DATA SCIENCE FINAL PROJECT

Yelp Data set

November 6, 2019

Please use these data sets and complete following tasks. You need to write a report where you explain the preprocessing/data exploration steps for, the method(s) you used, the result you received (plots and/or numbers). Remember, you need to return the report as a single PDF file (5-8 pages) and you do not need to include your code. You are allowed to apply a different method, but you need to be able to explain the method in a way that your peer can understand it. You may use all the material provided during the course, built-in functions in different packages, and online sources. Remember to cite the sources. The outline of the report is given in the slide "Final project and Peergrade".

1 Description of the data set

The data set has been collected from the Yelp dataset: `https://www.yelp.com/dataset/documentation/main`. The reviews have been randomly selected from the file *review.json* for this project, and it contains reviews of different businesses written by Yelp users. These reviews are under "text" column of the file. Additionally, this file includes columns **business_id, cool, date, funny, review_id, stars, useful, user_id**, where the columns "cool", "funny", and "useful" refer to how many users found the review cool, funny, or useful. The column "stars" contains a star rating associated with the review. The original file is quite large, therefore; some preprocessing has already been done for you. You will get two files:

1000_random_review.csv file contains 1000 randomly selected reviews. **bow1000.csv** includes a text matrix/bag of word of the text-field comments. The stop words and some anomalies have been removed from the bag of words representation of review texts. Please complete the following tasks and include your results and analysis in your report:

2 Preprocessing and data exploration

Start with preprocessing of the data sets. You are required to complete the following steps:

- For 1000_random_reviews.csv data set:
  - Load the data set 1000_random_reviews.csv using the Pandas library.
  - Visualize the distribution of the star ratings. Is there a bias or an imbalance in the distribution? Explain your observations.
  - Use the column "stars" to select the reviews with star rating equal or higher than 4, and generate a "wordcloud" plot for 100 most frequent words of the selected reviews. Then select the reviews with star rating lower or equal to 2 and do the same. Analyse your results.
  - Generate the pairs plots and observe correlation between features and feature-labels.
  - Use PCA for BoW1000.csv to plot the data as a two-dimensional scatter plot. Explain your plot. How many PCA components are need to reconstruct 50% of the original data?
  - Categorize the data in 1000_random_reviews.csv into two different classes: poor reviews and good reviews. A good review has the star rating above or equal to 3.5. A poor review has a rating of 3.0 or lower. Create a new column, named "category". If the star rating of a comment is 3.0 or less, then categorize the comment as 0 (poor), otherwise categorize the comment as 1 (good).
- For bow1000.csv data set:
  - Load the given bag-of-words data from the file.
  - Split the array into a training and test sets: 800 first rows for the training set and 200 last rows for testing.

- Find the 10 most frequent words. What can you say about them?

3 Regression

The goal of this task is to predict how well the column “useful" characterizes the review. In order to do so, you need to:

- fit (learn) a model on the training set by using features (BoW and other features such as funny) to predict the label (“useful" column). Then use the fitted model to predict the “useful" for the test set. Try one or two regression method(s), use plots to show your results. Observe and analyze methods and results.
- Apply PCA for reducing the dimension of your dataset. Plot the variance explained by components and determine the number of components to use in the prediction task. Remember to scale/normalize your dataset before applying PCA. Moreover, you need to obtain the components from the training set and apply them to project both your training and test sets to a lower dimension.
- Use the projected training and test sets to obtain the prediction for “useful" column.
- Compare your results after and before adapting PCA. Discuss your observations.

4 Classification or(and) clustering

In this task, you will predict the label (good or poor) of a review based on the review text. As features (**X**), use the given text matrix. As label use the new “category" column created in the preprocessing task. As before, use the training set to fit a model.

- Try one or two method(s) for predicting the “category" column for the test set. Use plots (such as a confusion matrix) to show your results and discuss your observations.

Tips: “sklearn.linear_model" and namely “LogisticRegressionCV" may prove useful. See the documentation and online sources for more information.

For clustering,

- Try one or two method(s) and different amounts of clusters (10-15). See, if the clusters created have something in common. How many of them are not related to catering businesses? If you are using the “kmeans" -method you may want to try to run the code several times and compare the results. Try to plot the wordcloud for each cluster and compare the plots.

After finishing all of the tasks, write a report based on the results you obtained in different tasks and upload the report to the Peergrade platform.

If you are interested in data science and looking for some challenges to participate, you may find this link interesting “`https://www.yelp.com/dataset/challenge`". Additionally, you might want to check **Kaggle** for Yelp data set.