M2.851 TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRÁCTICA 1

Web Scraping

Carmelo León Suárez

Vanesa Navarro Oronoz

Abril 2021

Contenido

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.

1. 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.

2. Definir un título para el dataset

El titulo escogido para el juego de datos extraído es: **Tarifas eléctricas tarifa normal en España entre Enero y Marzo de 2021**.

3. 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*). 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*) para asistir en la localización de los contenidos que nos interesan.

El dataset está compuesto por el precio por hora en cada día como se puede ver a contunuación:

PRECIO DEL KWH DE LUZ POR HORAS

00h - 01h: 0.11583 €/kWh

01h - 02h: 0.11587 €/kWh

02h - 03h: 0.11428 €/kWh

03h - 04h: 0.11595 €/kWh

04h - 05h: 0.11692 €/kWh

05h - 06h: 0.12461 €/kWh

06h - 07h: 0.13154 €/kWh

07h - 08h: 0.13488 €/kWh

08h - 09h: 0.14589 €/kWh

09h - 10h: 0.12985 €/kWh

10h - 11h: 0.12044 €/kWh

11h - 12h: 0.11203 €/kWh

12h - 13h: 0.10753 €/kWh

13h - 14h: 0.10472 €/kWh

14h - 15h: 0.10181 €/kWh

15h - 16h: 0.10197 €/kWh

16h - 17h: 0.10435 €/kWh

17h - 18h: 0.11356 €/kWh

18h - 19h: 0.12128 €/kWh

19h - 20h: 0.12402 €/kWh

20h - 21h: 0.13441 €/kWh

21h - 22h: 0.13289 €/kWh

22h - 23h: 0.12762 €/kWh

23h - 24h: 0.1259 €/kWh

Como se puede ver en cada hora se establece un precio en €/kWh.

En el caso que nos ocupa y como la web nos permite pasar la fecha extraer, vamos a generar un dataset con los datos de Enero, Ferbrero y Marzo de 2021 de la tarifa normal. Con este conjunto de tres meses creemos que es sufuciente para poder hacer un estudio del precio de la tarifa por franjas horarias e incluso por día de la semana si así se estimara conveniente.

4. *Representación gráfica*

TABLA RESUMEN E HISTOGRAMA DE FRANJAS HORARIAS MAS CARAS Y BARATAS EN EL PERIODO

(Enero – Marzo 2021)

Franja más cara	Frecuencia	Franja más barata	Frecuencia
19h - 20h	31	04h - 05h	24
20h - 21h	30	16h - 17h	15
18h - 19h	8	03h - 04h	11
08h - 09h	6	15h - 16h	11
07h - 08h	4	14h - 15h	8
00h - 01h	3	05h - 06h	6
09h - 10h	3	13h - 14h	4
21h - 22h	3	23h - 24h	3

5. Contenido

El dataset de compone de 90 registros y 27 atributos. Este conjunto de datos referentes a las tarifas eléctricas normales en España desde el mes de Enero hasta el mes de Marzo de 2021 extraídos en el fichero **tarifa_electrica_EneMar2021.csv** son:

- **Fecha:** fecha en formato '*dd de MM de yyyy'*
- **Franja más cara:** franja horaria del día donde la el precio del kW es más caro
- **Franja más barata:** franja horaria del día donde la el precio del kW es más barato
- **00h - 01h:** Precio del kWh para la franja horaria 00h - 01h
- **01h - 02h:** Precio del kWh para la franja horaria 01h - 02h
- **02h - 03h:** Precio del kWh para la franja horaria 02h - 03h
- **03h - 04h:** Precio del kWh para la franja horaria 03h - 04h
- **04h - 05h:** Precio del kWh para la franja horaria 04h - 05h
- **05h - 06h:** Precio del kWh para la franja horaria 05h - 06h
- **06h - 07h:** Precio del kWh para la franja horaria 06h - 07h
- **07h - 08h:** Precio del kWh para la franja horaria 07h - 08h
- **08h - 09h:** Precio del kWh para la franja horaria 08h - 09h
- **09h - 10h:** Precio del kWh para la franja horaria 09h - 10h
- **10h - 11h:** Precio del kWh para la franja horaria 10h - 11h
- **11h - 12h:** Precio del kWh para la franja horaria 11h - 12h
- **12h - 13h:** Precio del kWh para la franja horaria 12h - 13h
- **13h - 14h:** Precio del kWh para la franja horaria 13h - 14h
- **14h - 15h:** Precio del kWh para la franja horaria 14h - 15h
- **15h - 16h:** Precio del kWh para la franja horaria 15h - 16h
- **16h - 17h:** Precio del kWh para la franja horaria 16h - 17h
- **17h - 18h:** Precio del kWh para la franja horaria 17h - 18h
- **18h - 19h:** Precio del kWh para la franja horaria 18h - 19h
- **19h - 20h:** Precio del kWh para la franja horaria 19h - 20h
- **20h - 21h:** Precio del kWh para la franja horaria 20h - 21h
- **21h - 22h:** Precio del kWh para la franja horaria 21h - 22h
- **22h - 23h:** Precio del kWh para la franja horaria 22h - 23h
- **23h - 24h:** Precio del kWh para la franja horaria 23h - 24h

6. 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 con la librería BeautifulSoup para poder extraer la información alojada en el sitio web. Los datos son de acceso público y gratuito.

En lo referente a investigaciones previas, se incluyen alguno de los artículos y análisis que han servido para justificar la creación del dataset:

– *Estudio comparativo de las tarifas eléctricas en el mercado libre* [en línea] [fecha de consulta:10 de abril 2021]. Disponible en: **https://www.facua.org/es/documentos/electricasnoviembre2016.pdf**

- *¿Beneficia la nueva tarifa de luz por horas sin cambiar de hábitos de consumo?* [en línea] [fecha de consulta:10 de abril 2021]. Disponible en: **https://www.facua.org/es/documentos/electricasnoviembre2016.pdf**

- *¿Tarifa plana de luz? #NoCuela* [en línea] [fecha de consulta:10 de abril 2021]. Disponible en: **https://www.ocu.org/vivienda-y-energia/gas-luz/noticias/tarifas-planas-electricidad**

Todas estas publicaciones tienen en común el debate sobre si es más conveniente contratar un servicio de tarifa plana o, por el caso contrario, un servicio de precio regulado con discriminación de tarifa por horas, que es lo que se muestra en el dataset.

Y basado en todo lo anterior, la pregunta que nos ha llevado a trabajar con este dataset es:

> ¿Cuáles son la franjas horarias más baratas y cuales más caras en el mercado regulado de energía electrica?

7. Inspiración

El conjunto de datos obtenido puede utilizarse con una gran variedad de objetivos. Los que predominan para nuestro interés y han sido inspirados por los análisis que han servido de referente, son los siguientes:

- Estudiar las variaciones en el precio de la luz entre estaciones, meses, días de la semana, horas del día, etc.
- Clusterizar zonas horarias para estudiar la diferencia y margen de las horas pico, valle y llano estipuladas por las tarifas más básicas.
- Ayudar a entender el gasto de consumo eléctrico real en una vivienda.
- Usar los datos obtenidos para descubrir si a un consumidor le conviene más una tarifa fija o una con discriminación horaria.
- Recomendar las franjas horarias más habituales y económicas para el consumidor.
- Permitir comparar al consumidor el precio del kWh en el mercado con el precio que paga en su tarifa contratada fuera del mercado regulado.
- Realizar predicciones a futuro sobre los datos tarifarios.
- Promocionar el conocimiento y uso de la tarifa de discriminación horaria.
- Fomentar otros estudios que indaguen en el ahorro a la hora de la aplicación de la tarifa de discriminación horaria.

Las cuales van totalmente realcionados con la pregunta realizada en el apartado seis y los estudios preliminares que hemos aportado. En digfinitiva lo que le importa al consumidor final es saber si con el precio por franjas horarias y el uso que necesita de la energía electrica es más rentable una tarifa regulada que controla el gobierno o una las tarifas fuera del mercado regulado que ofertan los distintos operadores.

8. 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.

9. Código fuente

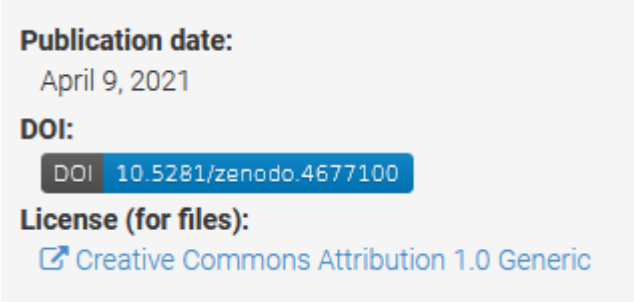
El código creado para generar este dataset se encuentra publicado en el repositorio de GitHub:

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

10. Dataset – DOI Zenodo

Para consultar el dataset y citarlo mediante DOI se puede acceder al siguiente enlace:

https://zenodo.org/record/4677100#.YHCbUT-UXIU



11. Contribuciones

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

12. Referencias bibliográficas

SUBIRATS MATÉ, Laia y CALVO GONZÁLEZ, Mireia. *WebScrapping* [en línea]. Barcelona: UOC, (s/f). Disponible en: https://materials.campus.uoc.edu/daisy/Materials/PID_00256970/pdf/PID_00256970.pdf

LAUSON, Richard.*Web Scraping with Python* [en línea]. Birmingham: Packt Publishing Ltd, 2015. ISBN 9781782164371. [Consulta: 10 de abril de 2021]. Disponible en: https://ebookcentral.proquest.com/lib/bibliouocsp-ebooks/detail.action?docID=4191102