Assignment 3

DSCI 222 – Data Science Workflow using Python
School of Mathematical and Data Sciences
West Virginia University

Instructions

This is a homework assignment to be solved in a small group (3-4). It is a great opportunity to show your teamwork skills. Enjoy it! All material needed for the assignment can be found in Github

- Even though you will be working in groups, each student should submit their own, own, albeit identical, set of deliverables. All deliverables must be in a folder created in your Google Drive account. The folder name should follow this format: `DS_X_LastName_FirstName`, where **X** represents the assignment number.

  Files included in your deliverables folder

  - Deliverable 1 Video Recording
  - Deliverable 2 Colab Notebook
  - Deliverable 3 Colab Notebook
  - Deliverable 4 Report
  - Anything additional or supplementary that contributed to the assignment (this includes .csv files, .json files, etc.)

- Include your **full name** as a Python comment at the top of the notebook and at the top of your report.
- All data manipulation should occur within the Python script. No manipulation of the supporting file(s) should occur prior to importing the file(s) into your script. You may check your work by manually performing data analysis.
- Include your report in PDF format, written in LaTeX.
- Set share folder permissions so chood@mix.wvu.edu can access and run every file and notebook.
- Everything counts! Include as much as you want in your deliverables, even if the activity is not fully complete by the deadline. *Important:* Review the grading policy and course policies in the online syllabus.
- **Total: 100 points.**

Activity 1: Video Recording (30 points)

Scenario

With this activity, we will be learning yet another visualization technique. We will be expanding on what we already know by learning how to create word clouds that visually represent the frequency of words in a text document. You will need to learn about the wordcloud library, a Python package that can generate word clouds directly from strings of text. Your video task is to provide a lesson on how to create a word cloud for a given text.

Deliverable 1:

Video Recording

- In the Assignment 3 folder in Github, you will find two files `word_cloud.ipynb` and `Country_Roads.txt`.
- Create a video explaining the .ipynb file.
- Your video should clearly explain what each line of code is doing in its entirety. It is expected for every aspect of the script to be addressed and nothing should be ignored.
- Be sure to pay special attention to the color function and each parameter used in wc.
- Your video must be a screen recording of the code as it is being explained, to help follow along. You may include yourself in the video as well, but you do not have to.
- You may add code, comments, and text to the .ipynb file to aid in your explanation.
- All members of the group must speak in the video

Activity 2: Python Source Code (30 points)

Scenario

Time to make your own Word Cloud! You have the choice of analyzing the Bible or the Complete Works of Shakespeare. Your job is to create a Word Cloud of the most used names in the work you select.

Deliverable 2:

Google Colab Notebook

- Create a Word Cloud of the most used character names in the work you select.
- You must us regex at some point in your code
- Explain in your analysis how you determined the number of names to include. This means some sort of data analysis should be conducted prior to creating the word cloud.

- Before each block of code, include text explaining how that block works.
- Within each block of code, include comments explaining how each component of the code works and why it was used.
- Any libraries, functions, methods, etc., used that are not explicitly discussed in class must be clearly explained.

Activity 3: Python Source Code and Report (40 points)

Scenario

data.gov is a great resource for to find all kinds of data. For this third activity, we will be focusing on their collection of JSON formatted data. Using the .json file of your choosing, import the file and perform some sort of data analysis.

Deliverable 3:

Google Colab Notebook

- Import your chosen .json file and perform some sort of data analysis.
- The level of complexity of your analysis will play a large role in the evaluation of this deliverable.
- Before each block of code, include text explaining how that block works.
- Within each block of code, include comments explaining how each component of the code works and why it was used.
- Any libraries, functions, methods, etc., used that are not explicitly discussed in class must be clearly explained.

Deliverable 4:

Report Write a 2-3 paragraph report on what conclusions you are able to draw from this analysis.