

**M2.851 TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS**

***PRÁCTICA 1***

**Web Scraping**

**Carmelo León Suárez**

**Vanesa Navarro Oronoz**

***Abril 2021***

Contenido

[1. Contexto 3](#_Toc68187102)

[2. Definir un título para el dataset 3](#_Toc68187103)

[3. Descripción del dataset 4](#_Toc68187104)

[4. Representación gráfica 4](#_Toc68187105)

[5. Contenido 4](#_Toc68187106)

[6. Agradecimientos 4](#_Toc68187107)

[7. Inspiración 4](#_Toc68187108)

[8. Licencia 5](#_Toc68187109)

[9. Código fuente 5](#_Toc68187110)

[10. Dataset – DOI Zenodo 5](#_Toc68187111)

[11. Contribuciones 5](#_Toc68187112)

[12. Referencias bibliográficas 5](#_Toc68187113)

*El objetivo de esta actividad es la creación de un dataset a partir de los datos contenidos en una web. A continuación se desarrollan los apartados a los que hay que dar respuesta para su correcta realización.*

# Contexto

“**El precio de la luz y el gas se disparan en el inicio de 2021 en plena ola de frío**.”

*El País (07-01-2021).*

“**El precio de la luz se dispara de nuevo en marzo y acabará el mes un 12% más caro que en febrero**.”

*El Mundo (28-03-2021).*

¿Cómo podemos saber la variación del precio de la luz de un día para otro? ¿Y de una hora para otra?

En España existen dos mercados de la electricidad: el mercado libre y el mercado regulado. A diferencia del mercado libre, en el mercado regulado, funciona la tarifa de luz por horas. Esto quiere decir que cada hora de cada día del año, el precio de la luz cambia. En el mercado libre, es la compañía comercializadora la que fija el precio del kWh. En este caso existe un precio fijo pactado con antelación entre cliente y comercializadora. Por ello, el cliente siempre sabe cuánto le cuesta el kWh en cada momento del día y cada día de la semana.

Si se hace seguimiento a lo largo del año de los precios en el mercado regulado se puede observar que los precios varían claramente. Existen varios motivos por los que el precio del kWh sube o baja:

* La generación de energía tiene un coste variable: como toda actividad, generar electricidad tiene un precio.
* La generación de energía no es siempre constante: no siempre se produce la misma cantidad de energía.
* La demanda de energía no es siempre constante. Por ello, a mayor demanda de energía, mayores serán los precios que las empresas comercializadoras pagarán por el kWh.

El precio del kilovatio hora lo fija diariamente el OMIE, a través del llamado “pool eléctrico”, que funciona como una subasta en la que, según la oferta y la demanda de energía se establece un precio para cada hora del día siguiente. A este precio, el Gobierno le añade diversas tasas de acceso y los costes de distribución eléctrica y de transporte. El precio de la energía fluctúa en función de varios factores, como por ejemplo los meteorológicos, los periodos de mayor consumo o el precio de las materias primas.

Con el dataset resultante de la práctica se pretende obtener los datos de los precios de la luz degún el tipo de tarifa aplicada y en un período de tiempo de interés del usuario para su posteior reutilización en diferentes análisis o procesos más complejos.

# Definir un título para el dataset

El titulo escogido para el juego de datos extraído es: **Tarifas eléctricas horarias en España**.

# Descripción del dataset

Este dataset, como se ha mencionado previamente en el contexto, tiene como contenido los datos relativos a los precios horarios de la luz en España, seleccionados para un rango de fechas definido por el usuario, pudiendo así tener una visión global de la fluctuación de las tarifas y comportamiento del mercado.

Para la extracción de los datos se ha tenido que examinar el fichero “robots.txt”, con la finalidad de saber que accesos se permiten a robots ([*https://tarifaluzhora.es/robots.txt*](https://tarifaluzhora.es/robots.txt)). En este caso el sitio web permite que para cualquier rastreador se habiliten todos los elementos gráficos, código y estilos y lo que se deshabilita es la parte para registro de usuarios y configuraciones de la web.

Se adjunta el fichero “robots.txt” en el directorio “/*docs*” del repositorio generado para el proyecto.

Como parte de la evaluación inicial también se ha examinado el mapa del sitio web ([*https://tarifaluzhora.es/sitemap.xml*](https://tarifaluzhora.es/sitemap.xml)) para asistir en la localización de los contenidos que nos interesan.

WORK IN PROGRESS

# Representación gráfica

WORK IN PROGRESS

Sacar esquema o plot del dataset.

# Contenido

WORK IN PROGRESS

Atributos del dataset final.

# Agradecimientos

Los datos han sido recolectados desde la base de datos online que pone a disposición del público el grupo Selectra. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de web scraping para poder extraer la información alojada en el sitio web. Los datos son de acceso público y gratuito.

WORK IN PROGRESS

# Inspiración

WORK IN PROGRESS

# Licencia

La licencia escogida ha sido **Released Under CCO: Public Domain License - CC0 1.0 Universal (CC0 1.0) Dedicación de Dominio Público** lo que significa que:

* La actividad está dedicada al dominio público, mediante la renuncia a todos sus derechos bajo las leyes de derechos autorales en todo el mundo, incluyendo todos los derechos conexos y afines, en la medida permitida por la ley.
* El dataset se puede copiar, modificar, distribuir e interpretar, incluso para propósitos comerciales, sin pedir permiso.
* A menos que esté expresamente señalado, no da garantías sobre la obra, y se exime de toda responsabilidad por los usos de la misma, en la medida permitida por la ley.
* Al usar o citar la obra, no debe pedirse aprobación de la autora o la afirmadora.

Se ha escogido la licencia explicada anteriormente debido a que los datos son públicos y se han obtenido de manera gratuita. Asimismo, la finalidad de este proyecto es puramente académica y sin ánimo de lucro, por lo que no se pretende obtener ningún tipo de beneficio del trabajo realizado. Por último, se ha considerado que está licencia es adecuada para que terceras personas puedan utilizar el juego de datos con distintos fines como podrían ser académicos, introducción al mundo del web scraping y mamipulación de los datos.

# Código fuente

El código creado para generar este dataset se encuentra publicado en el repositorio de GitHub:

<https://github.com/vnoronoz/web-scraping/tree/main/src>

WORK IN PROGRESS

# Dataset – DOI Zenodo

WORK IN PROGRESS

# Contribuciones

|  |  |
| --- | --- |
| Contribuciones | Firmas |
| Investigación previa | CLS, VNO |
| Redacción de las respuestas | CLS, VNO |
| Desarrollo código | CLS, VNO |

# Referencias bibliográficas

**Subirats, L., Calvo, M.** (2018). “*Web Scraping”*. Editorial UOC.

**Lawson, R.** (2015). *Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data*.