Course ID: CSE-506
Course Name: Data Mining

Bonus Assignment 2

September 14, 2022

Instructions

1. The assignment is to be attempted in groups, ranging from one to five students.
2. Programming Language: Python.
3. For Plagiarism, institute policy will be followed.
4. You need to submit the report.pdf, readme.pdf, source code files, images, model files and a PPT.
5. Report, models and code in .py format should be submitted in the Google Classroom in a zip folder with the name 'BA2_Rollnumbers separated by underscore.zip'.
6. You **can use any library** for data collection, pre-processing, and experimentation.
7. One member should submit on Google Classroom while other members can mark turned in without the attachment.
8. In case of doubts, please comment on Google Classroom.
9. You need to **make a presentation** of up to 20 slides in which the first slide should **clearly mention the contribution of each group member**. Penalty will be imposed on slides' limit violation.
10. **You should be well aware of theory of the algorithms mentioned and chosen by you. You will end up losing marks in case you haven't prepared well**.
11. You will need to update in a Google Sheet your weekly progress and this will be graded.

Total Marks: 100 (will be scaled later on)

Deadline: 5th December (This will be strictly followed). We will conduct demos from 6th Dec onwards.

This is a open ended problem, and you will be graded on how you use everything that you learn in class and bring in your own direction, strategies and ingenuity in solving this problem.

For this assignment you need to do review analysis of at least **5000** doctor reviews in English from India.

- Do this for at least **500** doctors and at least **15** specialities.
- Collect the data. **Make sure your data does not match any other groups'.**
- Process and store the data in human readable format, as well as in a format where computational operations can be easily performed. These can be two separate datasets, but there has be a one-to-one mapping between entries.
- The data should have features **including but not limited to** *location, availability days, timings, cost, speciality, patient review score, patient review text, doctors' years of experience, doctors' qualifications.*
- The analysis should take into account positive/negative sentiment of review, number of words, correlation between years of experience, cost, review, location and other features as you deem necessary.
- We will share a Google Form in which you will update your progress every week, what challenges you faced and how you mitigated them. Thus, for the project you will need to define weekly tasks and adhere to the timelines you set.
- Your progress will be monitored every week.

Bonus Assignment 2

N.B.

1. You may use any form of infographic that you feel represents your results and analysis best.
2. Detailed explanation of assumptions made for solving the mentioned problem need to provided.
3. Provide the various parameters and hyperparameters used.
4. Your zip file should contain a folder “visualizations”, “code”, “models”, report.pdf, a readme.pdf file and a ppt.
5. The folder “visualizations” should contain all the images and folder “code” should contain all the codes including notebooks if used.
6. Make a section “Learning’s” in the report.pdf and describe your learning’s from this assignment.
7. In the readme.pdf file, you will have to mention steps to recreate all your experiments and results.
8. Please provide references to all the sources including but not limited to libraries, GitHub Repositories, Research Articles, Blog Posts used in completing this assignment.