DS203-2024-S2: Exercise – 6 (Project)

- **This Exercise (Project) carries 30 marks and 30% weightage**
- **It should preferably be done in groups**
  - **A group can have maximum 4 members**
  - **Team of '1' is strongly discouraged, but not barred.**
  - **Identify one member of the team as the Group's 'anchor'**
- **Read the instructions and the evaluation criteria carefully.**
- **Note and adhere to the submission requirements and deadlines carefully.**
- **Submissions due by: April 13, 2025, 11:55pm**

Evaluation Criteria

Evaluation Criteria	Marks
Process and Results • How good is the EDA and its documentation • Has there been any creative thinking and innovation while solving the problems? • Quality of Feature Engineering / Feature Creation in terms of relevance to the problem • Steps of Data Science: Have they been correctly applied, and backed up with proper metrics / reasons / explanations? • Completeness and correctness of the results achieved (Approximately equal weightage will be given to all the above aspects)	20
Documentation (Presentation) • Completeness and preciseness in the documentation of the process and results [5 marks] • Design, formatting and readability of the slides. [3 marks] • Documentation of 'Executive Summary', successes, failures, hurdles and learnings [2 marks]	10
Penalty • **If the presentation contains raw and hyped-up outputs generated using LLM tools like ChatGPT / Gemini etc.** • **If a project is seen to be 'copied' from another team, both teams will get ZERO marks.**	**-10**
• Viva will be conducted, if deemed necessary, to ascertain originality of work, and to ascertain contribution by team members. • The project will be evaluated **ONLY** by reviewing the presentation. Source code / Jupyter Notebook will **NOT** be reviewed to understand your work; they will be only used to *verify* the claims you have made in the presentation. Therefore, **if you forget to mention some part of your work / analysis in your presentation, it will be concluded that you have not done it**! • If you do not submit your source and data files, your project becomes unverifiable and the submission will not be given any credit.	

Follow the '**Reporting and Presentation Guidelines'** outlined later in this document

Problem Description

In a Data Science course, the students have been creating and submitting summaries of the lecture sessions to the instructor. Unfortunately, the Instructor has not organized the summaries properly with the consequence that the summaries are all jumbled up and it is not known which summary belongs to which session!

Summaries are contained in the spreadsheet: **Session-Summary-for-E6-project.xlsx**

In spite of the 'jumbled up' situation, the instructor is still mighty hopeful that appropriate steps can be taken to recover the lost association and that the summaries can still be used to enable the following tasks / create the suggested applications / visualizations:

1) Thorough EDA of the given data.
2) Re-create the lost connection between the sessions and their summaries.
   - Two different text featurization methods should be tried and the best one finally selected.
3) Create visualization(s) that show the overlaps between sessions.
4) Rank the summaries within a session in terms of their detail and relevance to the overall session.
5) Create visualizations as follows:
   i) A visualization that shows all the sessions as nodes / bubbles, whose size reflects the number of important keywords relevant to that session.
   ii) When a session node / bubble is selected, it should create the following **word-clouds**:
      - A word cloud where the 'prominence' of words is as per their importance only within the session
      - A word cloud where the 'prominence' of words reflects their importance across all the sessions.
      - (If there are ideas / schemes that combine both these in a single visualization, that's fine too.)
6) Create an application to do the following: (Note: The instructor is really looking for a simple interactive application, with not many frills!)
   - Given a set of keywords that describe a topic of interest (no limit on the number of keywords), it should **identify** the most relevant session relevant to that topic and display the top 3 summaries of that session in a separately opened text window.

Clustering of summaries into sessions first of all.

The instructor expects the students to diligently follow Data Science steps while solving the problems, including a thorough EDA phase and good Feature Engineering. In addition to applications / visualizations that really work, **the instructor expects a well-organized and complete yet non-verbose PRESENTATION that succinctly captures EVERYTHING that is done in the project - including failures and challenges encountered**.

Solution guidelines

- Doing justice to this project involves much work, including thoroughly researching into feature creation methods for encoding words, sentences, paragraphs, documents, etc.
- Do what it takes to submit well researched, well-designed, and well-reasoned solutions to the above problems – and effectively communicate them!
- Re-visit all major aspects of Data Science that you have learned so far, and check if they can be / need to be used to solve the posed problems.
- It will really help to work in a team, and divide work amongst the members – to do a good job.
- It is important to have an initial **overall solution design** in place to guide your steps, distribute work among team members, and ensure that you are not overwhelmed, and that you don't get lost!

Reporting and Presentation Guidelines

1. As often mentioned, succinctly communicating your work and results is a very important part of the Data Science process. The most important submissions is the presentation – that summarizes your approach, work, results, achievements, learnings, and possibilities. Budget adequate time for this activity and design your presentation well; last minute work will invariably be shoddy.
2. **Include the names and roll numbers of all group members on the title slide. No credit will be given to members who are not mentioned on the title slide.**
3. Provide an executive overview (1-2 slides) at the start of the presentation.
4. DO NOT use verbose paragraphs, or storytelling, to explain your steps, observations, results, and recommendations. All these should be presented precisely and point-wise.
5. Summarize your observations and results using charts / Tables / metrics and explain them and draw conclusions from them. Merely including plots / Tables, with no associated comment, is not acceptable.
6. Slides should be well designed. Use as many slides as required to completely convey your work.
7. If you do not include something in your presentation, it will be deemed that you have not done it. **Source code / Notebooks will NOT be reviewed to *understand* your work**.
8. Towards the end of the presentation, include slides that clearly answer all the questions posed in the ***Evaluation Criteria*** table. In addition to focussing the evaluator's attention, this will also ensure that you have covered all the expected points in the presentation.
9. Finally, include a slide or two outlining your learnings from this project, and your experiences and hurdles while doing the project.

oooOOOooo