SI100B Project Report - Crawler Week 1

Workload Division

- Team members:

  李炳楠 libn@shanghaitech.edu.cn

  秦一帆 qinyf1@shanghaitech.edu.cn

  陈溯汀 chenst@shanghaitech.edu.cn

- Workload division:

  李炳楠：Decide what each part of the URL means; Figure out the basis of web crawler; Store data in a csv file; Consult online and offline resources; Ask for more accurate description on Piazza

  秦一帆：Decide what each part of the URL means; Add comments for code readability; Figure out the basis of web crawler; Figure out wireless connection of Pi; Clearify the process

  陈溯汀：Figure out the basis of web crawler; Decide what each part of the URL means; Consult online and offline resources; Store data in a csv file; Write the project report

Preliminary Comment

- Online resources:
  1. https://blog.csdn.net/qq_29011249/article/details/109154418
  2. https://blog.csdn.net/u012813201/article/details/70210812
  3. https://requests.readthedocs.io/zh_CN/latest/user/quickstart.html#id5
  4. https://blog.csdn.net/feit2417/article/details/85925586
- Offline resources:
  1. Eric Matthes (2016, July). Python Crash Course: A Hands-on, Project-Based Introduction to Programming. No Starch Press, p.312-337
- The main difficulty is to figure out what each part of the URL means. It has to be mention that valid information can be acquired only if 'air' is set to '1'.

Data Source

- Write down the URLs you requested for the flight data from FlightRadar24 or FlightAware and describe what each part of the URL means and the HTTP method used for each URL.

https://data-live.flightradar24.com/zones/fcgi/feed.js?bounds=32.6794,29.6794,120.09043,123.09043&faa=1&satellite=1&mlat=1&flarm=1&adsb=1&gnd=1&air=1&vehicles=0&estimated=1&maxage=14400&gliders=1&stats=0

bounds = max_latitude,min_latitude,min_longitude,max_longitude

faa=1&satellite=1&mlat=1&flarm=1&adsb=1 means   how the website gather the data from the airlines:

faa means data comes from Federal Aviation Administration

satellite means satellite is used

mlat means Multilateration is applied

flarm means 'flight alarm', an electronic device which is in use as a means of alerting pilots

adsb means Automatic Dependent Surveillance Broadcast, enables communication between flights

estimated=1 means displaying flights that are predicted from previous data

gnd=1&air=1 means collecting airlines both in the air and on the ground

vehicles=0 means ignoring vehicles on the ground

maxage=14400 means setting the maximum age of the cookie to 14400 seconds

gliders=1 means gliders are taken into account

stats=0 filters data which is not related to the detail of the flights

- In which format was the response of each requested URL you sent?

  After sending the requested URL above, the website responses with a Json file.

- How did you convert it to a structured way that Python could understand?

  With build-in package 'json', the texts in Json files can be directly changed into type 'dict' or 'list'

- What does each part of the response means? Use a minimal response example to explain.

   'For value"0",we consider it as ICAO-24bit address, because we directly see the ICAO-24bit address which is provided on the website, by comparing the value, we can know the ICAO-24bit address is equal to value"0".',

   'For value"1",we consider it as latitude, because the value in value"1" are all in the range between 30.90 and 31.53.',

   'For value"2",we consider it as longitude, because the value in value"2" are all in the range between 120.91 and 122.17.',

   'For value"3",we consider it as track, because we directly see the track which is provided on the website, by comparing the value, we can know the track is equal to value"3".',

   'For value"4",we consider it as altitude, because we directly see the altitude which is provided on the website, by comparing the value, we can know the altitude is equal to value"4".',

'For value"5",we consider it as ground speed, because we directly see the ground speed which is provided on the website, by comparing the value, we can know the ground speed is equal to value"5".',

'For value"6",we consider it as squawk number, although the squawk number is not provided by the website for free, we can see the types of all the values in value"6" are string, and all the strings are four numbers. We can guess that the four numbers are the squawk number, because we already knows that "7700" is the squawk number for mechanical fault from the movie 《中国机长》. We can also know that squawk number is always four numbers by searching the Internet. So by guessing, the squawk number is equal to value"6".',

'For value"7",we consider it as the type of radar, by searching the Internet, we can now that this part of data shows the type of radar used to communicate with the flight. And from the website we can know that the type is chosen in all the receivers that detect the flights.',

'For value"8",we consider it as the type of the airplane, by comparing the data on the website UI, searching the Internet and common knowledge, for example, "A320" and "A321" are common type of the airplanes, we can now that this part of data display the type of the airplane.',

'For value"9",we consider it as registration , by comparing the data on the website UI, we can now that this part of data display the registration of the airplane.',

'For value"10",we consider it as the last time stamp when radars detected the flight, by changing the timestamp into local time, we can see that the local time is close to the time we request the url, it also changes when we requests again. So our team guess that this part of data shows the last time the website detect the airplane.',

'For value"11",we consider it as departure airport's IATA, by comparing the data on the website UI, we can now that this part of data display the IATA of the departure airport of the airplane.',

'For value"12",we consider it as arrival airport's IATA, by comparing the data on the website UI, we can now that this part of data display the IATA of the departure airport of the airplane.',

'For value"13",we consider it as flight number(IATA), by comparing the data on the website UI, we can now that this part of data display the flight number of the airplane and the airline company is shown in IATA code.',

'For value"14",we consider it as whether the airplane is landed，1 stands for landed，0 stands for in-air, by comparing the other data of the flights, we can see that the flight with value"14" is 1 all have 0 in altitude and the flight with value"14" is 0 all have positive values in value"14". So our team guess this part of data shows whether the airplane is on the ground.',

'For value"15",we consider it as the rate of climb of the airplane, by comparing data shown on the flightradar24 website UI and the data on the flightaware website UI, we can know that this part of data shows that the rate of climb of the airplane, and the unit of the data is miles per minute',

'For value"16",we consider it as flight number(ICAO), by comparing the data in value"13", the only change is the letters before the numbers. By searching, the three letters(ICAO) and two letters(IATA) both mean one airline company. So this part of data also means the flight number and the airline company is shown in ICAO code.',

'For value"17",we guess it is a kind of status code for a flight, 0 stands for normal, 1 stands for abnormal, it is totally a guess, because we can not find the data that this part of value is "1". And we can only find "0" data in this part.',

'For value"18",we consider it as the ICAO code of the airline company of the flight, according to the value"13" and the Internet, we can know that this part of data shows the ICAO code of the airline company of the flight.'

Implementation

- What packages and modules did you use in this part to get flight data from FlightRadar24 or FlightAware server and parse the result? List them and describe the functionality of each of them in this project.

  1. csv

     Write data collected from the website into a csv file for storage.

  2. json

     Transfer json files into easy formats like dict and list

  3. requests

     Send requests to URL with a given header to prevent error[451]

  4. time

     Make the interval of gathering data