PRÁCTICA 1: WEB SCRAPING

Máster e Ciencia de Datos. Tipología y ciclo de vida de los datos

Evolución de las reservas hídricas de los embalses de España

Daniel Cabello Vázquez

Objetivo.

Creación de un dataset a partir de los datos contenidos en la web de Embalses.net mediante la técnica de web scraping.

El dataset contendrá información sobre las reservas hídricas de los embalses (>5Hm^3) con periodicidad semanal, almacenándolo de forma tabular en un archivo csv para su posterior análisis junto con otros datos.

Contexto.

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

Embalses.net es un portal web donde se muestra información hidrológica y pluviométrica de los embalses de la Península Ibérica con periodicidad semanal. La información es obtenida del boletín hidrológico que publica el Mapama semanalmente, proporcionando información actualizada sobre el nivel de los embalses, su capacidad y la variación a nivel nacional, de demarcación hidrográfica y de embalse. La información en la web se actualiza semanalmente, dejando de mostrarse el de la semana previa.

https://www.embalses.net/

Título del dataset.

Elegir un título que sea descriptivo

Evolución de las reservas hídricas de los embalses de España.

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).

El conjunto de datos recoge información sobre la capacidad y el agua embalsada de cada embalse con periodicidad semanal, su variación respecto a la semana previa y la demarcación en la que se encuentra.

Representación gráfica.

Presentar una imagen o esquema que identifique el dataset visualmente

Figura 1 Embalse del Atazar (Comunidad de Madrid)

Contenido.

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

Campos del dataset:

- **Cuenca**: Demarcación hidrográfica en el que se encuentra el embalse.
- **Pantano**: nombre del embalse (> 5 hm^3 de capacidad).
- **Capacidad**: capacidad del embalse en Hm^3.
- **Embalsada**: volumen de agua en el embalse en Hm^3.
- **Variación**: Variación de agua embalsada en porcentaje respecto a la semana pasada.
- **Fecha**: Fecha a la que hace referencia el agua embalsada y la variación respecto a la semana pasada.

Periodicidad: Semanal (campos: embalsada y variación). La capacidad del embalse puede verse modificada bien porque se produce aterramiento por deposición de sedimentos bien porque hay un recrecimiento de la presa.

La información se ha obtenido del portal de embalses.net (https://www.embalses.net/) mediante técnicas de web scraping en lenguaje de programación Python 3. El script programado generaría, para cada semana, un dataset y su correspondiente archivo csv del agua embalsada en los embalses. (recordar que el agua embalsada y su variación se actualiza cada semana no existiendo la posibilidad de recuperar la información de semanas pasadas).

Agradecimientos.

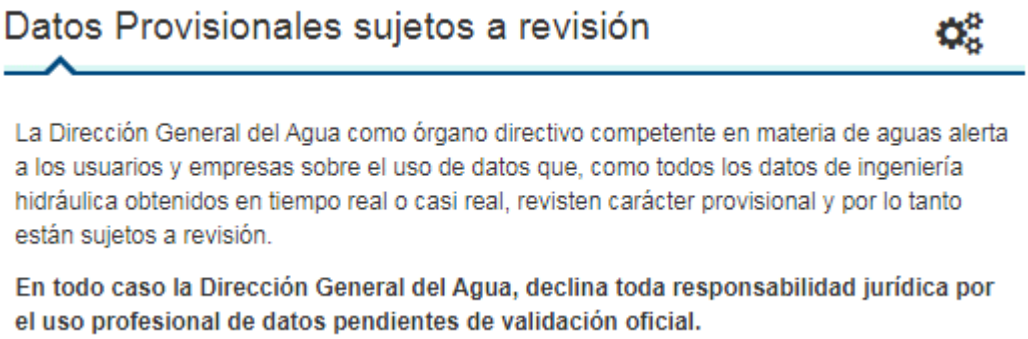
Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El portal de embalses.net muestra la información hidrológica y pluviométrica actualizada (semanal) de los embalses de la península Ibérica presentándolos en un formato amigable para los usuarios interesados, pero no es el propietario de la información, solo hace de vehículo citando la fuente de los datos. Los datos los obtiene del boletín hidrológico que es publicado con periodicidad semanal por *el área de información hidrológica del Ministerio para la Transición Hidrológica* (Miteco), organismo responsable y propietario de los datos que pone a disposición pública en aplicación de la Ley 27/2006 de acceso a la información ambiental.

Estos datos son generados por las Confederaciones Hidrográficas y las Administraciones hidráulicas intracomunitarias, la Agencia Estatal de Meteorología (AEMET) y la Red Eléctrica de España, realizando el tratamiento técnico de la información para su presentación como soporte de las decisiones de gestión hídrica que se deben tomar a nivel nacional.

El objetivo del boletín hidrológico es el de proporcionar información sobre las reservas hidráulicas en tiempo real, el seguimiento, análisis y publicación de los datos hidrológicos que permiten conocer el estado de los volúmenes almacenados en todos los embalses con capacidad mayor a 5 hm^3, la situación de los sistemas de explotación, de las reservas destinadas a riego y abastecimiento de poblaciones, los caudales fluyentes en los principales ríos de cada cuenca, las precipitaciones y la energía hidroeléctrica almacenada (calculada) y la producida real.

A la hora de reutilizar dicha información hay que tener en cuenta que todos *estos datos son provisionales y sujetos a revisión y validación*, tal como se indica en la web del MITECO.



Inspiración.

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder

Importancia. Uno de los objetivos principales de los planes hidrológicos de cuenca es el de garantizar la satisfacción de las demandas de agua que los distintos usuarios requieren (urbanos, agrarios, industriales, ambientales…). En este sentido, es esencial la monitorización continuada del agua disponible en los embalses para su adecuada gestión y la planificación de los recursos hídricos, haciéndola compatible con el otro gran objetivo de alcanzar el buen estado de las masas de agua y de sus ecosistemas asociados. Todo ello ante unos escenarios de cambio

climático para el ámbito mediterráneo de reducción de precipitaciones e incremento en la frecuencia e intensidad de los eventos extremos (sequías, inundaciones, olas de calor) junto con escenarios de desarrollo económico y de incremento de la población.

Preguntas a responder: ¿Porcentaje de agua embalsada respecto a su capacidad? ¿Cuál es la Evolución del agua embalsada y cómo influye en esta el régimen de precipitaciones y la demanda de agua de los distintos usos? ...

Licencia.

La licencia elegida para la publicación del conjunto de datos ha sido la ***CC BY-SA 4.0 License***:

- *El beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y cite la obra de la forma especificada por el autor o el licenciante*. Esto está en consonancia con la normativa que regula la reutilización de la información del sector público (ver más abajo) y, por otra parte, garantiza la trazabilidad de los datos derivados hasta su fuente original. También redunda en una mejora en la calidad y en el tratamiento de los datos pues son más los usuarios que lo revisan en función de su mayor o menor utilidad desde diferentes perspectivas.
- *El beneficiario de la licencia tiene el derecho de distribuir obras derivadas bajo una licencia idéntica a la licencia que regula la obra original*. Teóricamente, esto garantiza la continuidad de servicio público con el que estos datos son obtenidos y cedidos. También facilita la trazabilidad de los datos y la mejora en la calidad en el tratamiento de los mismos.

Normativa que regula la difusión y reutilización de la información ambiental

La información del boletín hidrológico del MITECO está sujeto a la Ley 27/2006 de 18 de julio, por la que se regulan los derechos de acceso a la información, de participación pública y de acceso a la justicia en materia de medio ambiente, en particular en su artículo 8; la Ley 37/2007 sobre reutilización de la información del sector público y su Reglamento de desarrollo, Real Decreto 1495/2011, en particular su artículo 7 así como el artículo 4 del Real Decreto 208/1996, de 9 de febrero, por el que se regulan los servicios de información administrativa y de atención al ciudadano.

Por tanto, se permite la reutilización de la información hidrológica del MITECO que conlleva la cesión gratuita y no exclusiva de los derechos de propiedad intelectual, autorizándose la realización de actividades de: "reproducción, distribución, comunicación pública o transformación, necesarias para desarrollar la actividad de reutilización autorizada, en cualquier modalidad y bajo cualquier formato, para todo el mundo y por el plazo máximo permitido por la Ley".

La reutilización de los conjuntos de datos por parte de los usuarios lo harán bajo su propia responsabilidad y riesgo, teniendo que responder frente a terceros por daños que pudieran derivarse de su reutilización. La administración pública propietaria de los datos no se hará responsable del uso que hagan de los datos los agentes reutilizadores, ni tampoco de los daños o pérdidas económicas que pudiera provocar la información reutilizada.

Código.

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

Se adjunta el código en Python.

Dataset.

Presentar el dataset en formato CSV.

Se adjuntan los datos en formato CSV. Los campos se separan con ‘,’.

Autores

Por razones de disponibilidad de tiempo, y de poder aplicarlo al campo que a mí me interesaba, opté por hacerlo de forma individual.

Contribuciones	Firma
Investigación previa	DCV
Redacción de las respuestas	DCV
Desarrollo código	DCV