Pràctica 1: Web scraping

1. **Context.** Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.

   Per a fer el web scraping s'ha escollit el lloc web www.airport-london-heathrow.com, que proporciona informació relacionada amb l'aeroport de Londres-Heathrow. En l'apartat "arrivals" (/lhr-arrivals), es pot consultar l'estat dels vols que hi aterraran, per al mateix dia, el dia següent o el dia anterior en franges de 6 hores. El web citat mostra aquesta informació per a que es pugui consultar l'estat dels vols a temps real, útil per a persones que esperen l'arribada d'algú que aterra a Heathrow.

2. **Títol.** Definir un títol que sigui descriptiu pel dataset.

   El títol del dataset és London-Heathrow_arrivals_DATAHORA, on DATAHORA indica el dia i hora en què s'han extret les dades.

3. **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 escollit.

   El dataset mostra les dades relatives a l'estat de cada vol que està programat per aterrar a l'aeroport de Londres-Heathrow, de la mateixa manera que es mostra en els panells d'informació als aeroports. Aquestes dades inclouen el número de vol, origen, hora programada i estat del vol.

4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

5. **Contingut.** Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

Els camps que inclou el dataset són els següents:

- ID: Índex del vol dins del dataset
- Origin: nom de l'aeroport d'origen
- Origin ID: Codi (3 lletres) de l'aeroport d'origen
- Arrival: hora d'arribada programada
- Flights: codi(s) del vol (múltiples en cas de codi compartit)
- Airline: aerolínia/es que opera el vol
- Terminal: Número de terminal on aterra l'avió
- Status: Estat del vol
- Date: Data programada

Aquesta informació es recull per a un dia sencer, tot i que és possible també extreure les dades del dia anterior i següent al dia de la consulta. Les consultes es fan per franges de 6 hores, però els resultats són concatenables per tal de poder analitzar períodes de temps més llargs.

La informació s'ha recollit mitjançant tècniques de web scraping, amb les llibreries de Python pandas, xmlx, datetime, re, requests i BeautifulSoup. El codi es pot consultar mitjançant l'enllaç de l'apartat 10.

6. **Agraïments.** Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

   Tot i que el propietari de la pàgina que s'està utilitzant per a fer el web scraping és l'aeroport de Londres-Heathrow, les dades d'operativa de les aeronaus són de difusió oberta, ja que es transmeten mitjançant l'ACARS (Aircraft Communications Addressing and Reporting System).

   El tema de data ownership al món de l'aviació és un tema a l'ordre del dia, ja que els fabricants d'aeronaus i els seus components estan començant a encriptar les dades per controlar la seva difusió [1]. Altres notícies inclouen el d'un jove difonent la ubicació del jet privat d'Elon Musk mitjançant un bot de Twitter, a través de les dades del seu ACARS, que per regulacions de la FAA (Federal Aviation Administration) emet en obert [2].

   Existeixen altres pàgines que mostren informació similar, entre les quals:

   - https://www.heathrow.com/arrivals
   - https://www.flightradar24.com/data/airports/lhr
   - https://www.flightstats.com/v2/flight-tracker/arrivals/LHR
   - https://aerfortel.com/flight-arrivals-heathrow-airport/
   - https://www.skyscanner.net/flights/arrivals-departures/lhr/london-heathrow-arrivals-departures

   Totes mostren aproximadament la mateixa informació, tot i que algunes de forma incompleta, al no mostrar les operacions en codi compartit, o en formats menys còmodes per a l'usuari.

7. **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.

   Aquest conjunt de dades pot ser interessant des de dues perspectives:

   - En temps real, pot integrar-se amb altres plataformes i notificar en temps reals d'actualitzacions en l'estat de vol per a usuaris finals que estan interessats perquè esperen algú que aterra a l'aeroport de Heathrow.
   - Si s'extreuen les dades periòdicament, l'anàlisi de dades històriques pot ser rellevant per determinar el nivell de servei de les aerolínies que operen a Heathrow. Això pot ser útil tant a les pròpies companyies, per fer avaluació del rendiment propi o per conèixer la seva situació respecte de la competència, com per el propi aeroport, per a optimitzar les seves operacions i avaluar l'assignació de slots actual.

Els exemples mencionats en l'apartat 6 estan enfocats principalment en el primer cas: presenten la informació per a que el consumidor final consulti a temps real.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:

La llicència del dataset resultant seria CC BY-SA 4.0, ja que, com s'ha comentat en l'apartat 6, les dades són públiques i no hi ha impediments per a treure rendiment econòmic a partir d'elles (de fet, si no fos així es perdria la utilitat d'aquestes dades per a les empreses). Sí que és necessari citar els autors i no modificar les dades per assegurar-ne la integritat.

9. **Codi.** Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi es pot consultar en el següent repositori de github:
https://github.com/Camporrio/London-Heathrow-Airport-WebScraping

10. **Dataset.** Publicar el dataset obtingut*(*)* en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

El dataset s'ha publicat a Zenodo, amb el següent DOI: 10.5281/zenodo.6453166

11. **Vídeo**. S'ha de lliurar un vídeo explicatiu de la pràctica on cadascun dels integrants del grup expliqui amb les seves pròpies paraules tant les respostes del projecte com el codi utilitzat per a dur a terme l'extracció. El vídeo ha de ser enviat a través d'un enllaç a Google Drive que heu de proporcionar, juntament amb l'enllaç al repositori Git, al moment de lliurar la pràctica.

El vídeo es pot veure a través del següent enllaç:
https://drive.google.com/file/d/1dbHgYM_1U8iyHrd76T3gh3sQPKb1aXte

Referències

[1] IATA Data ownership panel [en línia]. Disponible a: https://www.iata.org/contentassets/81005748740046de878439e6c54f2355/d2-1130-1230-data-ownership-panel-discussion.pdf

[2] The New York Times. A Teenager Tracked Elon Musk's Jet on Twitter. Then Came the Direct Message [en línia]. Disponible a: https://www.nytimes.com/2022/02/03/technology/elon-musk-jet-tracking.html

ANNEX

Contribucions	Signatura
Investigació prèvia	GCF, AYC
Redacció de les respostes	GCF, AYC
Desenvolupament del codi	GCF, AYC