Assignment 4

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

- Create and share with me a single folder with all members of the group clearly listed.

  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

The collapse of Enron in 2001 is one of the most infamous corporate scandals in history. If you are unfamiliar, this scandal was a massive corporate fraud in which the energy company Enron used accounting loopholes and deceit to hide billions in debt, leading to its collapse.

Luckily for researchers, a large dataset of internal emails was made public during federal investigations, due in part to Andrew McCallum. McCallum, a professor in the computer science department at University of Massachusetts Amherst, purchased and published these records for $10,000.

This dataset can now be found in a variety of forms all throughout the internet, including places like Kaggle (which many of you are already familiar with from working with your Project), and Hugging Face. To help assist, I have already written a short script that extracts this dataset from Hugging Face for you to use throughout this project. It contains data for over 500,000 emails from over 150 Enron employees over the course of the years leading up to the company's 2001 collapse.

To help better understand the dataset, you can visit CMU's Computer Science page and click on "May 7, 2015 Version of dataset" to download the raw version of the data that we have since cleaned up. Files here are separated by employee, mailboxes, and each email can be viewed as its own file.

Deliverable 1:

Video Recording

- In the Assignment 4 folder in Github, you will find `enron.ipynb`.
- 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.
- This file has three .csv files as outputs. What benefits do each provide, and why would we not just have a single output of the entire email collection?
- What information is included in the output files?
- While the script to extract this data is short, it includes many important elements that may not have been explicitly discussed in class. Be sure to explain what each aspect of the code does, but more importantly, WHY it has been included.
- We have not worked with Hugging Face yet either. At the start of your video, you should include a description of what this is and how it can be useful to you and the data science community.
- 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 should 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

As we conduct analysis on this data, please realize how large 500,000 emails is. My suggestion would be to build your code around our sample dataset, to ensure it works properly on the smaller scale. Then once you have it fully working, implement it on the entire dataset. This should hopefully alleviate a lot of headaches for you.

How can we make this data useful? We are going to first do some exploratory cleaning and analysis to see what is actually contained in this dataset.

Deliverable 2:

Google Colab Notebook

- For this assignment, we are only interested in email correspondence between two Enron employees. So you should remove any unnecessary emails from the dataset. How many emails does this leave us with?
- Add columns to the dataset that list the name of the email's sender and the name of its recipient.
- Determine how many emails were sent and received by each employee. Which 5 employees sent the most emails, and how many did each send? Do the same for received emails. Which 5 employees had the most total email correspondences?
- Determine how many emails were sent each month. Create a visual of your choosing displaying this information. Do you notice any trends?
- If you identify more than one interpretation of a question listed above, make sure to explain why the chosen approach was selected over others.
- 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

Now we are going to start looking for connections. By using NetworkX to analyze the Enron email dataset, we will work to uncover hidden structures of influence, communication patterns, and unusual group dynamics.

Deliverable 3:

Google Colab Notebook

Problems

Problem 1: Build an undirected graph where each node is an employee, each edge is an email connection, and edge weights are determined by the number of emails sent between the two parties.

Problem 2: Compute the weighted degree of each node to determine the 5 most connected employees. Does this match your findings in deliverable 2?

Problem 3: Find the 5 individuals with the highest betweenness centrality, the 5 individuals with the highest eigenvector centrality, and the 5 individuals with the highest closeness centrality

Problem 4: Draw a subgraph of the top 50 most active employees. Use node color to highlight communities and node size to represent centrality.

- If you identify more than one interpretation of a question listed above, make sure to explain why the chosen approach was selected over others.
- 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 report on what conclusions you are able to draw from the problems answered in deliverable 3. Do some quick research on the Enron scandal to see if your findings match real-life convictions (don't worry too much if our relatively short class assignment doesn't align with results of a full-fledged FBI investigation)