SI 618 Fall 2022 Homework 7 (100 points)

We are going to analyze the best hotels based on Trip Advisor reviews for this homework. The data was originally from the Hotel-Review Datasets. There are ~880, 000 reviews in the original dataset. I've cleaned and preloaded the datasets to the 618 folder on the Great Lakes cluster:

```
/scratch/siads618f22_class_root/siads618f22_class/shared_data/hw7_data/offering_cleaned.txt
/scratch/siads618f22_class_root/siads618f22_class/shared_data/hw7_data/review_cleaned.txt
```

The hotel class is the only column that might have a null value in the dataset and please only include the hotels that have a hotel class for all of the analyses in this homework. (Do not drop the null values from the dataset directly, instead, please use only SparkSQL query to filter out the null values).

In the .zip file, there is a readme file with the header information of the dataset that might be useful to you.

Best Hotels based on Trip Advisor reviews

Load the datasets with SparkSQL, and name the temporary view as 'offering' and 'review'. You may want to print out the schema of the table first to get a basic idea of the columns in the dataset. Then, use SparkSQL to solve 3 questions.

Q1. (30 pts) Write 1 SparkSQL query to find the top 100 hotels based on the average ratings. In case of ties, please sort the result by the id of the hotel. The output should have the format of the id of the hotel, the average rating of the hotel, the hotel class, the locality, the name of the hotel, and then the region id. Save the result to uniqname_hw7_q1.csv.

Q2. (30 pts) Write 1 SparkSQL query to find out the localities that have the most luxury hotels (hotels that have at least 4 stars). Please include the locality, number of luxury hotels, and the average rating of luxury hotels in that locality in the result. Save the result to uniqname_hw7_q2.csv.

Q3. (40 pts) For all the localities in the dataset, find out the average rating of hotels of each hotel class. Sort the result by locality based on **the average rating of all the hotels of each locality** irrespective of the hotel class in descending order, then by hotel class in descending order. Only use the reviews that have at least 2 votes and their author have more than 10 total votes. Please try to write only 1 SparkSQL query to solve this problem. The output should include locality, hotel class, and average rating.
Save the result to uniqname_hw7_q3.csv.

For each of the 3 questions, we provided a file that contains the header and the first 5 rows in the desired result. Please utilize these files to check the correctness of your result and the format of your result.

Your Spark code should run as a standalone application on the Great Lakes cluster. That is to say, we should be able to run your code with the **sbatch batch-job.sh** command from the command line. You MUST use **SparkSQL** to do this homework. Other solutions will not get any credit.

What to submit:

1. si618_hw7_youruniquename.py
2. uniqname_hw7_q1.csv
3. uniqname_hw7_q2.csv
4. uniqname_hw7_q3.csv

Submit these as separate files and not as a zipped folder. It helps the IAs to grade your work more efficiently