Josep Queralt
Aïda Piñol Noguero

PRÀCTICA 1 – web scraping

1. Context:

Adreça del lloc web:
https://meetingorganizer.copernicus.org/EGU23/sessionprogramme/pg-selection

L'enllaç que hem triat pertany a la **pàgina web de l'Assemblea General anual de la EGU**, l'esdeveniment més gran i important en el camp de les geociències celebrat a Viena, del 23 al 28 d'Abril. A mode de contextualització, aquest compta cada any amb més de 16000 presentacions orals, pòsters i Pics (presentacions interactives que combinen els avantatges de les orals i els pòsters), els quals abasten totes les disciplines de les ciències de la Terra, planetàries i espacials.
En aquest sentit, el cas pràctic que ens ocupa seria el següent:

"*Un laboratori d'anàlisis isotòpic envia una treballadora a l'assemblea per tal de veure les noves tendències en el camp d'investigació i recerca de les geociències i per identificar potencials usuaris dels serveis del laboratori.*
La treballadora, amb l'objectiu d'optimitzar la seva assistència, obre la pàgina web per analitzar el programa i per buscar quines sessions de cada dia són les més importants. La sorpresa que troba és que la pàgina no és molt intuïtiva i la quantitat d'opcions és aclaparadora. Ens demana si la podem ajudar a obtindre un dataset amb les dades més rellevants".

2. Títol Dataset: "sessions_EGU2023"

Escollim aquest títol general ja que engloba tant les exposicions, com els pòsters i altres formats de presentació dels papers i treballs dels investigadors/es.

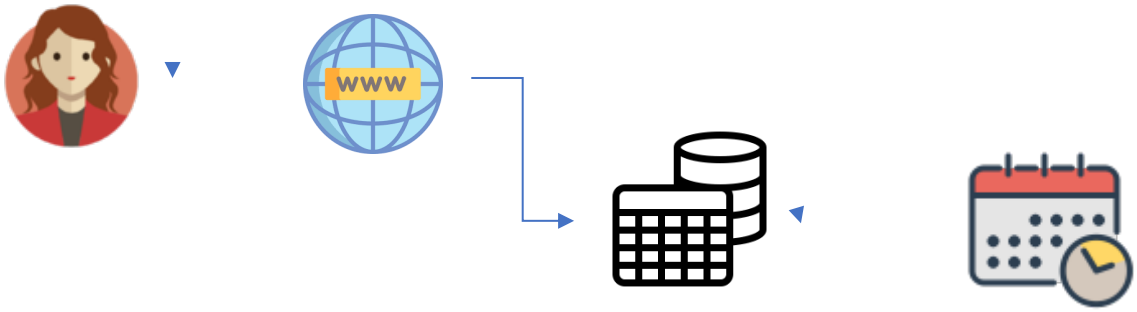
3. Descripció del dataset:

En termes generals, el dataset compta amb 36877 registres i 17 columnes ("Topic", "Informació", "Títol", "Autors" -i 13 columnes amb característiques-), les quals detallarem posteriorment a l'apartat de contingut.

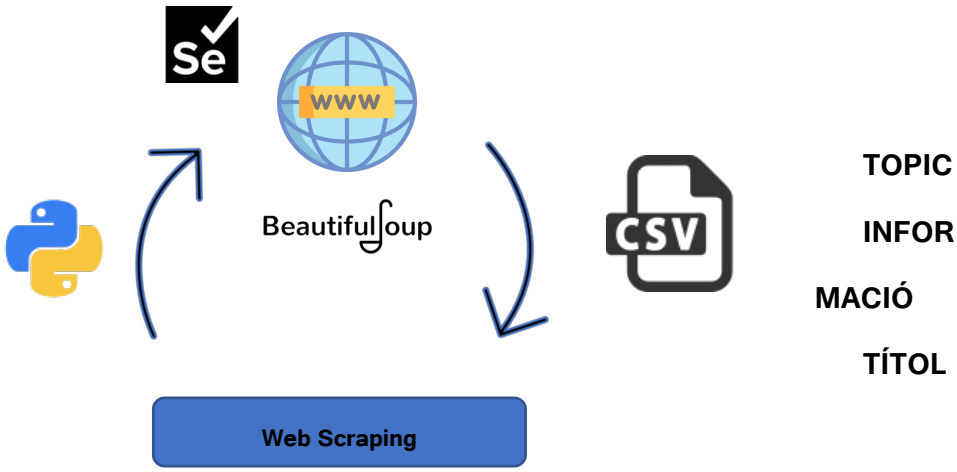
4. Representació gràfica:

A continuació, us presentarem una representació visual per sintetitzar el projecte escollit i el dataset obtingut:

PROBLEMA:



PROCÉS i DATASET:



5. Contingut:

El Dataset que hem obtingut conté informació referent als aspectes rellevants que ens demanava la treballadora, dividida en 4 columnes clau:

- **Topic:** type char
  Informació referent a la classificació general en la qual s'ha inserit el projecte d'investigació.
  Exemple: "*Artificial Intelligence for Natural Hazard and Disaster Management*", "*Compound weather and climate events*", etc.

- **Informació**: type char
  Especifica diferents detalls associats a cada projecte. En aquest cas, ens detalla el *tipus* de presentació (Orals/Posters on site), el *dia i franja horària* (ex: Tue, 25 Apr, 08:30–12:10 (CEST), 14:00–15:35 (CEST)) i la *sala* on es durà a terme (ex: Room 1.34).

- **Títol**: type char
  Títol del projecte/ paper.

- **Autors**: type char
  Nom dels autors/es que han realitzat el projecte/paper.

Altrament, comptem amb 13 columnes extra que hem definit com *"Característica1, Característica2..."* que ens han servit per tractar el problema de que les etiquetes div tenien el mateix nom i la informació no estava sempre al mateix lloc. En aquest cas, en la fase de neteja posterior, aquestes seran eliminades.

6. Propietari:

El propietari és copernicus.org (Copernicus GmbH) i la data d'actualització 2023-03-22T11:34:46Z [1].

Principis ètics i legals:

En primer lloc, es va revisar el conjunt de termes i condicions d'ús [2] i la informació de la pàgina web, actualitzada el 14 de Març de 2023.

Més concretament, vam trobar la informació següent:

> *"Teniu dret a que les dades que processem en funció del vostre consentiment o en compliment d'un contracte us l'entreguin automàticament o a un tercer en un format estàndard llegible per màquina."*
>
> *"Al públic en general, les dades personals dels usuaris de Copernicus Office només es mostren quan ocupen llocs concrets en conferències organitzades o revistes publicades per Copernicus. Els editors de les revistes es mostren automàticament al lloc web del consell editorial de les revistes i els convocants de les sessions de les conferències es mostren al programa en línia de la conferència respectiva per tal de contactar-hi. D'aquesta manera, els usuaris només es mostren al públic de manera predeterminada amb el nom, l'afiliació i el país. La visibilitat de qualsevol altra dada personal està controlada per l'usuari i la visibilitat s'ha de permetre de manera autònoma i explícita."*

[1] Enllaç: https://www.dondominio.com/es/whois/ consultat el 20/04/2023.
[2] Enllaç: https://www.copernicus.org/data_protection.html consultat el 21/04/2023.

A més a més, no només es va tenir en consideració aquest principi de verificació de les condicions d'ús, sinó que també vam respectar els principis de rastrejar només la informació pública exposada per la web (informació específica del programa en línia de la conferència), no vam causar cap dany al no sobrecarregar cap servidor, i sobretot, vam buscar fer un ús just i responsable de la informació obtinguda.

Altrament, podríem considerar com a anàlisis similars la pròpia opció que ofereix la pàgina web, ja que existeix la possibilitat de seleccionar sessions específiques i franges horàries per crear un document PDF i que l'usuari el descarregui amb la informació seleccionada.

En definitiva, hem pogut fer l'extracció sota els paràmetres ètics i legals, ja que com a usuaris de la web teníem aquest marge d'actuació.

7. Inspiració:

A grans trets, el dataset "sessions_EGU2023" pot ser interessant per a tots els assistents, estudiants, investigadors i empreses públiques i/o privades relacionades amb l'àmbit de les geociències. Més concretament, els motius (i algunes preguntes d'exemple) podrien ser:

a) Planificació i gestió del temps: són moltes les sessions simultànies que es duen a terme en els 6 dies que dura el congrés. És per això que el dataset pot actuar com una eina d'optimització del temps, en el sentit que ens permet identificar les sessions per franja horària i sala en la que es celebrarà, facilitant l'elecció de la nostra assistència.

- Quines sessions es fan a la sala "*Room 1.34*" el Dilluns 24 per la tarda?
- En quina sala es celebraran més sessions del tema "*Precipitation and urban hydrology* " el Dimarts 25?

b) Organització: el dataset brinda informació que els membre de l'organització poden emprar per preveure on hi haurà més assistents, o de cara a l'organització de l'any vinent, per destinar unes sales amb unes característiques o unes altres i les instal·lacions necessàries en funció del tipus de presentació, i els temes més repetits o rellevants.

- Quin format de presentació és el més utilitzat?
- En quina franja horària es celebren més sessions?

c) Interès acadèmic: el dataset especifica l'àmbit general i el títol del projecte. Així doncs, podem assistir a les presentacions que més ens interessin per tipologia, àmbit de recerca o tema.

- Quines sessions hi ha sobre el tòpic "*Physics-based earthquake modeling and engineering*"?
- Quants posters aborden el tema del "*Monsoon onset*"?

d) Interès laboral: al detallar el nom del projecte i els investigadors/es d'aquest, es poden identificar usuaris o clients potencials. En el cas que ens ocupa, la treballadora que ens ha demanat el dataset, pot identificar, per exemple, els projectes que han utilitzat anàlisis isotòpics però que no han treballat amb la seva empresa. Així ella es podrà posar en contacte amb ells de cara a presentar els serveis que ofereix la seva companyia i les seves condicions. Altrament, un estudiant o investigador pot assistir a una sessió concreta per veure com treballen o com han abordat un tema en concret altres investigadors.

- Quins autors parlen del tema "*Stone Heritage and Geological Heritage Sites*"?
- Quins autors han utilitzat anàlisis d'isòtops estables en la seva investigació?

e) Anàlisis de tendències: amb el dataset es poden identificar els temes més abordats durant el congrés i també les noves eines i metodologies més utilitzades per la comunitat científica.

- Quants articles utilitzen solucions d'IA?
- Quins són els temes més abordats en l'àmbit del "*Modelling and Monitoring Complex Urban Systems*"?

f) Networking: amb el dataset ens pot resultar més fàcil identificar una persona o grup de recerca per ficar-nos en contacte amb aquests, ja que tenim informació dels temes que tracten, el nom dels autors i el lloc i hora de la presentació.

- En quines franges horàries presenta el grup de recerca de Claudia Rodríguez-Pérez, Nemesio M. Pérez, Fátima Rodríguez i Carmen Solana?

- En quina sala i quan es presentarà el projecte " Using machine learning to emulate the hydrodynamic model for flood inundation modelling"?

8. Llicència:

Pel dataset resultant s'ha seleccionat la llicència Creative Commons **"Atribució-NoComercial 4.0 Internacional (CC BY-NC 4.0)[3]"**. En aquest sentit, els usuaris del dataset seran lliures de compartir-lo (copiar i redistribuir) i adaptar-lo (transformar, modificar), però no en podran fer un ús comercial i hauran de donar llicència als creadors.

[3] Enllaç: https://creativecommons.org/licenses/by-nc/4.0/deed.es consultat el 21/04/2023.

9. Codi:

- Aspectes tècnics:

Llenguatge de programació: Python

Llibreries i versions utilitzades:

Utilitzem la comanda `pip3 freeze > requirements.txt` (resultat als Annexos)

- Procés i dificultats:

Les dificultats més destacables han derivat de la pròpia estructura de la pàgina web, en el sentit que la pàgina principal estava organitzada en 4 apartats, amb 39 "form-check" (formularis). Per obtenir un enllaç de la sessió, primer s'ha d'accedir a un d'aquests "form-checks" (fent ús de la llibreria Selenium). En funció de quin escollim, ens pot sortir o no una llista desordenada (<ul>) d'items de tipus <li> HTML.

Ara bé, el que vam inspeccionar va ser el codi HTML de la pàgina web, identificant els elements ancla <a>. En aquest sentit, ens adonem que la informació que necessitem referent a les presentacions, es troba dintre de cada sessió. Aquestes sessions les identifiquem amb . Cada sessió té un atribut amb un enllaç únic que ens condueix a la informació que volem al dataset (que a la pàgina web no és visible).

Quan accedim a aquest darrer enllaç, ens trobem que els detalls no sempre es troben organitzats igual ni en la mateixa posició. Tampoc sabem quins són més rellevants, així que hem optat per recollir-los tots, creant les columnes Característica1, Característica2.. per no perdre'n cap.

10. Dataset:

Zenodo: https://zenodo.org/record/7857166#.ZEVqU-xBxjM

DOI: https://doi.org/10.5281/zenodo.7857099

GitHub: https://github.com/josquefi/PRA1

11. Vídeo:
https://drive.google.com/file/d/1AkwHQomgjEqOFq7Wrh67qU5k4W8HAepg/view?usp=share_link

Contribucions Signatura

Investigació prèvia Aïda Piñol Noguero, Josep Queralt

Redacció de les respostes Aïda Piñol Noguero, Josep Queralt

Desenvolupament del codi Aïda Piñol Noguero, Josep Queralt

Participació al vídeo Aïda Piñol Noguero, Josep Queralt

Enllaç vídeo:

12. Annexos:

anyio==3.6.2

argon2-cffi==21.3.0

argon2-cffi-bindings==21.2.0

arrow==1.2.3

asttokens==2.2.1

async-generator==1.10

attrs==23.1.0

backcall==0.2.0

beautifulsoup4==4.12.2

bleach==6.0.0

certifi==2022.12.7

cffi==1.15.1

charset-normalizer==3.1.0

colorama==0.4.6

comm==0.1.3

debugpy==1.6.7

decorator==5.1.1

defusedxml==0.7.1

exceptiongroup==1.1.1

executing==1.2.0

fastjsonschema==2.16.3

fqdn==1.5.1

h11==0.14.0

idna==3.4

ipykernel==6.22.0

ipython==8.12.0

ipython-genutils==0.2.0

ipywidgets==8.0.6

isoduration==20.11.0

jedi==0.18.2

Jinja2==3.1.2

jsonpointer==2.3

jsonschema==4.17.3

jupyter==1.0.0

jupyter-console==6.6.3

jupyter-events==0.6.3

jupyter_client==8.2.0

jupyter_core==5.3.0

jupyter_server==2.5.0

jupyter_server_terminals==0.4.4

jupyterlab-pygments==0.2.2

jupyterlab-widgets==3.0.7

MarkupSafe==2.1.2

matplotlib-inline==0.1.6

mistune==2.0.5

nbclassic==0.5.5

nbclient==0.7.3

nbconvert==7.3.1

nbformat==5.8.0

nest-asyncio==1.5.6

notebook==6.5.4

notebook_shim==0.2.2

numpy==1.24.2

outcome==1.2.0

packaging==23.1

pandas==2.0.0

pandocfilters==1.5.0

parso==0.8.3

pickleshare==0.7.5

platformdirs==3.2.0

prometheus-client==0.16.0

prompt-toolkit==3.0.38

psutil==5.9.5

pure-eval==0.2.2

pycparser==2.21

Pygments==2.15.1

pyrsistent==0.19.3

PySocks==1.7.1

python-dateutil==2.8.2

python-json-logger==2.0.7

pytz==2023.3

pywin32==306

pywinpty==2.0.10

PyYAML==6.0

pyzmq==25.0.2

qtconsole==5.4.2

QtPy==2.3.1

requests==2.28.2

rfc3339-validator==0.1.4

rfc3986-validator==0.1.1

schedule==1.2.0

selenium==4.8.3

Send2Trash==1.8.0

six==1.16.0

sniffio==1.3.0

sortedcontainers==2.4.0

soupsieve==2.4.1

stack-data==0.6.2

terminado==0.17.1

tinycss2==1.2.1

tornado==6.3

tqdm==4.65.0

traitlets==5.9.0

trio==0.22.0

trio-websocket==0.10.2

tzdata==2023.3

undetected-chromedriver==3.4.6

uri-template==1.2.0

urllib3==1.26.15

wcwidth==0.2.6

webcolors==1.13

webencodings==0.5.1

websocket-client==1.5.1

websockets==11.0.2

widgetsnbextension==4.0.7

wsproto==1.2.0