

Homework

for Junior Data Extraction Specialist position

Using the given URL <http://homeworktask.infare.it/> create a web scraper that extracts roundtrip flight data from this API and saves information to a local CSV file.

Notes:

- *The website acts as a front-end for a JSON based API. The result page doesn't have a front-end, so in order to better understand the JSON data, use the front-end example provided on the start page. The example visualizes data you will get from the API with given search parameters.*
- *Feel free to use any coding language.*
- *Prefer HTTP requests over headless browsers.*
- *Data needs to be saved in CSV format. The CSV data structure should be the same as in the CSV example file provided on the start page.*
- *Reviewing CSV file content will help you understand what data is required.*
- *It is ok not to complete all the steps in order for us to accept it.*

Steps of the task:

1. Implement a pageload to get the JSON data with your scraper.
2. Extract outbound and inbound flight data flying from MAD to AUH. You may choose any dates.
3. Make outbound and Inbound flight combinations for each price category (roundtrip flights).
4. Extract **all** available prices and calculate taxes for each combination.
 - Data can be saved into a CSV file for examination.
5. Find the **cheapest** price option **for each flight combination**.
 - Data can be saved into a CSV file for examination.
6. Make sure that scraper can work with flights having 1 connection (example routes: JFK-AUH, CPH-MAD). Flights having 2 connections must be skipped.
7. Make sure that scraper can work with any search parameter set (origin, destination, dates).
8. Save extracted data into CSV file using multiple search parameter sets (choose 10 of any search parameters you want).

Bonus step:

9. Implement an additional filter allowing only flights having a specific connection airport (provided with the search parameters) OR direct flights.

Data to extract (CSV example visible on the start page):

- Departure and arrival airport three-letter IATA codes for each flight (including connections).
- Departure and arrival dates with times for each flight (including connections).
- Flight numbers of each flight (two-character airline company designator with flight number in digits ex. BA4040)
- All roundtrip flight combinations with the **cheapest** full price.
- Roundtrip flight combination total taxes.

Suggestions:

- HttpClient library to make HTTP request if using C# or something similar for other languages.
- Newtonsoft library for working with JSON format if using C# or something similar for other languages.
- <http://jsonviewer.stack.hu/> as visual JSON viewer.
- <https://www.telerik.com/fiddler> for network traffic tracking.