# **Pràctica 1: Web scraping**

1. **Context**.

Hem triat una web de codi obert (Dades Obertes) de la Generalitat de Catalunya. Disponible a Internet a: <http://governobert.gencat.cat/ca/dades_obertes/>.

La Generalitat de Catalunya ofereix un portal de dades d'accés públic, que permet pensar que és una bona opció per analitzar les dades obtingudes mitjançant Web Scraping.

Així doncs, l'objectiu concret d'aquesta activitat és la creació d'un dataset a partir de les dades obtingudes en la següent web disponible a Internet a: [https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/jj6z-iyrp/data,](https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/jj6z-iyrp/data,%20) extreta de la web arrel Govern Obert de la Generalitat de Catalunya i que conté el registre de casos de COVID-19 a Catalunya.

Aquestes dades provenen de diversos sistemes d'informació del Departament de Salut i del Servei Català de la Salut, detallats a <https://dadescovid.cat/documentacio> , i mostren per a cada dia, municipi, sexe i procediment diagnòstic el nombre de casos identificats com a positius COVID mitjançant alguna prova diagnòstica o per estudi epidemiològic. Tots els casos s’activen pels serveis de vigilància epidemiològica.

La dades s’actualitzen diàriament. Només es mostren les dades fins a 3 dies abans.

1. **Definir un títol pel dataset**.

Dades\_COVID\_Catalunya

1. **Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

La recollida de dades s’inicia l’01/03/2020.

**TipusCasData:** Data de detecció del cas, format dd/mm/aaaa. La data del cas és la data d’inici de símptomes, no la data de realització de la prova diagnòstica.

**ComarcaCodi:** Codi de la comarca.

**ComarcaDescripcio:** Nom de la comarca.

**MunicipiCodi:** Codi del municipi.

**MunicipiDescripcio:** Nom del municipi. En els casos en què no ha estat possible identificar el municipi de residència de la persona identificada com a cas positiu, el valor de la variable 'MunicipiDescripcio' és 'No classificat'. I en els casos de persones residents en municipis amb una població inferior a 200 habitants, per evitar la seva identificació i garantir-ne la confidencialitat, en aquesta variable s'hi fa constar el valor 'Altres municipis'.

**DistricteCodi:** variable per millorar el detall territorial de la informació facilitada per aquest conjunt de dades per a la població resident en el municipi de Barcelona. La identificació del districte s'obté a partir del sector sanitari, que per al municipi de Barcelona coincideix amb el districte municipal, i la codificació dels districtes que es mostra és la pròpia de l'Ajuntament de Barcelona. Per als registres corresponents a altres municipis, el camp 'DISTRICTECODI' està en blanc.

**DistricteDescripcio:** Nom del districte. Per als registres corresponents al municipi de Barcelona, aquests camps mostren els valors corresponents al districte municipal; per als registres corresponents a altres municipis conté el valor “No classificat”.

**SexeCodi:** 0/1

**SexeDescripcio:** Home/Dona o no classificat.

**TipusCasDescripcio:** especifica el procediment diagnòstic.

**NumCasos:**

1. **Representació gràfica.** Presentar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.
2. **Contingut**. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.
3. **Agraïments**. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.
4. **Inspiració**. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l’apartat 6..
5. **Llicència**. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

* Released Under CC0: Public Domain License
* Released Under CC BY-NC-SA 4.0 License
* Released Under CC BY-SA 4.0 License
* Database released under Open Database License, individual contents under Database Contents License
* Other (specified above)
* Unknown License

1. **Codi**. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
2. **Dataset**. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

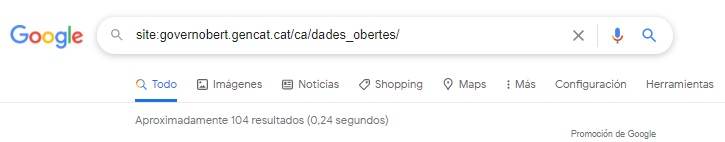
|  |  |
| --- | --- |
| Contribucions | Signa |
| Recerca prèvia | Maria Begoña Felip, Vicenç Pio |
| Redacció de les respostes | Maria Begoña Felip, Vicenç Pio |
| Desenvolupament codi | Maria Begoña Felip, Vicenç Pio |

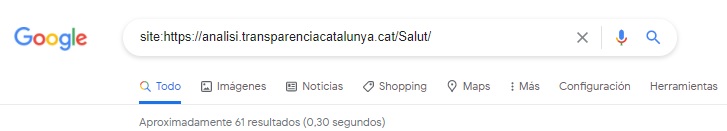
# **Avaluació inicial**

* Arxiu robots.txt
* Mapa del lloc web

He generat el fitxer sitemap.xml amb això: https://www.xml-sitemaps.com/

* La seva grandària





* La tecnologia emprada
* El propietari del lloc web