Universitat Oberta de Catalunya

uoc.edu

Tipologia i cicle de vida de les dades

Pràctica 1: Web Scraping

Albert Torres Rubio

M2.951

14/04/2020

Estudiant: Albert Torres Rubio

Índex

1. Context

El lloc web https://airfleets.net recopila informació sobre les aeronaus que tenen i han tingut les diverses aerolínies comercials que existeixen. La web, per exemple, mostra el seu estat, a quina aerolínia pertany, si aquesta ha estat transferida, entre d'altres característiques. La web també conté un apartat d'accidents aeris.

El context de recol·lecció de la informació és purament acadèmic i analític. Potser en un futur es podria utilitzar amb fins d'anàlisi de mercats/competència per diverses empreses (amb previa consulta amb l'autor i canviant la llicència escollida per distribuir el dataset). La idea era recopilar les dades de les diferents aeronaus i veure l'estat d'aquestes, la seva evolució entre les diferents companyies, poder extreure la mitjana d'edat del tipus concret d'aeronau o la mitjana d'edat de les aeronaus per aerolínia, saber quina es la companyia que "reutilitza" més les aeronaus, etc.

2. Títol del dataset

El scraper implementat te la capacitat de generar diversos datasets segons l'elecció del usuari (a l'apartat 9 s'explicarà el seu funcionament). L'usuari pot escollir, si vol, de quins tipus d'aeronau vol obtenir les dades, segons els paràmetres que introdueixi, així dons el títol del dataset esdevindria:

- Per un únic tipus d'aeronau:

{tipus d'aeronau} status.csv

- Per diversos tipus d'aeronau:

{tipus d'aeronau1}-{tipus d'aeronau2}-{...} status.csv

En canvi, si no s'escull cap tipus d'aeronau el scraper retorna totes les dades relatives a tots els tipus d'aeronau que conté el lloc web (https://airfleets.net). En aquest cas el títol esdevindria:

civil airfleets status.csv

2.1 Dataset lliurat

El dataset lliurat té com a títol *civil airfleets estatus.csv*. Però cal remarcar que NO conté tots el tipus d'aeronaus disponibles al lloc web https://airfleets.net. Això es degut a que per extreure totes les dades de totes aeronaus del lloc web hagués calgut mantenir el web scraper funcionant durant un període molt llarg de tems. Degut a la limitació de recursos i la potencia limitada del meu ordinador personal això no ha estat possible.

Si més no, a la carpeta *data* del projecte es poden veure els diferents datasets extrets pels diferents tipus d'aeronaus. També es pot veure l'arxiu *civil airfleets estatus.csv* recopilant totes les dades d'aquests datasets. Els tipus d'aeronaus de les que s'han extret les dades són:

- Airbus A220
- Airbus A300
- Airbus A310
- Airbus A318
- Airbus A319
- Airbus A320
- Airbus A330
- Airbus A350
- Airbus A380
- Airbus A321
- Boeing 717
- Boeing 737
- Boeing 747
- Boeing 757
- Boeing 777
- Boeing 787
- Comac ARJ21
- Comac C919
- Concorde
- Lockheed L-1011 TriStar
- McDonnell Douglas DC-10
- McDonnell Douglas MD-11
- Sukhoi SuperJet 100

Com a punt a destacar, per extreure el total de dades obtingudes, el web scraper ha estat funcionant al voltant de 2 dies ininterrompudament.

3. Descripció del dataset

Com s'ha comentat prèviament, el dataset general o els diferents datasets dels diferents tipus d'aeronaus intenten recollir la informació disponible de les diferents aeronaus mostrades al lloc web https://airfleets.net, concretament al apartat de aircraft > Supported planes (https://airfleets.net/recherche/supported-plane.htm). Aquest apartat recull tots els tipus d'aeronaus que conté el lloc web, amb els seus respectius links a les diferents aeronaus.

El dataset conté informació sobre les aeronaus civils dels principals fabricants d'aquestes, com per exemple, Boeing, Bombardier, Airbus... Les dades que conté són, per exemple, l'estatus, moviments entre operadors, registre, data del primer vol, motors...

4. Imatge representativa

Diferent tipus d'aeronaus pertanyents a diferents aerolínies.

Figura 1 Font: https://airlinersgallery.wordpress.com/

5. Contingut

Les dades recollides pel mètode del web scraping són:

- **serialNumber:** Número de sèrie de l'aeronau.
- **lnNumber:** Número de línia de l'aeronau.
- **familyType:** Família a la qual pertany l'aeronau.
- **Type:** Tipus específic de l'aeronau.
- **firstFlight:** Data del primer vol de l'aeronau.
- **planeAge:** Edat en anys de l'aeronau.
    - Format: Decimal
- **testRegistration:** registre ICAO (International Civil Aviation Organization) de l'aeronau per les proves de test.
- **eginesType:** Tipus de motors.
- **enginesNumber:** Quantitat de motors de l'aeronau.

- **registration:** registre ICAO (International Civil Aviation Organization) de l'aeronau.
- **deliveryDate:** Data d'entrega de l'aeronau al opererador.
    - Format: dd/mm/aaaa
- **operator:** Aerolínea propietària de l'aeronau.
- **remark:** Comentari sobre l'aeronau.
- **status:** Estatus final de l'aeronau amb l'operador indicat.
    - Opcions possibles:
        - Active: Aeronau activa, actualment en ús.
        - Stored: Aeronau emmagatzemada.
        - Scrapped: Aeronau desmantellada.
        - Written off: L'aeronau ha sofert un accident irreparable.
        - Transferred: Aeronau transferida a un altre operador.
        - On order: Aeronau pendent de lliurament a l'operador.

6. Agraïments

Com s'ha comentat prèviament les dades pertanyen al lloc web https://airfleets.net. Aquest lloc web relaciona aerolínies amb aeronaus i informa dels seus traspassos. Descriu com estan composades les flotes de les aerolínies. Per altra banda, també conté un registre d'aeroports i d'accidents en el món de l'aviació.

El lloc web porta actiu des del 2002, però conté dades d'aeronaus fabricades des del segle passat.

7. Inspiració

La inspiració de la cerca d'aquestes dades ha vingut donada per dos motius, el primer, és que un dels exemples exposats a l'assignatura sobre aquesta pràctica estava relacionat amb els accidents aeris. El segon, es què jo sóc enginyer aeronàutic i durant l'estudi de la meva carrera i a la meva vida professional he cercat informació en el lloc web indicat moltes vegades. Així dons, he considerat que el lloc web en qüestió pogués ser una bona font d'informació.

Si més no, amb aquest conjunt de dades es podria intentar respondre a diferents preguntes, com per exemple, Quines són les aerolínies que transfereixen més les sever aeronaus? Quines han tingut més accidents? Quines aerolínies tenen més despeses segons l'adquisició de noves aeronaus? Quin tipus d'aeronaus pateix

més accidents? Quina es la mitjana de traspassos fins que una aeronau s'emmagatzema o es desmantella? Quin operador té la flota de més edat? Constitueix això un problema en quant a la probabilitat d'accident?, etc.

8. Llicència

Tant el dataset com el web scraper es distribueix amb llicència: **CC BY-NC-SA 4.0**. Aquesta llicència ha estat escollida per la bona adequació de les seves clàusules amb l'autor del projecte:

- Es permet la copia i la redistribució del material en qualsevol medi o format.
- Es permet la mescla, transformació i creació a partir del material publicat.
- Es obligatori reconèixer adequadament l'autoria del material, proporcionar un enllaç a la llicència i indicar els canvis realitzats sobre el material.
- No es permet utilitzar el material amb una finalitat comercial.
- Si és mescla, transforma o crea a partir del material, es deurà distribuir les noves contribucions sota la mateixa llicència que l'original.

9. Codi

Estructura de carpetes:

- code: Ubicació del codi a executar.
- data: Emmagatzematge dels datasets que crea el scraper.
- log: Registre d'incidències durant l'execució del scraper.
- report: Ubicació d'aquest document.

Cal executar el arxiu *main.py* per iniciar el web scraper. Hi ha diverses opcions per executar el codi, si executem el codi:

python3 main.py -h

Podrem veure les opcions que hi ha

Web scraper for airfleets.net supported aircraft

optional arguments:
-h, --help show this help message and exit
-a [...], --aircraft [...] Aircraft type to scrape (Ex.: "Comac C919, Boeing 737") / All types selected if no argument

-q, --quiet Selenium performs silently

- Si executem amb la opció *-a* podem afegir una llista amb el tipus d'avió del que volem extreure les dades. Si no indiquem el paràmetre el web escraper

extreurà les dades de tots els avions suportats a la pàgina https://airfleets.net. Exemple de com introduir els diferents tipus d'aeronaus:

-a "Airbus A220, Concorde, Boeing 717"

Llista de tipus d'avions suportada:

- Airbus A220
- Airbus A300
- Airbus A310
- Airbus A318
- Airbus A319
- Airbus A320
- Airbus A321
- Airbus A330
- Airbus A340
- Airbus A350
- Airbus A380
- ATR 42/72
- BAe 146 / Avro RJ
- Beech 1900D
- Boeing 717
- Boeing 737
- Boeing 737 NG / Max
- Boeing 747
- Boeing 757
- Boeing 767
- Boeing 777
- Boeing 787
- Canadair Regional Jet
- Comac ARJ21
- Comac C919
- Concorde
- Dash 8
- Embraer 120 Brasilia
- Embraer 135/145
- Embraer 170/175
- Embraer 190/195
- Fokker 50
- Fokker 70/100
- Iliouchine Il-96
- Lockheed L-1011 TriStar
- McDonnell Douglas DC-10
- McDonnell Douglas MD-11
- McDonnell Douglas MD-80/90
- Saab 2000
- Saab 340
- Sukhoi SuperJet 100

Com a punt afegir, no està recomanat l'ús del web scraper sense indicar una llista de tipus d'aeronaus, ja que executar el codi per tots els tipus d'aeronaus pot demorar dies.

- Si executem amb l'opció *-q* el navegador controlat per la llibreria Selenium actuarà en segon pla. En cas contrari, es podran veure per pantalla totes les acciones del navegador.

10. Publicacions a Zenodo

- DOI dataset: https://doi.org/10.5281/zenodo.3752107
- DOI software: https://doi.org/10.5281/zenodo.3752082

11. Link de Github

https://github.com/albert-torres/airfleet-scraper