CIS 8392
Topics in Big Data Analytics

#Assignment 3

Yu-Kai Lin

Assignment 3

Step 1. Identify one dataset from Kaggle

- Requirements for the dataset:
  - The dataset should be suitable for image/text mining or machine learning
  - Size requirement:
    - For image dataset: **1k** images
    - For text corpus: **5k** documents/messages
    - For json, csv, or other datasets with rows and columns: **>= 10k rows** and **>= 10 columns**. If the dataset has multiple files, the number of rows (or columns) combined across all files should be greater than 10k (or 10).
  - Code: **No public code written in R on Kaggle on this dataset**. Check the "Code" tab of the dataset page on Kaggle. Use filters to filter by language "R". If there are results, you cannot use this dataset. (Tip: Newer datasets are more likely to meet this requirement)
- Each student will find a unique dataset. No two students use the same data. Use CIS 8392 Assignment 3 Data Signup Sheet to check whether the dataset has been taken by another student. Once you are certain that no other student uses the same dataset, sign up yours in the data signup sheet.

Assignment 3

Step 2. Use R Markdown to achieve the following:

1. Specify author, date, and title in the YAML metadata of your document
2. Describe the Kaggle dataset (title, link to its Kaggle page, summary statistics)
3. Provide a screenshot of the Code section of the Kaggle dataset you use and filter by language "R". This screenshot is to show that no R code is available on Kaggle on this dataset.
4. Visualize the data with at least two appropriate plots (Note: If you are using an image/text dataset, you don't need to do this data visualization.)
5. Summarize and discuss the patterns found from the data visualization
6. Preprocess the dataset when necessary
7. Build a classifier for the data using `keras` or `h2o`
8. Summarize and discuss your findings

Assignment 3

Here are some additional notes about writing a RMarkdown report. Violating these rules may lead to a lower grade.

- Do not modify the file name of the data you downloaded from Kaggle. Since we do not include data in the submission, this ensures that I can download the data from Kaggle and knit your Rmd file directly.
- Put the data in the same folder as your Rmd file. Whenever we run/knit an RMarkdown file, it uses the folder with the Rmd file as the working directory.
- Read the data in your Rmd code chunk using relative path. If you use an absolute path, I will not be able to knit the Rmd file to an html file from my end.
- You will lose 5 points if for any reason (input path, error in code, etc.) the Rmd file cannot be knitted to an html file.
- All tables (any output of a data frame) must be formatted using kable in your R Markdown report.
- Distinguish headings (## heading) and normal text. We should not put all the text in headings.
- Do not put your discussions/explanations in code chunk. Write them as normal text.
- Do not use `include=FALSE` or `echo=FALSE` in your code chunk. I need to read your code. You may use `message=FALSE`, `warning=FALSE` to suppress messages/warnings.
- Do not write an excessively long line of code. Break it into multiple lines to improve readability.

Assignment 3

Step 3. Knit the R Markdown file (.Rmd) to an HTML file

Step 4. The Rmd and HTML files must follow the naming rule below:

- `Assignment3-YourLastName-KaggleDataTitle.FileExtension`
  - For example:
  - Assignment3-Lin-Black Friday.Rmd
  - Assignment3-Lin-Black Friday.html

Step 5 Submit the two files (individually) to iCollege

Assignment 3

Due by the beginning of next class

Extra credit: the student who has the best report (determined by the instructor) will be given 5 extra points towards the final grade

- Submissions that are too similar would not be considered for the extra credit

Grading is based on the following:

- Grading is based on the submitted files on iCollege. Do not wait till the last minutes before the deadline. You will lose 10 points for late submission. You will receive 0 point if you submit your assignment via email.
- Whether the submission matches the record in the data sign-up sheet
- Whether all required files were submitted to iCollege on time, following the naming rule
- Whether the Rmd file is syntactically correct and can render the html file
- Whether the report has a professional format and style (succinct and yet provides adequate and clear discussions about the data, plots, analyses, and findings)
- Whether the report meets the requirements specified in Step 2
- Whether there is no public kernel written in R/Rmd for this dataset on Kaggle