COMP9321 Assignment One: Welcome to Barcelona

~~Version:1.2~~ Version 2.2
Weight: 10 Marks

Due: **23:59:59 9th March, 2019**

Change log

- Version 1.1. Bonus mark mechanism changed. Now you can use the bonus mark to make up any assignments which you lose marks. However total marks for assignments will cap at 40
- Version 1.2. Add the late penalties
- Version 2.0. Add sample output format from Q1 to Q4. Format testing for Q1 is available on give, you can get some feedback about your output format. Q5's coordinate fixed.
- Version 2.1. Fixed output format checking issues check for Q1 and Q3. Removed checking for Q2 and Q4. Added some specific details regarding to formatting in **Q0**. Add some comm question's answer in corresponding question. Also we fix some floating point problem about the sample output.
- Version 2.2. Support submission your own map.png.

Pre-requirements

- You should able to write some workable python code fluently ($\leq$ 999 lines)
- Know how to use basic python data structure and packages like matplotlib, math, csv,re....
- Python version is **3.7.2** which is the version installed on CSE machine. Packages are not limited in this assignment, as long as it is pre-installed in standard CSE environment.
- You can always try your codes on CSE login servers[1]

Overview

In this assignment, you will be given some 'real experience' on your week 1-2 lecture content. You are asked to make some CRUD(create,read,update and delete) operations to the data-set which we provided. After finish this assignment you should have some idea about how to do data ingestion,cleansing and manipulation on a small-size data-set(less than 5MB). In this assignment, we will work on the data which from Barcelona.

[1] See https://taggi.cse.unsw.edu.au/FAQ/Logging_In_With_SSH/#CSE_Login_Servers

Tasks

Question 0.
(0 mark)
Download the zip file a1.zip which contain a python code template a1.py and a few csv files: Do **not** change the file name.
Noted:

- For assessment reason, printed data in table rows should be separated by one single space;
- Some inconsistent names in files **don't** need to be changed in this assignment(e.g. Sant Martí and Sant Marti, also Horta-Guinardó and Horta-Guinardo, "Meridiana" and "Av Meridiana");
- Names like ARAGÓ shall be converted to Aragó ;
- Names in dataset may be in BLOCK LETTERS(Upper cased) or lower cased, they should be corrected to Title Style, except for "la", "de", "d′" and "l′". e.g. "El Camp de l′Arpa Del Clot" ;
- Multiple street values shall not be changed and be kept as it is;
- For Q2 - Q5, invalid data like "Unknown", "–" shall be removed;
- In Q4, you need **only** match by *hour, day, month* and *district names* of stations and accidents; You can assume all data are in the **same year**; Accident outside the range of air stations **can** be ignored;
- Sample testing for Q1 and Q3 is updated; We are planning an update sample testing and release a less buggy version for Q4 part 1,2;
- Human review will be involved in terms of final marking for this assignment, which means some formatting disagreement in the sample auto testing will still be marked right.

Question 1.
(1 mark)
Barcelona is second largest city on Spain, accidents are quite normal in a city of such scale. In this question, you are require to read the accident data("accidents_2017.csv") correctly and print table head with first 10 lines of data.

Sample output is as follow:

```
Id "District Name" "Neighborhood Name" Street Weekday Month Day Hour "Part of the day" "Mild injuries" "Serious injuries" Victims "Vehicles involved" Longitude Latitude
2017S008429 Unknown Unknown "Número 27" Friday October 13 8 Morning 2 0 2 2 2.12562442 41.34004482
2017S004615 "Sant Martí" "El Camp de l'Arpa Del Clot" "Las Navas de Tolosa" Thursday May 25 14 Afternoon 1 0 1 3 2.1852720000000003 41.416365
...
```

Question 2.
(2 mark)
You may noticed during the first question that some fields in this file is "unknown". In this question you need to remove all lines with "unknown" fields, and save to "result_q2.csv".

Sample output is as follow:

`result_q2.csv :`

```
"Id","District Name","Neighborhood Name","Street","Weekday","Month","Day","Hour","Part of the day","Mild injuries","Serious injuries","Victims","Vehicles involved","Longitude","Latitude"
"2017S004615 ","Sant Martí","El Camp de l'Arpa del Clot","Las Navas de Tolosa ","Thursday","May",25,14,"Afternoon",1,0,1,3,2.185272,41.416365
"2017S007775 ","Sant Martí","El Camp de l'Arpa del Clot","Indústria / Trinxant ","Wednesday","September",20,12,"Morning",1,0,1,2,2.183245,41.416336
...
```

Question 3.

(3 mark)

Statistics of accidents can be useful, and in this question, you are asked to produce and print a table of total numbers of accidents in different district("District Name" in the dataset) with names, descending ordered. Note: Using the data which don't have the "unknown" field Sample output is as follow:

```
"District Name" "Total numbers of accidents"

Eixample 3029

"Sant Martí" 1104

...
```

Question 4.

(4 marks)

It is also interesting to view different data together. "air_stations_Nov2017.csv" contains air quality station information while "air_quality_Nov2017.csv" is the air quality logs. Firstly, print the air station names with its district names, in json format; Secondly, print the first 10 records that the air quality is **NOT** "Good"; Finally, save the **accident** data when the air quality is **NOT** "Good", into 'result_q4.csv', in the same format of the original "accidents_2017.csv".(Using the cleaned data)

Sample output is as follow:

```
[...,{"Station":"Barcelona - Vallvidrera, El Tibidabo I Les Planes","District Name":"Sarri\u00e0-Sant Gervasi"}, {"Station:  ..., "District Name":"..."}, ...]

Station "Air Quality" Longitude Latitude "O3 Hour" "O3 Quality" "O3 Value" "NO2 Hour" "NO2 Quality" "NO2 Value" "PM10 Hour" "PM10 Quality" "PM10 Value" Generated "Date Time"

"Barcelona - Eixample" Moderate 2.1538 41.3853 0h Good 1.0 0h Moderate 113.0 0h Good 36.0 "01/11/2018 0:00" 1541027104

"Barcelona - Eixample" Moderate 2.1538 41.3853 20h Good 1.0 20h Moderate 92.0 21h Good 17.0 "02/11/2018 21:00" 1541189103

...
```

`result_q4.csv :`

```
"Id","District Name","Neighborhood Name","Street","Weekday","Month","Day","Hour","Part of the day","Mild injuries","Serious injuries","Victims","Vehicles involved","Longitude","Latitude"

"2017S005097 ","Eixample","Sant Antoni","Corts Catalanes / Vilamar ","Saturday","June",10,16,"Afternoon",1,0,1,2,2.15178,41.377044

"2017S009475 ","Eixample","Sant Antoni","Paral·lel / Floridablanca ","Wednesday","November",15,17,"Afternoon",1,0,1,2,2.154167,41.375

...
```

Question 5.

(Bonus 3 marks, with all other assignments, capped to 40 marks)

New map file available for downloaded.

Plot a Heat Map to show the total accident data on the provided map("Map.png"), where the coordinates start from **UTM 31T 409584 4594121 to 31T 451699 4570324**. The ploted map should be saved as "plot.png" with the same size of the original map (some data may be outside the range and they should be considered as outliers).

You are free to use your own map and ignore the coordinates above. And when submission, simply

```
give cs9321 assn1 a1.py map.png
```

Submission

Submit you python source code through give commend on CSE machine(or VLAB/SSH):

```
give cs9321 assn1 a1.py
```

or

```
give cs9321 assn1 a1.py map.png
```

Note: If your code fails to run, you will get 0 on this assignment. Please make sure your python version is 3.7.2, and it works properly on the CSE machine. We don't accept any reason like:' It works on my own computer, but I don't know why it doesn't work on CSE'.

Late Penalties

You will lose 2 marks for each day the assignment is late.

Plagiarism

Plagiarism will result in 0 marks in Term 1, 2019 for this course.

Finally, and also very importantly, have fun!
COMP9321 administration team.