M2.851 - Tipología y ciclo de vida de los datos: Práctica 1: Web scraping

Olga Garcés Ciemerozum y Carlos Acosta Quintas

Abril 2021

Contents

Contexto **1**

Título del dataset **2**

Descripción del dataset **3**

Representación gráfica **3**

Contenido **3**

Agradecimientos **5**

Inspiración **6**

Licencia **6**

Código **7**

Dataset **7**

Referencias **7**

INTRODUCCIÓN

El presente informe forma parte de la primera práctica de la asignatura M2.851 - Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de Datos impartido por la Universitat Oberta de Catalunya.

En esta práctica se realizarán técnicas de Web scraping aplicadas a una Web en concreto y se analizarán dichos datos para extraer información relevante y útil.

A su vez, se entregará, junto con la presente memoria, una serie de archivos con el código necesario para la realización de dicho Web scraping y varios juegos de datos reales y actualizados con el que el usuario podrá realizar diferentes estudios analíticos a posteriori.

Contexto

---

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

Contexto en la recolección de la información.

La página web seleccionada para realizar las técnicas de Web scraping es: https://www.bonarea.com/

BonÀrea es empresa con una amplia experiencia en el sector agroalimentario, donde su principal negocio se desarrolla en actividades ganaderas, industriales y comerciales con el fin de poder llegar al consumidor sin intermediarios.

Una de las divisiones de la empresa, bonÀrea Energía, dispone de más de 55 gasolineras que venden más de 430 millones de litros al año con un ratio calidad/precio que hace ahorrar más de 40 millones de euros al año a sus clientes.

Las gasolineras bonÀrea son conocidas por su precio económico debido a sus reducidos márgenes de beneficio y al gran volumen de carburantes vendido, y además debido al uso de economías de escala entres sus diferentes líneas de negocio.

Para ésta práctica, se ha decidido hacer un estudio geográfico y temporal de los precios de las gasolineras de bonÀrea Energía, para determinar, no solamente las variaciones diarias de los precios de los diferentes productos, sinó también realizar un registro de las economías de escala asociadas a cada gasolinera (servicios complementarios aportados por cada gasolinera).

Información extraída del sitio web.

El dominio www.bonarea.com aporta una web dinámica que referencia a todas las divisiones de negocio del grupo, implementando diferentes esquemas de datos para que la experiencia del usuario final a nivel de visualización de la información sea clara y concisa.

En el apartado de los establecimientos, y en particular en el de las gasolineras, la página web aporta diferentes datos sobre ellas, tanto a nivel geográfico mediante un mapa interactivo para que el usuario sepa en todo momento dónde está la gasolinara más cercana, como a nivel económico (se muestran los precios de los diferentes carburantes disponibles en cada una de ellas).

La página permite al usuario seleccionar uno o varios tipos de establecimientos y muestra la ubicación sobre el mapa de todos establecimientos de los tipos seleccionados.

Además, haciendo click en el icono **"+ info"**, cada gasolinera dispone de una página web anidada para que el usuario pueda tener información adicional como el horario de apertura, dirección , teléfono, etc. de cada una de los establecimientos.

En definitiva, el sitio web proporciona dicha información con solamente un motivo: conseguir el compromiso y la fidelización del cliente a una página web práctica y bien diseñada, que ayuda a obtener una información "a la carte" útil.

Hay que tener en cuenta **dos puntos claves** relacionados con la interacción del usuario:

-Durante la interacción del usuario con la web, los datos se generan de forma dinámica: la web envía una solicitud jquery de información y devuelve un conjunto de puntos en el mapa que cumplen con el criterio impuesto por el usuario.

- Al hacer click en los distintos puntos que aparecen en el mapa, la página realiza otra solicitud jquery para una instancia de establecimiento específica y visualiza un recuadro con su información básica.

Título del dataset

---

2. Definir un título para el dataset. Elegir un título que sea descriptivo..

El proyecto puede generar el dataset en dos modalidades:

1. Dataset en dos archivos csv: csv con datos de las gasolineras y csv actualizable con histórico de precios diarios de las diferentes gasolineras.

3. Dataset completo en un archivo csv (combinación de los dos primeros datasets)**

Por tanto, habrá 3 títulos propuestos para los diferentes datasets. Los títulos se muestran en orden de aparición en referencia a la anterior lista:

1. "Información descriptiva de las gasolineras bonÀrea."

2. "Valores diarios del precio de los carburantes en gasolineras bonÀrea."

3. "Gasolineras bonÀrea: Información descriptiva y valores diarios del precio de los carburantes."

Descripción del dataset

---

3. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido)..

///»> PENDIENTE DE ELABORACION

Representación gráfica

---

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

///»> PENDIENTE DE ELABORACION

Contenido

---

5. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El dataset contiene los precios diarios de los carburantes en distintas gasolineras de la red BonÀrea en el período de tiempo desde:

23 de marzo de 2021 hasta la fecha de entrega de la práctica.

Los diferentes datasets pueden ser generados en dos modalidades:

Dos ficheros csv:

'bonarea_gasolineras_prices.csv' - datos históricos del precio de los carburantes por gasolinera

'bonarea_gasolineras.csv' - datos de las gasolineras

Un fichero csv:
'bonarea_gasolineras_data_and_prices.csv' - datos históricos del precio de los carburantes por gasolinera con información detallada de cada gasolinera.

-> **'bonarea_gasolineras_prices.csv'**

id - número identificador de la gasolinera.

Fecha - fecha

Dia_Semana - día de la semana

latitude - latitud

longitude - longitud

minutsLatitude - latitud en minutos

minutsLongitude - longitud en minutos

GASOIL A - precio diario de este tipo de combustible

GASOLINA S/P 95 - precio diario de este tipo de combustible

GASOLINA S/P 98 - precio diario de este tipo de combustible

ADBLUE - precio diario de este tipo de combustible

-> **'bonarea_gasolineras.csv'**

id - número identificador de la gasolinera

type - tipo de establecimiento de BonÀrea

url - url de la gasolinera

street - calle donde se ubica la gasolinera

city - ciudad donde se ubica la gasolinera

postalCode - código postal de la dirección de la gasolinera

raoSocial - razón social del establecimiento

latitude - latitud

longitude - longitud

minutsLatitude - latitud en minutos

minutsLongitude - longitud en minutos

RENTADOR - tiene túnel de lavado de coches (1 - sí)

CANVI - tiene cambio para dinero, booleano (1 - sí)

SUPER - tiene supermercado (1 - sí)

LAVABO - tiene lavabo (1 - sí)

PARKING - tiene parking (1 - sí)

VENDING - tiene vending (1 - sí)

-> **'bonarea_gasolineras_data_and_prices.csv'**

id - número identificador de la gasolinera.

Fecha - fecha**

Dia_Semana - día de la semana**

type - tipo de establecimiento de BonÀrea

url - url de la gasolinera**

street - calle donde se ubica la gasolinera

city - ciudad donde se ubica la gasolinera

postalCode - código postal de la dirección de la gasolinera

raoSocial - razón social del establecimiento

latitude - latitud

longitude - longitud

minutsLatitude - latitud en minutos

minutsLongitude - longitud en minutos

RENTADOR - tiene túnel de lavado de coches (1 - sí)

CANVI - tiene cambio para dinero, booleano (1 - sí)

SUPER - tiene supermercado (1 - sí)

LAVABO - tiene lavabo (1 - sí)

PARKING - tiene parking (1 - sí)

VENDING - tiene vending (1 - sí)

GASOIL A - precio diario de este tipo de combustible**

GASOLINA S/P 95 - precio diario de este tipo de combustible

GASOLINA S/P 98 - precio diario de este tipo de combustible

ADBLUE - precio diario de este tipo de combustible

Procedimiento de colección de datos

La página permite al usuario seleccionar uno o varios tipos de establecimientos y muestra la ubicación sobre el mapa de todos establecimientos de los tipos seleccionados.

Al hacer click en los distintos puntos que aparecen en el mapa, la página realiza otra solicitud jquery para una instancia de establecimiento específica y visualiza un recuadro con su información básica.

Durante la interacción del usuario con la web, los datos se generan de forma dinámica: la web envía una solicitud jquery de información y devuelve un conjunto de puntos en el mapa que cumplen con el criterio impuesto por el usuario.

Para realizar el webscraping se realiza en dos pasos:

Id gasolineras - encontramos el string que representa la solicitud jquery y usamos Postman (https://www.postman.com/) para obtener los campos `url`, `data`, `header`.

Descripción detalle de las gasolineras - usamos una lista de ids de gasolineras para solicitar información detallada de las gasolineras.

Agradecimientos

Los datos han sido recolectados de la página web de BonÁrea. Para ello se han utilizado `requests` de Python y la aplicación `Postman` (https://www.postman.com/).

Cabe mencionar que www.bonarea.com no dispone de ningún tipo de archivo “robots.txt”, hecho que ha facilitado el scraping de la información con total libertad por parte de los estudiantes.

6. 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.

///»> PENDIENTE DE ELABORACION

Inspiración

---

7. Explique 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 ELABORACION...

El juego de datos generado nos aporta la información sobre los precios diarios de los carburantes (por productos y por gasolineras) de la empresa BonÀrea.

Esta información es realmente interesante para el usuario (y también para la competencia) puesto que se puede generar diferentes aplicaciones o visualizaciones que pueden ser base en toma de decisiones tanto particulares (el usuario puede elegir qué gasolinera repostar por ser más barata) como a nivel estratégico (la empresa competidora puede utilizarlos en su beneficio).

Ejemplos de posibles gráficas que se podría generar, serían las siguientes:

Heatmap geográfico de precios (siendo el color rojo indicador de productos más caros):

Evolución temporal del precio del carburante por producto y por gasolinera:

Estadísticos de los precios de los carburantes por producto y por día de la semana:

Como se puede observar, la información disponible es valiosa y útil en futuras tomas de decisiones.

Por otro lado, el dataset también aporta información de los servicios asociados a cada gasolinera, donde éstos pueden ser:

RENTADOR - Tiene túnel de lavado de coches

CANVI - Tiene cambio para dinero, booleano

SUPER - Tiene supermercado

LAVABO - Tiene lavabo

PARKING - Tiene parking

VENDING - Tiene vending

Y en consecuencia, además de la infomación adicional que el usuario pueda obtener de los diferentes servicios que aporta cada gasolinera, estadísticas sobre los precios de los carburantes mediante la agrupación de los diferentes servicios provistos, puede aportar información oculta, que a primera vista no trivial.

Licencia

---

8. 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

///»> PENDIENTE DE DECISION.

Código

---

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

La generación del dataset se ha generado utilizando el lenguaje de programación Python.

El código utilizado y archivos auxiliares se pueden encontrar en https://github.com/ bajo la siguiente dirección web:

https://github.com/Carlos-Acosta/webscraping.git

Dataset

---

10. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

///»> PENDIENTE DE ELABORACIÓN DE LA ÚLTIMA VERSIÓN DEL DATASET...

Referencias

---

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.