"Eina de creació de bases de dades de bibliografia mèdica específica"

Autors: "Xabier Urra Nuin i Roger Ribas Gimeno"

Data: "3 de novembre de 2021"

1. Context

Els treballs científics com ara escriure un article o elaborar una tesi doctoral tenen com a part prèvia indispensable fer una revisió exhaustiva de la literatura disponible sobre el tema estudiat. La revisió de la literatura s'ha simplificat en les darreres dècades per l'existència de webs d'accés públic on poder fer cerques bibliogràfiques. A l'àmbit de la medicina, el recurs més important de cerca d'informació és PubMed, una pàgina web gratuïta per a accedir al MEDLINE, una base de dades bibliogràfiques de citacions i resums d'articles de recerca en biomedicina i ciències de la vida.

2. Títol

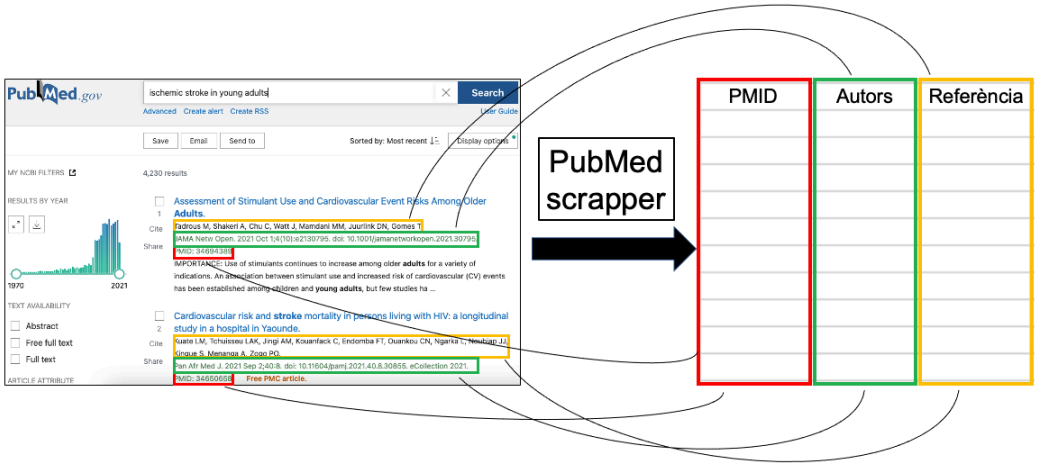
Eina de creació de bases de dades de bibliografia mèdica específica.

3. Descripció del dataset.

L'eina que hem desenvolupat permet a l'usuari introduir els termes desitjats per fer la cerca bibliogràfica (per això parlem al títol del "bibliografia mèdica específica"), i retorna com a resultat un llistat imprès i una base de dades en format "csv" que recull els articles científics que contenen els termes de la cerca, en concret tres camps que permeten la seva identificació i divulgació, com són: el PMID (identificador únic), llistat d'autors i referència bibliogràfica (per poder citar l'article, entre altres funcions).

4. Representació gràfica

En aquest esquema mostrem un exemple de cerca al web PubMed i l'output obtingut després d'executar l'eina "pubmed_scraper", amb els diferents camps d'informació codificats en colors diferents:



5. Contingut

La cerca que permet l'eina "pubmed_scraper" és personalitzable. En executar l'arxiu, s'obre una caixa de text en la que l'usuari especifica els termes de cerca, i l'eina retorna com a resultat el llistat d'articles científics que contenen aquests termes, a més d'exportar-los al mateix directori a una base de dades amb nom "pubmed.csv". Per a cada article, es recullen els camps següents:

- **PMID**: identificador únic de l'article a PubMed.
- **Autors**: llistat d'autors de l'article.
- **Referència**: referència bibliogràfica de l'article.

6. Agraïments

Les dades de l'aplicatiu s'obtenen a partir de PubMed, un recurs gratuït accessible a https://pubmed.ncbi.nlm.nih.gov que dóna suport a la cerca i recuperació de literatura biomèdica i de ciències de la vida. La base de dades PubMed conté més de 33 milions de cites i resums de literatura biomèdica.

S'ha d'agrair als organismes que han desenvolupat i mantenen PubMed: el Centre Nacional d'Informació Biotecnològica (NCBI), a la Biblioteca Nacional de Medicina dels Estats Units (NLM), situada en els Instituts Nacionals de Salut (NIH).

7. Inspiració

Aquesta eina permet la generació i exportació de cerques bibliogràfiques actualitzades de temes biomèdics fent servir un dels repositoris més grans de literatura biomèdica disponible mèdicament.

Encara que pugui haver-hi eines semblants en pàgines web com a PubMed o en aplicacions bibliogràfiques com Mendeley, la senzillesa d'aquesta eina la podria fer útil per la seva aplicació en petits grups de recerca que vulguin mantenir bases actualitzades de manera ràpida, amb uns pocs camps d'informació que permeten la identificació dels articles ja sigui pels noms dels autors o per disposar de l'identificador únic de cada article a PubMed.

8. Llicència

Encara que els articles trobats a la cerca a PubMed poden tenir diferents llicències, la llicència escollida pel joc de dades resultant ha estat Released Under CC0: Public Domain License.

El motiu és que les dades s'ajusten a les característiques d'aquesta llicència. La CC0 permet als usuaris i propietaris de continguts com a bases de dades renunciar a drets d'autor i posar-les de la forma més completa possible en el domini públic, de manera que uns altres puguin basar-se en elles, millorar-les i reutilitzar-les lliurement per a qualsevol propòsit sense restriccions. Això permetria a més de fer servir la base de dades resultant de la cerca per revisar la literatura disponible, que els usuaris poguessin actualitzar o ampliar la cerca o fusionar-la amb els resultats d'altres cerques sense cap limitació.

9. Codi

El codi l'hem preparat en Python, i està disponible a un repositori Git: https://github.com/xabimaster/pubmed_scraper

10. Dataset

Hem publicat el dataset obtingut amb la cerca "COVID-19 + ischemic stroke" en format CSV a Zenodo: https://zenodo.org/record/5639149#.YYFYEtaZO60

Enllaç del DOI: 10.5281/zenodo.5639149T

11. Taula de contribucions

Contribucions	**Signatura**
Investigació prèvia	XUN, RRG
Redacció de les respostes	XUN
Desenvolupament del codi	RRG

12. Recursos

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2010). El lenguaje Python. Editorial UOC.
- Tutorial de Github https://guides.github.com/activities/hello-world.