ISOM3400 - Assignment 6 (Group Assignment)
A Data Mining Project

Due Date & Time: July 16th, 2019 (11:59pm)

You will be given a review.csv dataset which contains 10000 real review data for businesses in the US. You have four tasks with this dataset.

1. Task 1: Import the dataset and explore the following features of the dataset

1.1. Shape of the dataset; descriptive statistics of all numeric columns; and check if there are any missing values in the dataset
1.2. Find out the distribution of star ratings (how many 1s, 2s, 3s, 4s, and 5s)
1.3. How many businesses were reviewed
1.4. List top 10 businesses with best average ratings (show business ids with ratings)
1.5. List top 10 businesses with best average ratings, but only those businesses with more than 10 reviews (show business ids with ratings)
1.6. list top 10 reviewers whose reviews got the most total number of useful ratings (show reviewer IDs with total number of useful ratings)
1.7. Plot a bar chart for star ratings (ordered 1 through 5)
1.8. Plot and see the relationship between review text's lengths and cool/useful/funny respectively (three plots) [Hint: you will need to create a new column for review text's length]
1.9. Any other useful/interesting statistics & plots you may get or draw from the data

2. Task 2: Use text mining to predict star ratings from review text

Your main goal in Task 2 is to find out the relationship between the review texts and the star ratings, as you are interested in knowing what makes people give good versus bad ratings of a business.

For easier comparison, please keep default percentage of test data, and random_state=1 for all train/test splits!

2.1. Train/test split data, and then predict. Report prediction accuracy and confusion matrix. **Briefly discuss your findings in a word document** (e.g., what do you think about the accuracy score; what can you observe from the confusion matrix?).
2.2. Repeat step 1, but only with reviews that give start ratings 1 or 5. **Briefly discuss your findings in the word document** (e.g., any insights from examining false positives and false negative reviews?)
2.3. Based on the result of step 2, find the top 10 token words in the training data that are most predictive of 5-star reviews; and top 10 token words that are most predictive for 1-star reviews. **Briefly discuss** any insights you can get from examining these token words **in the word document**.

Task 3: Try to improve prediction accuracy for:

3.1. Review data with ratings of 1 and 5 only (from the result obtained in task 2.2). For example, is there any other model that might give better prediction accuracy?
3.2. Review data with ratings from 1 through 5 (from the result obtained in task 2.1). For example, can you get better prediction accuracy by filtering out some unreliable reviews?

Discuss your findings in the word document with your observation and analysis. Please document all valid efforts you made, **even if some of them may NOT result in better prediction accuracy**!

Task 4: Anything else interesting that you can find in the dataset? I mean, ANYTHING :-) **Discuss your findings in the word document.**

Your Turn-ins:

1. A Jupyter file with all of your codes (with appropriate comments to assist understanding of your codes for grading)
2. A word document with your discussion and analysis. Limit to two pages, font size 12.

Grading Guidelines:

Jupyter file requirement:

- Correctness of the codes
- Proper comments to assist understanding of your codes

Word report requirement:

- Your discussion needs to be supported by your data and findings in the Jupyter file
- Quality of writing
- Insightful and interesting findings/observations
- For tasks 3 and 4, note that both successful and unsuccessful attempts are of value! It reflects your business senses, and data mining is inherently a trial-and-error process!

Peer Evaluation:

A peer evaluation form will be supplied upon request. For students who do not contribute their share to this group assignment, and take the advantage of others' work (i.e., free-riders), their scores will be adjusted base on the peer evaluation. Once a group member requests a peer evaluation, everyone in the group will need to fill out the peer evaluation form. Each student shall accurately report the contribution of every member in the group. Making false claims in the peer evaluation is an Academic Misconduct. The instructor will make a judgment of each case. No appeal is allowed after a judgment is made by the instructor. A request for peer evaluation has to be made on or before the due date of this assignment.