combustible y que por lo tanto emitan menos CO2. Esta extracción es interesante tanto para los consumidores, como para los propios fabricantes ya que estos pueden analizar la emisión de vehículos de la competencia.

La práctica se ha orientado para que un potencial consumidor pueda extraer información de todos los turismos pertenecientes a una misma marca, de modo los datos proporcionados le puedan ayudar a elegir el modelo en términos de consumos y emisiones.

2. Definir un título para el dataset.

ConsumoYEmisionesDeTurismos.

3. Descripción del dataset.

El conjunto de datos que se generado contiene detalles sobre el consumo de carburante y emisiones de CO2 de los turismos nuevos que se comercializan en España. Esta información se ha recogido de la página web *http://coches.idae.es* que forma parte programa del Instituto para la Diversificación y Ahorro de Energía (IDAE), organismo adscrito al Ministerio para la Transición Ecológica.

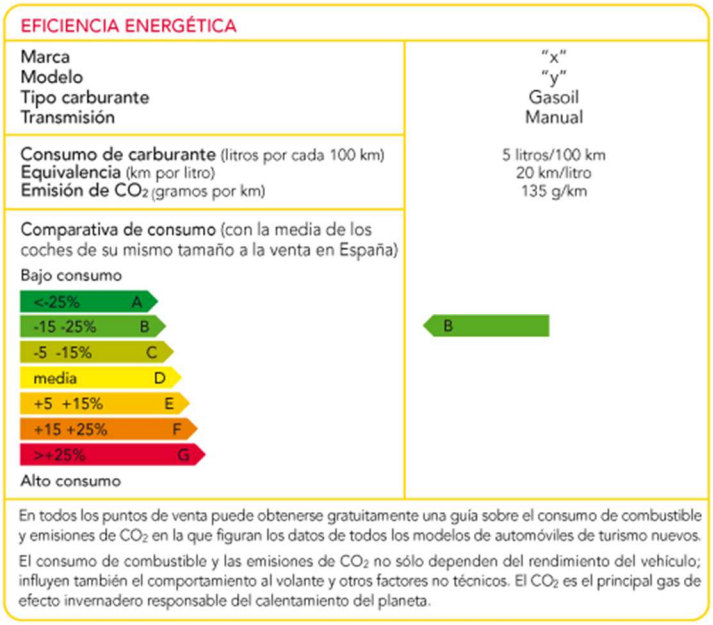
El dataset está constituido por el conjunto de modelos de una determinada marca de coches comerciables. Al ejecutarse el proceso se solicita una marca y se construye el dataset con todos los modelos disponibles para esa marca. El objetivo consiste en presentar los consumos y las emisiones CO2 de dichos modelos para ayudar al consumidor en la elección del modelo en términos de eficiencia energéticos.

El formato de salida es un fichero de texto del tipo csv (comma-separated values) donde cada línea contiene información energética de los modelos de la marca de coche seleccionado. Las características que se incluyen (modelo, clasificación energética, consumo mínimo, consumo máximo, emisiones mínimas y emisiones máximas) aparecen separados por comas. A excepción del modelo y la clasificación energética que son datos de tipo carácter los demás datos son numéricos.

La información que se presenta en la página web se actualiza cada seis meses desde el IDAE en colaboración con los fabricantes e importadores. Aunque la frecuencia se puede considerar adecuada, es cierto que la aparición de nuevos modelos durante el tiempo que no se actualicen los datos puede dejar al usuario sin la información energética necesaria para tomar decisiones adecuadas sobre su compra,

Los datos son consistentes y en principio no es necesario realizar una limpieza de los mismos, ya que esa tarea de recoger únicamente los daos necesarios se ha realizado en el código Python.

4. Representación gráfica



5. Contenido.

El dataset creado recoge una serie de información correspondiente a las mediciones de consumo de combustible y sobre las emisiones de CO2 según el ciclo WLTP del coche que se haya seleccionado. Cada fila corresponde a un modelo diferente de la marca seleccionada, y cada modelo contiene está compuesta por los siguientes campos:

- **Modelo:** Nombre del modelo, incluye la marca, los caballos y la potencia térmica.
- **Clasificación energética:** valor correspondiente a la clasificación por consumo relativo. Puede tomar los siguientes valores: *A, B, C, D, E, F* y *G*; donde *A* corresponde a un consumo mínimo y *G* a un alto consumo.
- **Consumo mínimo (l/100Km):** valor correspondiente al consumo mínimo del vehículo en litros por cada 100km.
- **Consumo máximo (l/100Km):** valor correspondiente al consumo máximo del vehículo en litros por cada 100km.
- **Emisiones mínimas (gCO2/km):** valor correspondiente a la emisión mínima de dióxido de carbono (CO_2) que emite el vehículo por km.
- **Emisiones máximas (gCO2/km):** valor correspondiente a la emisión máxima de dióxido de carbono (CO_2) que emite el vehículo por km.

Los datos ofrecidos por el Instituto para la Diversificación y ahorro de la Energía (IDAE) son elaborados anualmente, incluyendo todos los modelos de turismos nuevos existentes en la fecha

de publicación de la actualización, puestos en venta en los Estados miembros, clasificados por marcas y por orden alfabético.

Para la creación del dataset, primero hemos hecho una llamada a la web indicando la marca del coche del que queremos obtener la información. Los datos obtenidos de esta llamada son guardados en formato JSON; se recorre este JSON para obtener los datos que interesan y que se irán guardando en un dataframe. Una vez que se dispone de toda la información en el dataframe, se convierte en un fichero .csv. Para ello se ha hecho uso del lenguaje de programación Python y utilizado técnicas de Web Scraping.

6. Agradecimientos.

Este conjunto de datos ha sido creado por la Directiva y del Real Decreto, IDAE con el apoyo del programa SAVE de la DG TREN (Transporte y Energía) de la Comisión Europea. Lo datos correspondientes al consumo, emisión y otros datos técnicos han sido facilitados por ANFAC (Asociación Española de Fabricantes de Automóviles y Camiones), ANIACAM (Asociación Nacional de Importadores de Automóviles, Camiones, Autobuses y Motocicletas) e IEA (Instituto de Estudios de Automoción).

Estos conjuntos de datos están disponibles gratuitamente para los consumidores de nuevos coches, quienes podrán solicitarla en el punto de venta o ante un organismo designado en cada Estado miembro.

No se han encontrado análisis similares de scraping sobre la página web que se realiza en esta práctica, por lo que se considera que este trabajo puede servir de partida para otros posteriores donde se quiera ahondar en las características particulares de los vehículos (segmento comercial, motorización, cilindrada, tipo cambio…) con el fin de obtener un conjunto de datos más completo que pueda ayudar en el análisis del consumo desde varias perspectivas o dimensiones.

7. Inspiración.

Este conjunto de datos que informa del consumo de carburante y las emisiones de CO_2 de los turismos nuevos, es interesante tanto para las personas físicas como para las jurídicas de modo que tengan disponible esta información para poder considerar la adquisición de los vehículos más eficientes energéticamente.

El consumo de combustible y emisiones de CO2 específicos de los turismos puede influir en la decisión del consumidor en favor de los automóviles que consuman menos combustible y por lo tanto emitan menos CO2, impulsando de ese modo a los fabricantes a hacer lo necesario para reducir el consumo de los automóviles.

Entre otras aplicaciones, este conjunto de datos podría ser útil para empresas que tienen la necesidad de adquirir una flota de vehículos, a quienes les podrían interesar aquellos vehículos de menor consumo para poder ahorrar en costes.

Como fuentes de información previa para abordar este trabajo se ha utilizado los artículos que proporciona la propia página web de IDAE (*https://www.idae.es/)* y en particular el apartado dedicado al consumo de carburante y emisiones CO2 (*http://coches.idae.es/*).

Por otro lado hemos utilizado también otros enlaces que nos han permitido obtener más información sobre el objetivo de esta base de datos pública:

https://www.mapa.gob.es/ministerio/pags/biblioteca/revistas/pdf_AM/AM_2008_75_34_40.pdf

https://www.certificadosenergeticos.com/etiqueta-energetica-vehiculos-coches-combustible-tecnologias-alternativas.

El valor agregado que se proporciona en este trabajo es el de la posibilidad de obtener en un único fichero la información completa de una marca, para poder tratar esa información con otras herramientas visuales y analíticas con posterioridad.

8. Licencia

Autorizamos el uso de toda la información de nuestro dataset mediante ***CC BY-SA 4.0 License***. Esta licencia permite:

- Compartir**,** copiar y redistribuir el material en cualquier medio o formato.
- Adaptar, remezclar, transformar y crear a partir del material para cualquier finalidad, incluso comercial, siempre y cuando se difundan las contribuciones bajo la misma licencia que el original.
- Reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios.

De esta manera permitimos que nuestros datos sean útiles a terceros, ya sea para su uso informativo, de estudio o comercial siempre y cuando se distribuya bajo la misma licencia.

9. Código.

El código fuente utilizado para generar el dataset en formato Python, se encuentra en cualquiera de los siguientes dos repositorios de Github:

- https://github.com/eruizmartinez/ConsumoYEmisionesDeTurismos
- https://github.com/abayonv/ConsumoYEmisionesDeTurismos

10. Dataset.

El dataset en formato csv se encuentra en cualquiera de los repositorio de Github anteriormente mencionados, corresponde a los datos obtenidos de seleccionar la marca: **Jaguar**

Tabla de contribuciones al trabajo:

Contribuciones	Firma
Investigación previa	ABV, ERM
Redacción de las respuestas	ABV, ERM
Desarrollo código	ABV, ERM