Homework 4

MSDS 694, Diane Woodbridge

Description

Given the *app_name*, *input_file, output_file, day, start_time* and *end_time* in "user_definition.py", complete the pyspark script "hw4.py" to:

1. Print the total number of lines. (0.5 pt)
2. Print the total number of unique business information – Each information should be unique. (1 pt)
3. Print the number of unique business information for each day in descending order based upon
   1) numerical value and then 2) alphabetical value of the day. (2 pt)
4. The number of businesses open between *start_time* and *end_time* on *day.* (2 pt)
5. Follows pep 8 standards – Should not have any errors when running pycodestyle (0.5 pt).

You can assume that the file format is always the same.
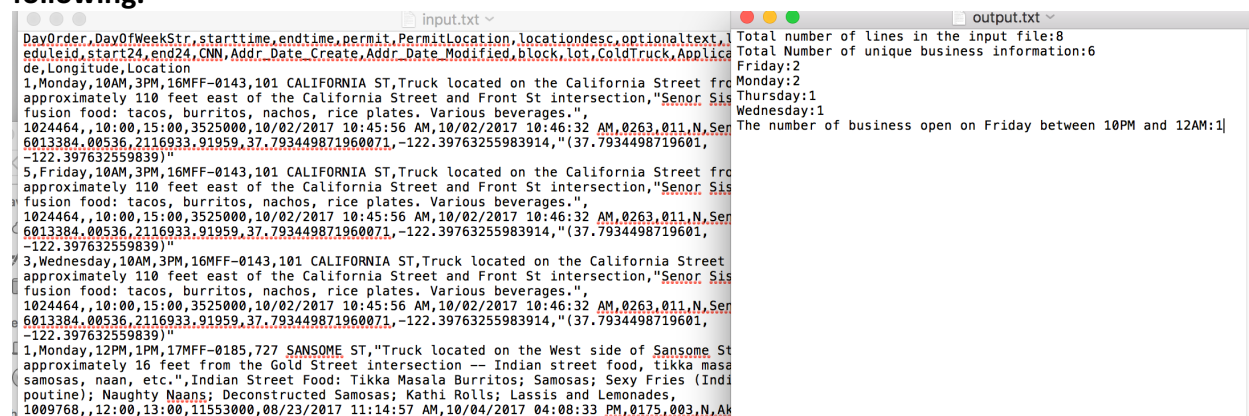Do not include any additional libraries other than those given in the file.
Please see the output.txt for the detailed formatting.
If you run diff output.txt input_n/ouput.txt, it should not return any messages(4pt).

Submit the hw4.py file (**ONLY**) - the name of your file should be **hw4_LastName_Firstname.py** on Canvas. Make sure it runs in **Python 3.6 and pyspark 2.3**.

We provide two example input files (input_1/input.txt, input_2/input.txt) and corresponding output.txt. **Please do not hardcode input_file, output_file, day, start_time and end_time (They are given in the user_definition.py).**
If you run spark-submit hw4 _Woodbridge_Diane.py, the output should be **same as the following.**