Lab One: Exploring Table or Text Data

Due Feb 4 by 11:59pm **Points** 10 **Submitting** a file upload

Lab Assignment One: Exploring Table or Text Data

You are to perform preprocessing and exploratory analysis of a data set: exploring the statistical summaries of the features, visualizing the attributes, and addressing data quality. This report is worth 10% of the final grade. Please upload a report (**one per team**) with all code used, visualizations, and text in a rendered Jupyter notebook. Any visualizations that cannot be embedded in the notebook, please provide screenshots of the output.

You can choose to analyze text data or table data. The requirements and rubric for each are explained below.

***A note on grading**: This lab is mostly about visualizing and understanding your dataset. The largest share of the points is from how you interpret the visuals that you make. Making the visuals is not enough to satisfy each of the rubrics below—you should appropriately explain what the implications of the visualizations are. In other words, expect about 20% of the available points for visuals that have no substantive discussion.*

***Graded Example from previous offering**: https://www.dropbox.com/sh/yinn3v3qnzlbep5/AAA-i1u-4ylyxwiCb8nyEDY7a?dl=0 (https://www.dropbox.com/sh/yinn3v3qnzlbep5/AAA-i1u-4ylyxwiCb8nyEDY7a?dl=0)*

Option 1: Table Data (Only Choose One Option)

Dataset requirements: Choose a dataset that is mostly ready to be analyzed. That is, it is already in the format of table data. The following requirements should be met:

1. The data includes categorical features (it can also include other forms of data, but must have categories)
2. The data must be 1,000 rows or larger
3. The data is **not** strictly image or text data
4. The dataset should have some prediction task associated with it (*i.e.*, labels to learn)

Table Data Grading Rubric

- Business Understanding (**15 points total**).
  - [**15 points**] In your own words, give an overview of the dataset. Describe the purpose of the data set you selected (i.e., why and how was this data collected in the first place?). What is the prediction task for your data and why are other third parties interested in the result? Once you begin modeling, how well would your prediction algorithm need to perform to be considered useful to these third parties?
    - Be specific and use your own words to describe the aspects of the data.

- Data Understanding (**30 points total**)
  - [**15 points**] Load the dataset and appropriately define data types. What data type should be used to represent each data attribute? Discuss the attributes collected in the dataset. For datasets with a large number of attributes, only discuss a subset of relevant attributes.
  - [**15 points**] Verify data quality: Explain any missing values or duplicate data. Visualize entries that are missing/complete for different attributes. Are those mistakes? Why do these quality issues exist in the data? How do you deal with these problems? Give justifications for your methods (elimination or imputation).
- Data Visualization (**45 points total**)
  - [**20 points**] Visualize attribute distributions. Choose and visualize distributions for a subset of single attributes. Choose any appropriate visualization such as histograms, kernel density estimation, box plots, etc. Describe anything meaningful or potentially interesting you discover from these visualizations. **Note**: You can also use data from other sources to bolster visualizations. Visualize at least 5 attributes, at least one categorical and at least one numeric.
  - [**25 points**] Visualize relationships between a subset of attributes. Use whichever visualization method is appropriate for your data. Explain any interesting relationships. **Important:** Interpret the implications for each visualization. Visualize at least three subsets of the attributes.
- Exceptional Work (**10 points total**)
  - You have free reign to provide any additional analyses.
  - One idea (**required for 7000 level students**): implement dimensionality reduction using t-SNE, then visualize and interpret the results. Give an explanation of t-SNE dimensionality reduction methods.

Option 2: Text Data (Only Choose One Option)

Dataset requirements: Choose a dataset that is comprised of text documents. That is, the dataset should not yet be pre-processed. It should contain only text divided into documents. The following are required for the dataset:

1. The data includes at least 30,000 words
2. The data has at least 500 documents
3. The data should have a well defined prediction task (i.e., a label to predict for each document)

Grading Rubric

- Business Understanding (**15 points total**).
  - [**15 points**] In your own words, give an overview of the dataset. Describe the purpose of the data set you selected (i.e., why and how was this data collected in the first place?). What is the prediction task for your data and why are other third parties interested in the result? Once you begin modeling, how well would your prediction algorithm need to perform to be considered useful to these third parties?
    - Be specific and use your own words to describe the aspects of the data.
- Data Encoding (**40 points total**)

- [**5 points**] Read in your document data as strings using python. You must read the data as raw text documents. That is, you cannot use an already processed dataset.
- [**15 points**] Verify data quality: remove words from the vocabulary that are not relevant or that you think should not be included. That is, choose a specific vocabulary to choose for your data. Explain why you chose this vocabulary.
- [**10 points**] Convert the data from raw text into a sparse encoded bag-of-words representation. Explain any parameters selected to convert to bag-of-words.
- [**10 points**] Convert the data into a sparse encoded tf-idf representation. Explain any parameters selected and why you chose different values.

- Data Visualization (**35 points total**)
  - [**20 points**] Visualize statistical summaries of the text data such as word frequencies, document lengths, most relevant words, vocabulary size, etc. Choose visualizations that you think summarize your data best. Explain all visualizations.
  - [**15 points**] For at least three target classes, visualize the most common relevant words and word frequencies. Are there any prevalent differences between your target classes? Is there separation in the visualizations that could be used to help classify the classes? If you have many target classes, choose a representative subset of classes to perform this visualization.
    - Word clouds can be powerful visualizations for this task.
- Exceptional Work (**10 points total**)
  - You have free reign to provide any additional analyses.
  - One idea (**required for 7000 level students**): visualize bigram distributions of word-pairs and implement stemming of the words to reduce redundancy in representing the words.