Assignment II – 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.

Due	Feb 15th by 8pm (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.*** • If you are attempting the bonus question, then you also need to upload the scraped data file

In the last problem set, you collected tweet data and created the `pandemictweets.csv` file. In this assignment, your task is to work with that collected data and loop through the remaining data science steps of asking questions, cleaning the data, analyzing data.

***Note**: If you were not able to collect this data as part of Assignment 1 or you were not happy with your collection or 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 1000 tweets.*

Q1. Asking questions

What are some compelling questions that you can ask with this dataset (list at least 2 questions)? Enter your responses and rationale behind choosing those questions in markdown cells. *Hint: You had collected pandemictweets.csv file using certain keywords and had provided a rationale behind the choice of the keywords. This can help you think through some questions.*

Q2. Inspect & Data Cleaning

2a). **Inspect**: Write code to inspect the data. What do you observe? Along with the code, write your observation in markdown cell.

2b). **Clean**: Write code to clean the data. Along with code, you need to write the rationale behind the cleaning process, i.e. what are you observing after first level of cleaning, what is still messy and needs additional cleaning, how are you deciding to do it, etc. At this stage, your cleaning should at least comprise 1). Common data cleaning steps and 2). Dealing with messy Twitter data.

2c). **Tokenize**: Write code to tokenize your entire dataset. Use at least 3 different types of tokenizers. Display results from all the tokenizers in a pandas dataframe so that you can visually compare them.

2d). Pick the best tokenizer. Which one do you think works best for your data and why?

Q3. Analyze Data for Sentiment

3a). **Sentiment analysis**: Pick at least 3 different ways of conducting sentiment analysis. Write code to loop through your data and find sentiment for each tweet, using each of these methods.

3b). **Quantitatively comparing methods**: Is there a way to do pairwise comparison of the methods that you picked? You need to report at least one pairwise comparison between the methods. Better if you are able to report all pairwise comparisons.
In you pairwise comparison, compute a quantitative measure to show what proportion of tweets match or does not match between two measures. You can also include the rationale behind the choice of your measure.

3c). **Qualitative + Quantitative comparison**: Write code to randomly pick 25 tweets. By hand, mark each of their sentiments. Now pick two sentiment analysis methods of your choice to automatically find sentiments of these 25 tweets. Considering that your hand-label is the absolute ground truth, write code to figure out which sentiment analysis method works better? It is possible that both methods that you picked are equally good. Provide rationale in markdown cell for your response. *Hint: the more the analysis method's output matches with your hand-label, the better it is.*

Q4. Analyze Data with LIWC

Pick at least two dimensions from LIWC that you would want to investigate with respect to your data. Write code to find what proportion of each of the two dimensions that you picked are present in your data. Motivate your choice of dimension with a research question. For example, if you are curious to know how many angry tweets are there, you can pick the "angry" dimension in LIWC. Another neat trick here would be make your code modular via python functions, so that later you can reuse this function.

Q5. Analyze Data over time

How does the polarity of your corpus change over time? Answer this question by showing plots. What do you observe from the plot?

BONUS Questions:

i). Pick an online site of your choice where you are eager to try out web-scraping skills to fetch data. These can be review websites, like Amazon, etsy, online donation platforms like Kickstarter or gofundme, etc. Fetch at least 10 web pages and associated **text data** using beautiful soup scraping technique. You are free to try other scraping techniques. Here are we are only concerned about the text data. For example, this can be reviews for a product page on Amazon, or project description in case of Kickstarter.

Save the scraped text data in a file and submit it to be graded for the bonus question.

Also, place the text data in a pandas dataframe. Choose appropriate index for the data. Display the pandas dataframe in your notebook.

ii). Clean the collected data and conduct sentiment analysis to show polarity of each of these 10 scraped data points.