Xavier Maltas Tarridas
Monica Ortiz Guijarro
Tipologia i cicle de vida de les dades

Universitat Oberta de Catalunya

Pràctica 1 : Web scraping

Context de la pràctica

La pàgina web sobre la que s'ha realitzat el scraping correspon servei meteorològic de Catalunya, concretament, la llista de la xarxa d'estacions meteorològiques automàtiques (XEMA) que podem trobar al següent enllaç: https://www.meteo.cat/observacions/llistat-xema

Ens ha semblat interessant realitzar el scraping sobre aquesta pàgina, ja que hi apareix tot el llistat d'estacions meteorològiques de Catalunya, un total de 234 estacions arreu del territori, que disposen de diferents sensors meteorològics que recullen diferents mètriques que ofereixen informació interessant com podrien ser: punts cardinals, altitud, temperatures, precipitacions, entre altres.

Però sobretot l'objectiu principal és el de poder obtenir les dades diàries actualitzades per a cada una de les estacions, concretament les magnituds meteorològiques següents:

- Temperatura mínima ºC
- Temperatura màxima ºC
- Temperatura mitjana del dia en qüestió ºC
- Humitat relativa mitjana
- Precipitació acumulada (mm)

Així doncs, mitjançant aquest scraping a la pàgina esmentada, podrem obtenir totes les magnituds desitjades per a cada una de les diferents estacions del territori sense haver d'accedir de forma diaria a la pàgina web.

Títol del dataset

El dataset generat a través d'aquest scraping és un conjunt de magnituds diàries de la xarxa d'estacions meteorològiques automàtiques. Així doncs que el títol escollit per aquest dataset és el següent: '*XEMA_dataset.csv*'

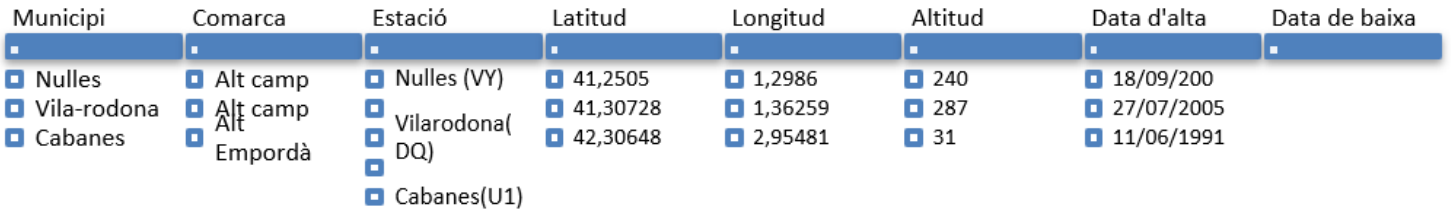
Descripció del dataset

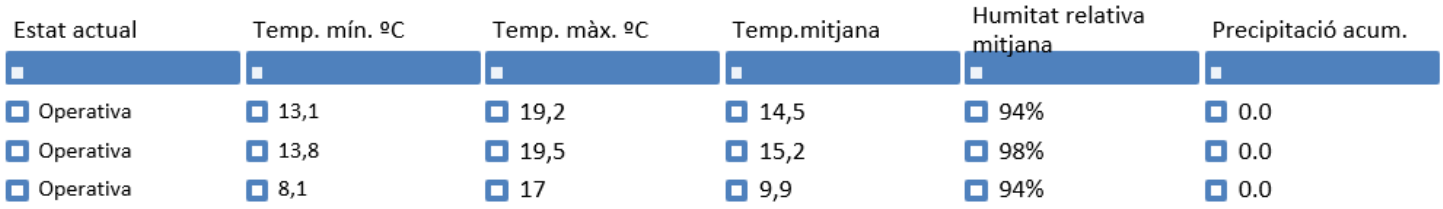
El conjunt de dades que conforma aquest dataset es compon d'un llistat de totes les estacions meteorològiques automàtiques de Catalunya, en total 234. Acompanyant el nom de l'estació, trobarem amb una sèrie de característiques i/o variables que descriuen una mica la posició i localització de l'estació, esmentades anteriorment en el punt 1, així com unes seguit de magnitud recollides que es van actualitzant cada 30 minuts. És per aquest motiu que podem afirmar que aquest scraping ens permet obtenir dades quasi en temps real.

Representació gràfica

Per tal de realitzar una correcta explicació del projecte, hem realitzat dos diagrames que permeten entendre de forma clara i visual el dataset i l'estructura del scraping.

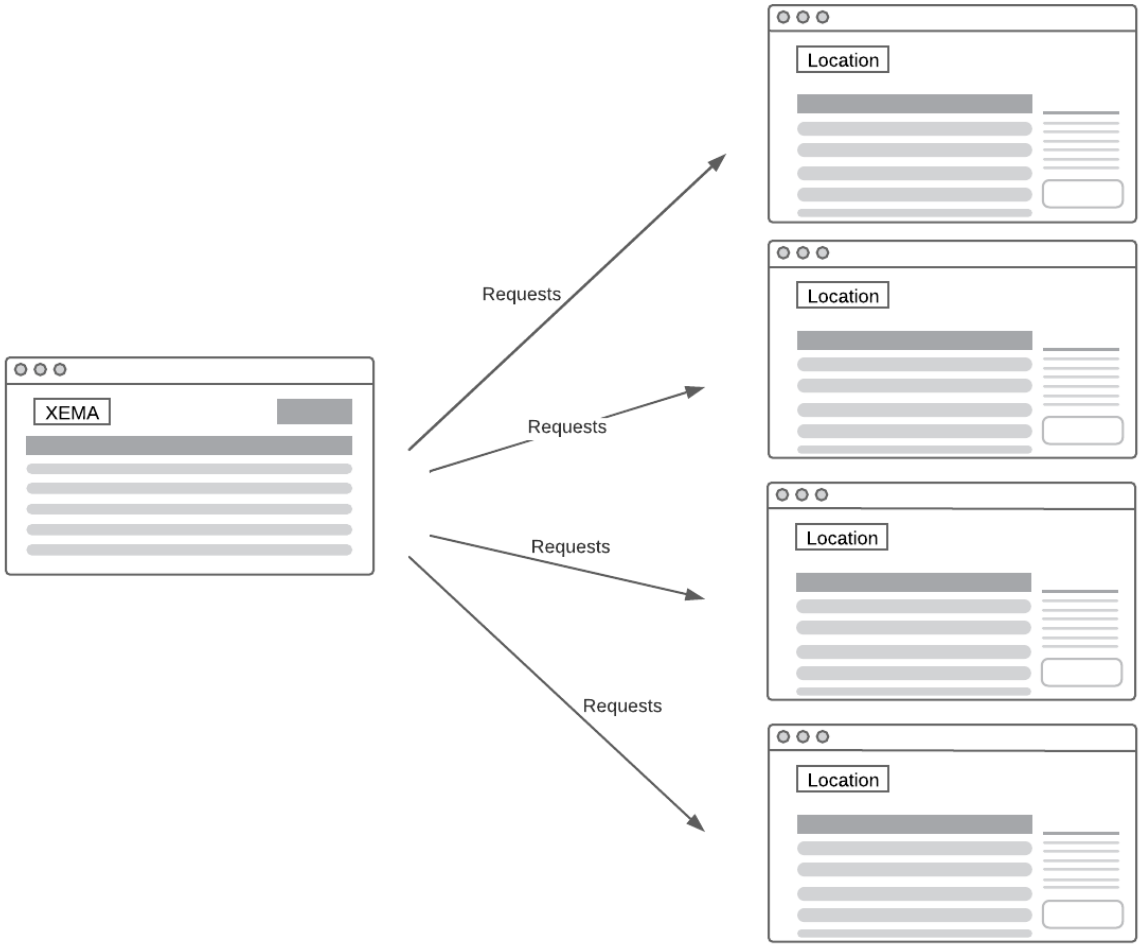
- Diagrama dataset





A l'anterior diagrama apareixen els camps i registres que inclou el dataset creat a partir de les dades de la pàgina web del Meteocat. Podem observar que trobem a la capçalera els camps que trobem al llistat de la XEMA juntament amb un seguit de magnitud recollits per tal de realitzar l'estudi.
Tal i com hem esmentant al punt anterior, el dataset es composa de 234 registres corresponent a cada una de les estacions meteorològiques. Aquest esquema és una mostra que representa la totalitat que ens permet tenir una idea de com estructurem el dataset i quines dades conté.

- Diagrama scraping



Aquest esquema ens permet entendre l'estructura de com s'ha realitzat el procés de scraping dins el nostre projecte.
Inicialment es realitza una petició a la pàgina del llistat de les estacions meteorològiques que ens permet extreure les dades descriptives de cada una de les estacions. A la vegada, mitjançant l'anàlisi d'aquest arxiu HTML podrem obtenir cada un dels links individuals de cada estació. Finalment, a través d'aquests links i l'anàlisi del corresponent fitxer HTML podem obtenir les magnituds meteorològiques de cada estació en particular.

Contingut

Els camps que conté el dataset són els següents:

- Estació: Nom de l'estació meteorològica.
- Codi: Codi de l'estació.
- Comarca: Nom de la comarca on està ubicada l'estació.
- Municipi: Nom del municipi on està ubicada l'estació.
- Latitud: Coordenades latitud de la ubicació de l'estació.
- Longitud: Coordenades longitud de la ubicació de l'estació.
- Altitud: Alçada respecte el nivell del mar on es troba ubicada l'estació expressat en metres.
- Data l'alta: Data quan va començar a estar operativa l'estació.
- Data de baixa: Data quan va deixar d'estar operativa l'estació
- Estat actual: Estat en què es troba l'estació. El valor pot ser 'operativa' o 'desmantellada'.
- Temp. mín. : Temperatura mínima registrada en ºC.
- Temp. màx. : Temperatura màxima registrada en ºC.
- Temp. mitjana : Representa la mitjana de temperatura del dia (es reben dades de les temperatures cada mitja hora, amb totes aquestes dades es fa la mitjana diària).
- Humitat relativa mitjana : Valor corresponent a la mitjana d'humitat relativa del dia en que ens trobem, igual que la temperatura, les dades es van recollint cada mitja hora.
- Precipitació acum. : Valor corresponent a la suma de precipitació acumulada expressat en mm (mililitres) durant el dia que ens trobem.

Les magnituds meteorològiques escollides que representem són actuals i gairebé informació en temps real, ja que com expliquem anteriorment, les dades recollides són de cada mitja hora.
Hem descartat altres magnituds com la irradiació solar o el gruix de neu, perquè no hi eren en totes les estacions. Hem preferit escollir només aquelles que apareixen per igual en totes les estacions.

Agraïments

Les dades han estat recollides pel servei meteorològic de Catalunya, concretament del llistat de la xarxa d'estacions meteorològiques automàtiques (XEMA).

Aquesta xarxa, d'entre altres, pertany a la Xarxa d'Equipaments Meteorològics (XEMEC) que es va crear com a mesurament de variables meteorològiques i per detectar fenòmens meteorològics amb l'objectiu de donar informació per avançat per la protecció de persones i béns davant a possibles fenòmens adversos, per la protecció del medi ambient i millorar el benestar econòmic i social. [1]
La gestió i manteniment d'aquestes xarxes es va encomanar al Servei Meteorològic de Catalunya (SMC).

Pel que fa als principis ètics, hem fet el scraping de dades que són de caire públic, no han sigut modificades, sinó que hem creat un dataset amb les dades meteorològiques de cada estació però les hem posicionat i estructurat de manera diferent de com es mostren al web.

La decisió d'utilitzar aquesta web prové d'un projecte per una companyia d'aigües on necessitem mantenir actualitzades la temperatura, humitat i precipitació per anar alimentant un sistema de clusterització i poder realitzar una classificació de municipis de Catalunya segons les magnituds extretes. El motiu d'elecció d'aquesta pàgina web és que les dades que hi trobem són fiables, quasi bé en temps real i es poden tenir en compte per fer la clusterització en grups de municipis que tinguin magnituds semblants.

Inspiració

La principal raó d'utilitzar aquest conjunt de dades meteorològiques té molt relació amb que s'ha esmentat a l'apartat anterior.
La motivació i inspiració la trobem en el projecte de classificació de municipis de Catalunya a través mitjançant mètodes de clusterització a través d'un seguit de variables meteorològiques.
Per realitzar aquest procés es tenen en compte les magnituds de les temperatures, humitat i precipitació, que ens permeten classificar-los en municipis secs o plujosos i càlids o freds.
Finalment, es realitza una estimació de dotacions d'aigua per a cada un dels grups tenint els paràmetres meteorològics esmentats.
L'objectiu és fer-ho per municipis que encara no tenen comptadors o l'aigua declarada és incorrecta.
És necessari pel projecte que es vagin actualitzant perquè un municipi que actualment està en un grup pot passar a un altre segons la variació de les dades.

[1] https://www.meteo.cat/wpweb/sobre-meteocat/xarxa-dequipaments-meteorologics-xemec/

Un altre motiu pel qual ens hem decantat a escollir aquest projecte és que el llenguatge de marques de la web es troba molt ben estructurat, el cos de la taula està ben organitzat incloent totes les estacions per cada fila, així doncs com que la informació de les magnituds de cada una de les estacions, es troba enllaçada amb el codi corresponent a l'estació i ens permet accedir de forma ràpid i sencilla a una altra pàgina web on es realitza l'extracció de les magnituds que es van actualitzant cada 30 minuts.

Llicència

Pel que fa a la llicència escollida per aquest projecte, hem pensat que la més adient és la Public Domain License, perquè en el dataset que hem creat no veiem els motius per posar limitacions de llicències, no hi ha dades sensibles ni que es puguin comercialitzar.
D'aquesta manera estarem oberts a possibles modificacions que podrien millorar el data set i per conseqüència el projecte.

Codi

https://github.com/xaviermaltas/webScrapingXEMA

Dataset

https://doi.org/10.5281/zenodo.5646289

Video explicatiu

https://drive.google.com/file/d/12eJWOrNTmPiVxNUk-IAu8v3r8quNBgVg/view?usp=sharing

CONTRIBUCIONS	SIGNATURA
Investigació prèvia	Xavier Maltas, Mónica Ortiz
Redacció de les respostes	Xavier Maltas, Mónica Ortiz
Desenvolupament del codi	Xavier Maltas, Mónica Ortiz