Pràctica 1: Web scraping

Context

Aquest web scraper emmagatzema en un arxiu csv les informacions relatives a totes les operacions, és a dir, tant arribades com sortides, ocorregudes a l'aeroport de Barcelona – El Prat el dia anterior. El web triat és https://www.barcelona-airport.com, que tot i no ser la pàgina oficial de l'aeroport ni d'AENA disposa de tota la informació que desitgem emmagatzemar.

Aquesta activitat ha estat realitzada de manera individual per Guillem Tudela Debrigode.

Definició del títol del dataset

El dataset creat sempre inclourà la data del dia de les operacions en el seu títol per a facilitar-ne una posterior diferenciació en cas de tenir datasets de diferents dies. L'script, un cop finalitzat, crea un arxiu csv seguint el següent format: BCN_Operations_AAAA-MM-DD.

Descripció del dataset

Com s'ha comentat anteriorment, el dataset inclou la informació sobre totes les operacions que es van dur a terme el dia anterior a l'aeroport de Barcelona – El Prat. Això implica enregistrar la companyia aèria, el número de vol, a quin aeroport opera així com informació relativa a l'hora programada i l'hora real en que es va iniciar l'operació. Permetent així en un posterior anàlisi el càlcul dels retards.

Representació gràfica

Vista de l'aeroport de Barcelona – El Prat
(font: https://ca.wikipedia.org/wiki/Aeroport_de_Josep_Tarradellas_Barcelona_%E2%80%93_el_Prat)

Contingut

El contingut del dataset inclou la següent informació:

- Scheduled Date: Data del vol programada.
- City: Nom de la ciutat d'origen o de destinació.
- IATA: Nom de l'aeroport d'origen o destinació en codi IATA, és un codi de tres lletres associat a cada aeroport.
- Airline: Nom de la companyia que opera la ruta.
- Flight: Número de vol de la ruta. Les dues primeres lletres estan associades a l'aerolínia (p.ex. IB per Ibèria, o VY per Vuèling) i els números posteriors a la ruta (aeroport d'origen, de destinació i horari). Una ruta operada un altre dia conserva el número de vol.
- Operation: Tipus d'operació, és a dir, si és sortida o arribada.
- Scheduled Time: Hora programada de l'operació.
- Terminal: Terminal de l'aeroport on es realitza l'operació, pot ser 1 o 2.
- Status: Status de l'operació, aterrat, en ruta, cancel·lat...

- Real Date: Data real de l'operació (si l'avió aterra abans d'hora o surt amb molt de retard pot ser que no coincideixi el dia programat amb el real).
- Real Time: Hora real de l'operació. La diferència entre aquesta columna i la "Scheduled Time" permeten calcular el retard del vol.
- Gate: Porta de la terminal des d'on s'ha embarcat (informació només disponible per a les operacions de sortides i en cas que l'avió hagi aterrat).

Agraïments

Les dades han estat extretes del web https://www.barcelona-airport.com mitjançant un script en llenguatge Python.

Les tècniques de Web Scraping aplicades s'han basat en el contingut dels següents llibres:
- **Mitchell, Ryan** (2018). *Web Scraping with Python, 2nd Edition*. O'Reilly Media, Inc.
- **Subirats, L., Calvo, M.** (2018). *Web Scraping*. Editorial UOC.
- **Heydt, Michael** (2018). *Python Web Scraping Cookbook*. Packt Publishing

Inspiració

Aquest dataset pot ser molt útil per a determinar dia rere dia indicadors de qualitat (o KPI - key performance indicator) tals com el percentatge de vols puntuals o els minuts de retard mitjà per cada operació. També ens permet veure quines són les companyies que més vols operen, les rutes més freqüents o quines hores l'aeroport opera al màxim de la seva capacitat. Executant l'script diàriament durant un període de temps prolongat ens permet determinar si els retards en una determinada ruta són puntuals o sistemàtics, també ens permet determinar la temporalitat de les rutes o en quines hores l'aeroport experimenta de mitjana més retards.

Aquesta informació pot ser interessant per a diferents tipus de persones:
- A nivell d'usuari: podem arribar a projectar la probabilitat que surti el nostre vol amb retard i de quants minuts; aquesta dada és especialment útil en cas de tenir algun vol de connexió en el mateix aeroport de Barcelona.
- A nivell de companyia aèria: permet analitzar la puntualitat de la competència i les rutes que opera.
- A nivell aeroport: ens permet establir indicadors de qualitat i fer-ne un seguiment. També ens pot interessar recrear el flux de moviment dels passatgers a partir de la informació emmagatzemada de les portes, determinant si hi ha zones poc usades o aquelles d'alta ocupació i optimitzar l'assignació de portes per cada vol.

Llicència

El motiu d'aquesta tria és que d'aquesta manera es po

Codi i dataset

El codi (en .py) com el dataset (en .csv) es troben ubicats al següent link:
https://github.com/GuillemTD/BCN_Airport_scraping

Nota final

Degut a l'arquitectura de la web triada, per a obtenir la informació relativa a la data i hora real de les operacions, així com la de la porta des d'on s'embarca, cal fer un request per cadascun dels vols operats usant el codi de vol, això comporta que per a obtenir el dataset complet es pot requerir més de 25 minuts (dependrà del nombre d'operacions que hagi experimentat l'aeroport). Per aquest motiu s'ha creat un segon script, anomenat BCN_Airport_scraping_limited.py que prescindeix de buscar aquesta informació i, en

conseqüència, el seu procés de creació es redueix a qüestions de segons. El nom del dataset, seguirà el format BCN_Operations_AAAA-MM-DD_limited.