Tipología y ciclo de vida de los datos

Práctica 1

Carlos Pintor (cpintorv)
Enrique Vera (everaor)

FORMAR
TRANS-
FORMAR

Universitat Oberta de Catalunya
25 anys

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

El contexto es el estudio de la generación y consumo de energía obteniendo una amplia información acerca de:

- Condiciones meteorológicas que permiten la generación de energía y anticipan el consumo de esta.
- Generación de energía desglosada por tipo de generación.
- Consumo de energía por un cliente de una operadora desglosado por horas.

Con estos objetivos se han seleccionado tres sitios web diferentes cada uno respondiendo a las necesidades expuestas anteriormente.

- Condiciones meteorológicas: AEMET ofrece distintos indicadores meteorológicos y climáticos en diferentes estaciones de España.
- Generación de energía: Red Eléctrica proporciona diariamente el volumen de energía generada desglosada por los diferentes orígenes.
- Consumo de energía por un cliente: i-DE como comercializadora permite consultar el consumo de energía diario por horas.

El motivo de esta extracción de datos es poder tener de manera unificada el mayor número de factores posibles que afectan a la generación y a la demanda de energía.

Título

Generación, oferta y demanda energética.

Descripción del dataset: Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

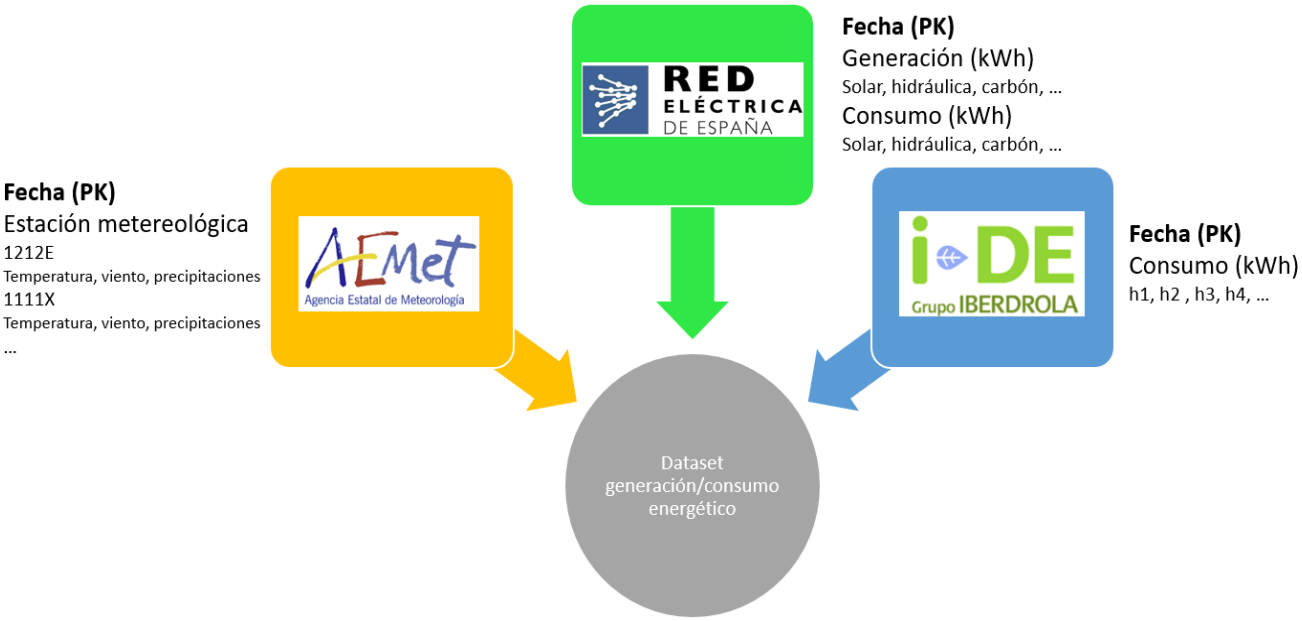
Se ha extraído información de tres conjuntos diferentes correspondientes a la generación y consumo energético:

- Indicadores climáticos.
- Fuentes de generación energética.

- Consumo detallado de un cliente.

El resultado del conjunto de datos extraídos es un data set que contiene en cada registro un día entre el 1 de enero de 2020 y el 31 de septiembre de 2021 y como columnas todos los indicadores nombrados en esta memoria.

Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



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

El dataset extraído está compuesto por 3 diccionarios diferenciados correspondiente a cada una de las fuentes de información que han sido utilizadas.

Cada registro corresponde a un día natural del 1 de enero de 2020 hasta el 30 de septiembre de 2021.

Las columnas son las siguientes:

Generación energética: Volumen de kW/h por cada una de las fuentes de generación:

- Hidraúlica
- Eólica
- Solar fotovoltaica

- Solar térmica
- Hidroeólica
- Otras renovables
- Residuos renovables
- Turbinación bombeo
- Nuclear
- Ciclo combinado
- Carbón
- Motores diesel
- Turbina de gas
- Turbina de vapor
- Cogenereación

También se ha separado la parte de generación renovable y no renovable, y además se ha recogido los datos de cada una de estas fuentes tanto para la generación como para el consumo.

Diccionario climático: Recoge información de 16 estaciones climatológicas (una por comunidad autónoma) sobre temperatura media, precipitaciones, velocidad del viento y situación geográfica de la estación. Las ciudades de las que se ha recogido la información, son las siguientes:

- A Coruña
- Huesca
- Barcelona
- Asturias (AEMET no especifica ciudad)
- Santander
- Bilbao
- Navarra (AEMET no especifica ciudad)
- Logroño
- Zaragoza
- León
- Madrid
- Badajoz
- Toledo
- Valencia
- Murcia
- Sevilla

Diccionario consumo de un cliente: Recoge los kW/h de cada una de las 24 horas del día de un cliente de Iberdrola.

La recolección de esta información se ha realizado de las siguientes maneras:

- Red eléctrica: Se comprobó si disponían de una API para las descargas masivas. Se vio que efectivamente se disponía y se analizó la documentación creando un script de Python que realizaba una llamada para cada día en el periodo definido.

Se estudió la estructura de la respuesta y se adaptó la lectura al esquema obtenido de manera que el resultado final era un registro para cada día con todos los datos obtenidos.

El bucle finalizaba apilando los sucesivos días obtenidos en un único dataset con la clave primaria de fecha.

- AEMET: En primer lugar se investigó la existencia de una API. Una vez estudiada la documentación se solicitó un código API de validez cinco días. Se realiza una llamada con Python a una dirección web incorporando la API, lo que obtenía por respuesta una URL temporal. Se realizó un código Python que parte de un diccionario con todas las estaciones y realiza 16 llamadas (una por estación) a esta URL temporal.

Por último, se adecuan los datos a un data frame y se unifican todas las estaciones mediante uniones por clave primaria.

- i-DE: En esta ocasión, existía la posibilidad de la descarga manual de un fichero CSV con el consumo diario con los contras de que sólo se podía seleccionar el último año y que por un motivo didáctico se realizó a través de llamadas individuales.

Había una primera llamada de establecimiento de sesión en la que se usaban las credenciales del usuario y un bucle por el que se lanzaba una llamada diaria, se recogía el resultado y se iba sumando al de los días anteriores.

Este ejercicio se realizó unas cinco ocasiones entre pruebas y la ejecución final y se descargó el resultado en un CSV para no interferir en el funcionamiento normal del portal.

Agradecimientos. 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. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Presentación de los propietarios de los datos:

- i-DE: Es la compañía distribuidora de Iberdrola. Se encargan de llevar la energía hasta los puntos finales de los consumidores facilitando además toda la información necesaria para sacar el máximo partido a los datos de cada consumidor. Tienen más de 11 millones de clientes distribuidos en 10 provincias. Más información sobre i-DE:

  https://www.i-de.es/i-de-grupo-iberdrola/conocenos

- AEMET: Agencia Española de Meteorología cuyo objetivo principal es contribuir a la protección de vidas y bienes basándose en predecir y vigilar los fenómenos meteorológicos.

Prestan servicios de calidad a la sociedad para poder planificar, dirigir y desarrollar las actividades meteorológicas. Más información en:

https://www.aemet.es/es/conocenos/quienes_somos

- Red eléctrica de España: Es la primera compañía del mundo que se dedica únicamente a la operación del sistema eléctrico y al transporte de electricidad. Actualmente gestiona más de 44.000 km de líneas eléctricas y se encarga de garantizar la seguridad y continuidad del suministro eléctrico manteniendo una red de transporte fiable que contribuya al progreso de la sociedad. Más información en:

  https://www.ree.es/es/conocenos/ree-en-2-minutos

Existen diversos análisis de consumo de energía, y cada vez más debido a las polémicas subidas de precios y el auge de las energías renovables. A continuación, se muestra algún ejemplo:

- Análisis del consumo eléctrico en España en 2016 y en Navidad desarrollado por el Instituto de Ingeniería del conocimiento. En este análisis se obtiene la demanda eléctrica horaria en España, el consumo y la procedencia de la energía para los años 2014, 2015 y 2016. Algunos de los indicadores que se obtienen son:
  - Día del año con mayor consumo.
  - Curva de la demanda de la electricidad un día laborable.
  - Curva de la demanda de la electricidad un día festivo.
  - Fuentes de energía con más generación en cada año.

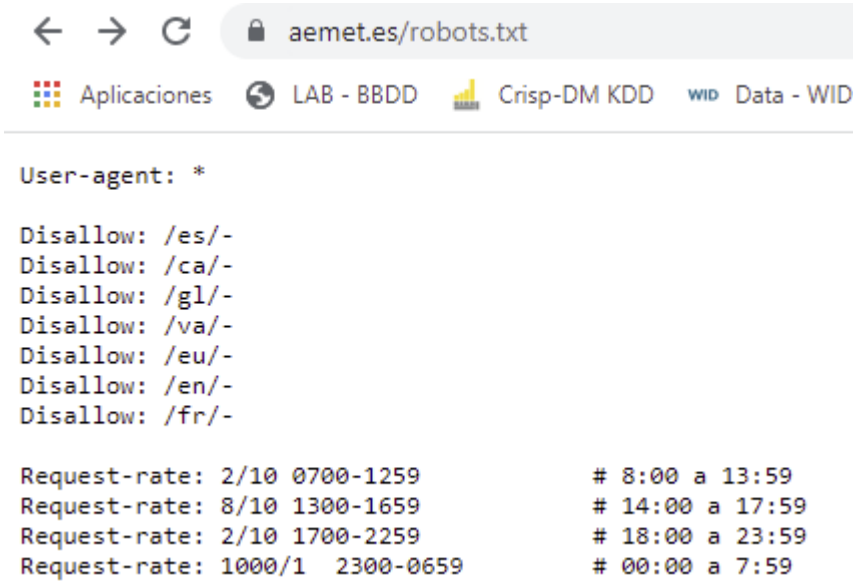
  Acceso al análisis: https://www.iic.uam.es/energias/analisis-consumo-electrico-espana-2016/

- Trabajo Fin de Máster Máster en Ingeniería Industrial - Generación de Escenarios para el Análisis de Redes de Distribución (María Molina Salvador): En este trabajo fin de máster se realiza un análisis detallado de las redes de distribución de energía, tanto urbana como rural, aplicando el caso de estudio de la meteorología y las redes renovables observando como afectan a las curvas de precio y al consumo.

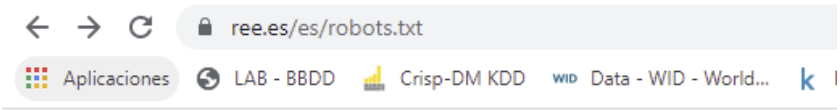
  Acceso al análisis: http://bibing.us.es/proyectos/abreproy/71496/fichero/TFM-1496-MOLINA.pdf

Se han estudiado los ficheros de robots.txt tanto para la AEMET como para Red eléctrica y se encuentra que en el primer caso no permite acceder a la web mediante web scraping lo que nos llevó al uso de la API, si bien por motivos didácticos se usó la API a modo de web para scrapear lanzando consultas masivas. En el de red eléctrica no hace referencia explícita al permiso o denegación de acceso a la página principal en la que se encuentran los datos, sin embargo disponía de una API que se ha utilizado a modo de web por los mismos motivos didácticos.

- Robots.txt de AEMET:



- Robots.txt de Red eléctrica:




```
#
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
# CSS, JS, Images
Allow: /core/*.css$
Allow: /core/*.css?
Allow: /core/*.js$
Allow: /core/*.js?
Allow: /core/*.gif
Allow: /core/*.jpg
Allow: /core/*.jpeg
Allow: /core/*.png
Allow: /core/*.svg
Allow: /profiles/*.css$
Allow: /profiles/*.css?
Allow: /profiles/*.js$
Allow: /profiles/*.js?
Allow: /profiles/*.gif
Allow: /profiles/*.jpg
Allow: /profiles/*.jpeg
Allow: /profiles/*.png
Allow: /profiles/*.svg
# Directories
Disallow: /core/
Disallow: /profiles/
# Files
Disallow: /README.txt
Disallow: /web.config
# Paths (clean URLs)
Disallow: /admin/
Disallow: /comment/reply/
Disallow: /filter/tips
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register/
Disallow: /user/password/
Disallow: /user/login/
Disallow: /user/logout/
# Paths (no clean URLs)
Disallow: /index.php/admin/
Disallow: /index.php/comment/reply/
Disallow: /index.php/filter/tips
Disallow: /index.php/node/add/
Disallow: /index.php/search/
Disallow: /index.php/user/password/
Disallow: /index.php/user/register/
Disallow: /index.php/user/login/
Disallow: /index.php/user/logout/
```

En el caso del i-DE de Iberdrola no se dispone de fichero robots.txt. Aún así se pidió autorización y se programó para que las llamadas no interfirieran en el funcionamiento normal de la web.

Los agradecimientos van para Iberdrola que han permitido que accedamos al i-DE y lanzar más de 3000 peticiones en una única ejecución sin bloquearnos la conexión. A continuación, se muestra el correo que les enviamos para poder llevar a cabo esta extracción:

Buenos días,

Estoy realizando un trabajo universitario por el que necesito lanzar consultas a su portal de i-DE para descargarme información de consumo desde enero 2020 a septiembre 2021.

Para no causar molestias había pensado en lanzar una consulta más para no crear más problemas y exportar el resultado a un csv y no hacer más pruebas.

Me gustaría pediros poder realizar **una prueba** y luego descargarme el resultado en csv para no volver a lanzarla.

Podría poner un retardo de 1 segundo por petición de manera que no perjudicara el acceso.

Si os parece bien por la tarde lanzo esas 600 peticiones con retardo, lo guardo en un csv y no vuelvo a lanzar más peticiones (salvo el uso normal que hago de la app, claro).

Gracias!

Inspiración. Explicar 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

Este proyecto nace del interés mutuo de los estudiantes por el funcionamiento del sistema energético y su impacto sobre el medio ambiente y el cambio climático. El objetivo 7 de los objetivos de desarrollo sostenible (ODS) indica la siguiente:

“Garantizar el acceso a energía asequible, segura, sostenible y moderna (además de no contaminante)”

Hubo un proyecto original cuyo objetivo era desarrollar un scoring para un usuario dado que representara como se adecuaba su consumo individual al impacto por las diferentes fuentes de generación de energía.

Este proyecto tenía un inconveniente: No realizaba una consulta masiva de información ni permitía el estudio conjunto de las variables algo que será objetivo en la próxima práctica.

Con el fin de proporcionar la mayor utilidad de la información que extraíamos y sabiendo que lo íbamos a publicar en un repositorio abierto decidimos darle varios enfoques al dataset resultante:

- Relación entre oferta y demanda a nivel global en España.
- Relación entre indicadores metereológicos y producción energética sostenible.
- Relación entre indicadores metereológicos y consumo de energía.
- La posibilidad de desarrollar un caso de uso en el que se utiliza a un cliente y se estudian sus hábitos de consumo energético (volviendo a la idea original).

Al tratarse de un conjunto de datos muy diverso existen un gran número de posibilidades en cuanto al análisis y a las posibles tomas de decisiones con esta información. Más aún teniendo en cuenta que incorpora todo el periodo desde el inicio del COVID-19 y todas las implicaciones que tiene sobre demanda energética.

En comparación con los demás análisis que se han presentado anteriormente, este conjunto de datos contiene la información para el consumo de un usuario en concreto. Esto permitirá analizar con más detalle al usuario en vez de un análisis global que es lo que se realiza en los análisis presentados.

Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

Released Under CC0: Public Domain License. Elegimos esta licencia acorde a los objetivos del proyecto ya que renunciamos a los derechos para permitir que cualquier analista pueda extraer valor de la información que hemos generado.

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

https://github.com/cpintorv/climate_power_generation

Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

DOI: 10.5281/zenodo.5601489
URL: https://zenodo.org/record/5601489#.YXhDy9pByUk

Contribuciones	Firma
Investigaciones previas	E.V, C.P
Redacción de las respuestas	E.V, C.P
Desarrollo del Código	E.V, C.P