Homework 3
(*Individual Homework)

MSAN 694, Diane Woodbridge
Due: Nov 9th, 11:59 PM

Description

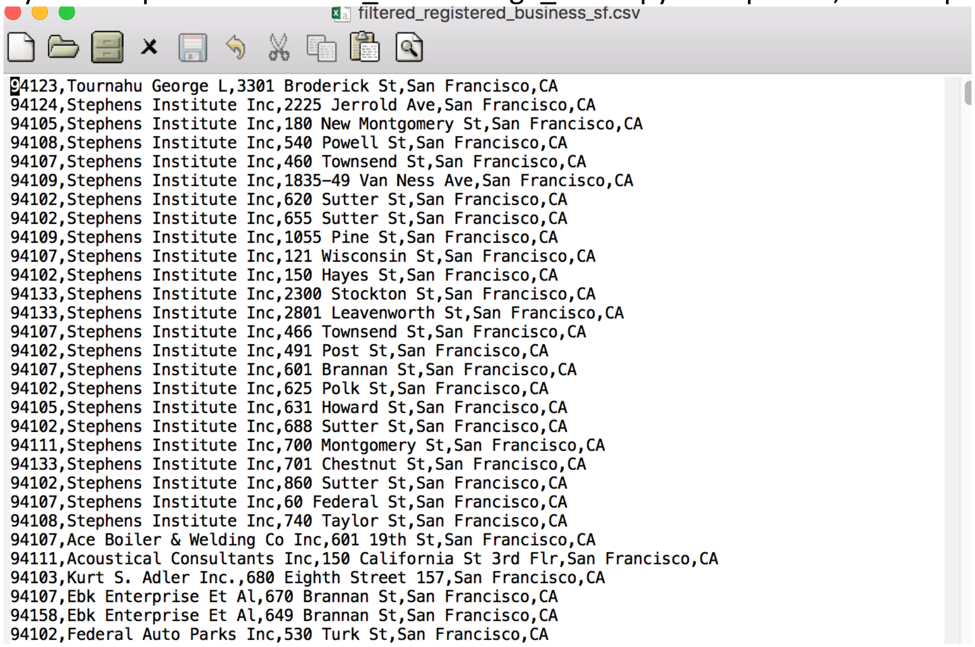
Given the *input_file* and *app_name* in “user_definition.py”, complete the python code “hw3.py” to:

1. Preprocess the input data.
    - Remove all duplicate lines.
    - The input file must have 5 fields including zip, name, street, city and state where name, street, city and state could be an empty string. Remove any lines that contain no zip field.
    - Zip must be a 5-digit number. - Remove any lines with malformed zip codes.
2. Print the ten zip codes with the most businesses and the respective number of businesses for each zip code.
    - Order by the number of business.
    - Each line should be formatted as *<zip> : <number of respective businesses>*
3. Print all zip codes that correspond to more than 1 state in the file.
    - Order by zip.
    - State names are case-sensitive (CA≠ca, CA≠California).
    - If a state name is an empty string, print out as *NULL*.
    - The list of states should be ordered with *NULL* appearing first.
    - Each line should be formatted as *<zip> : <list of ordered state names>*.
4. Print an empty line between step 2 and 3 (example below).

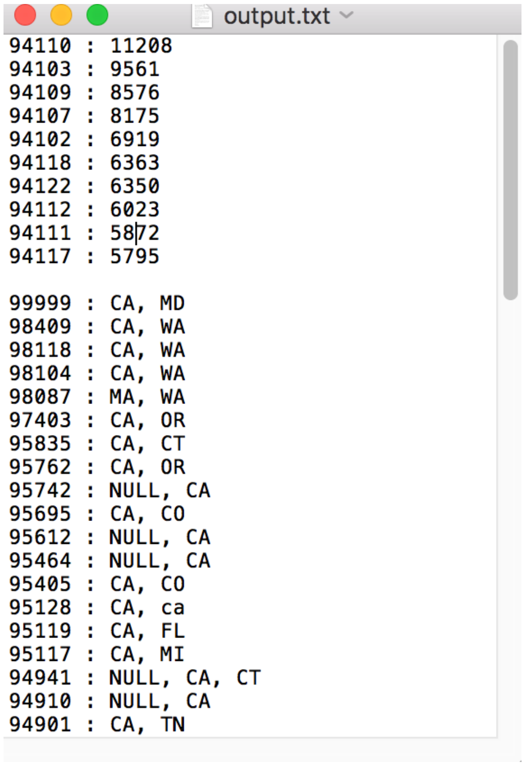
Submit the hw3.py file (**ONLY**) - the name of your file should be **hw3_LastName_Firstname.py** on Canvas. Make sure it runs in **Python 2.7**.

This assignment counts **6%** of your final grade.

We provide two example input file(input_1/ filtered_registered_business_sf.csv, input_2/ input.csv) and corresponding output.txt.
If you run spark-submit hw3_Woodbridge_Diane.py > output.txt, the output should be:



```
94123,Tournahu George L,3301 Broderick St,San Francisco,CA
94124,Stephens Institute Inc,2225 Jerrold Ave,San Francisco,CA
94105,Stephens Institute Inc,180 New Montgomery St,San Francisco,CA
94108,Stephens Institute Inc,540 Powell St,San Francisco,CA
94107,Stephens Institute Inc,460 Townsend St,San Francisco,CA
94109,Stephens Institute Inc,1835-49 Van Ness Ave,San Francisco,CA
94102,Stephens Institute Inc,620 Sutter St,San Francisco,CA
94102,Stephens Institute Inc,655 Sutter St,San Francisco,CA
94109,Stephens Institute Inc,1055 Pine St,San Francisco,CA
94107,Stephens Institute Inc,121 Wisconsin St,San Francisco,CA
94102,Stephens Institute Inc,150 Hayes St,San Francisco,CA
94133,Stephens Institute Inc,2300 Stockton St,San Francisco,CA
94133,Stephens Institute Inc,2801 Leavenworth St,San Francisco,CA
94107,Stephens Institute Inc,466 Townsend St,San Francisco,CA
94102,Stephens Institute Inc,491 Post St,San Francisco,CA
94107,Stephens Institute Inc,601 Brannan St,San Francisco,CA
94102,Stephens Institute Inc,625 Polk St,San Francisco,CA
94105,Stephens Institute Inc,631 Howard St,San Francisco,CA
94102,Stephens Institute Inc,688 Sutter St,San Francisco,CA
94111,Stephens Institute Inc,700 Montgomery St,San Francisco,CA
94133,Stephens Institute Inc,701 Chestnut St,San Francisco,CA
94102,Stephens Institute Inc,860 Sutter St,San Francisco,CA
94107,Stephens Institute Inc,60 Federal St,San Francisco,CA
94108,Stephens Institute Inc,740 Taylor St,San Francisco,CA
94107,Ace Boiler & Welding Co Inc,601 19th St,San Francisco,CA
94111,Acoustical Consultants Inc,150 California St 3rd Flr,San Francisco,CA
94103,Kurt S. Adler Inc.,680 Eighth Street 157,San Francisco,CA
94107,Ebk Enterprise Et Al,670 Brannan St,San Francisco,CA
94158,Ebk Enterprise Et Al,649 Brannan St,San Francisco,CA
94102,Federal Auto Parks Inc,530 Turk St,San Francisco,CA
```



```
94110 : 11208
94103 : 9561
94109 : 8576
94107 : 8175
94102 : 6919
94118 : 6363
94122 : 6350
94112 : 6023
94111 : 5872
94117 : 5795

99999 : CA, MD
98409 : CA, WA
98118 : CA, WA
98104 : CA, WA
98087 : MA, WA
97403 : CA, OR
95835 : CA, CT
95762 : CA, OR
95742 : NULL, CA
95695 : CA, CO
95612 : NULL, CA
95464 : NULL, CA
95405 : CA, CO
95128 : CA, ca
95119 : CA, FL
95117 : CA, MI
94941 : NULL, CA, CT
94910 : NULL, CA
94901 : CA, TN
```