Assignment III – IMT 547

This is an individual assignment and work submitted should be written solely by you. Do not copy-and-paste from other students' responses or code. Collaboration is often fun and useful and while it is Ok to discuss at a high-level the general approach to a problem, under no circumstance you should collude to complete this assignment by copy pasting or slightly tweaking someone else's already written code without you making any attempt at solving the question. In other words, each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. **The names of all collaborators must be listed on each assignment**. **This includes anyone you discussed the problem set at a high level.** At the top of your notebook include a markdown to list all collaborators. If none, say so.

Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please DO NOT submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow.

Hint: For most of the questions in this problem set, you can refer to the lab notebook that you all practiced on Week 9 – Advanced Topics & Ethics week.

Due	March 16th by 11:59pm (See late policy on Canvas)
What to submit on Canvas	• Jupyter notebook (pdf and/or html) with answers to the questions listed below. ***Use proper code formatting. Use markdown cells to write questions and descriptive answers. Use code cells to insert your code, run your code and show output.***

In the last two problem sets, you collected tweet data, you had created the `pandemictweets.csv` file, and then had looped through the remaining data science steps of asking questions, cleaning the data, analyzing data. In this problem set, your task is to do additional deep dive analysis on the data.

Note*: If you were not able to collect this data as part of Assignment 1 or if you were not happy with your collection, cleaning and/or analysis in Assignment II, or if you'd just want to work with some other dataset, here is a sample tweet file that you can work with: "trump_20200530.csv". This dataset contains President Trump's tweets from the moment he took office on January 20, 2017 to May 30, 2020. Download the data from Canvas. You are free to down sample this data to keep it within a reasonable number of tweets that will help you with the downstream analysis.*

Q1). Organizing the data into Document-term matrix

Write code to organize the cleaned data that you have from your previous assignment submissions into a document term matrix format. Make sure to exclude common English stopwords.

Q2). Most common words

Using the DTM matrix that you have just built, write code to find the top 30 most common words. What can you infer by seeing the top 30 words?

Q3). Visualize the data

Repeat your inference by visualizing the most frequent words in your corpus. What can you infer from the visual (or visuals)? If you decided to slice the data in some way or take a smaller sample for visualizing, provide a rationale for how you sampled and why. Present your response in the QQQ format. That is, provide a rationale for your choice of visuals, and/or slicing of data, next write code, next draw inference from analysis and visuals.

Q4). Profanity in your data

In this question you will determine the amount of profanity in your data and you need to report your response in the QQQ format. That is, first say in markdown how you define profanity, what decision did you make to detect profanity in your data. Next, write code to find the amount of profanity and show output. And finally draw inference from your analysis and visuals.

Q5). Find topics

What are some key topics you can find in your dataset? Report your response in the QQQ format. That is, first say what all decisions you are making before even trying to run topic modeling, then provide code for topic modeling and show output, then provide inference. Did you get reasonable results in your first attempt? If not, show additional attempts (at least one more) and provide your rationale, code, and inference in QQQ format again.

Q6). Answering additional questions with this data

In addition to the questions that you have answered so far with the data, pick one additional question of your choice that you can answer with additional analysis. In assignment II, you had already listed several compelling questions. You can pick one from there or you can come up with a new question. Use QQQ format to answer this question