CUSP-GX-5008.001: Big Data Management & Analysis

SPRING 2016

Homework 2

This homework is for practicing Python's *generators* and *streaming*. We're going to use the Citibike dataset posted on NYU Classes. **Please make sure to download the citibike.csv file again** as it has just been corrected recently to have all the records sorted by starting time. This is needed for task 2 of the homework.

TASK 1

Your task is to compute the median age of the Citibike's subscribed customers. You are required to read data line by line and are not allowed to store the entire data set in memory. Indeed, you should not have any containers (e.g. list, dictionary, DataFrame, etc.) with more than 100 elements in memory. You can use the **citibike.csv** data file that we have on NYU Classes for testing, but we will evaluate your code on a much larger input to ensure it's streaming capability.

You will turn in a stand-alone Python program named **hw2_task1.py** (i.e. just a single .py file, no notebooks) that takes a CSV file as arguments and print out a single number showing the median age of the subscribed customers in that file.

EXAMPLE:

```
python hw2_task1.py citibike.csv
39
```

TASK 2

Your task is to write a generator to extract the first ride of the day from a Citibike data stream. The data stream is sorted based on starting times (similar to the **citibike.csv** file uploaded on NYU Classes). The first ride of the day is interpreted as the ride with the earliest starting time of a day. For the sample data, which is a week worth of citibike records, your generator should only generate 7 items (one for each day).

You are given a template **hw2_task2.py** with a template generator **first_ride**. The generator currently takes in csv.DictReader generator and output its first element. Please adjust this generator to output the first ride of the day for the entire stream as specified above. The output of the generator must be in the same format as csv.DictReader. You can think of this generator as a filter only passing certain records out. Please turn it your modified **hw2_task2.py** as a submission for this task.

EXAMPLE:

```
python hw2_task2.py citibike.csv
1,,801,2015-02-01 00:00:00+00,2015-02-01 00:14:00+00,521,8 Ave & W 31
St,40.75044999,-73.99481051,423,W 54 St & 9 Ave,40.76584941,-
73.98690506,17131,Subscriber,1978,2
…
```