UNSTRUCTURED DATA MANAGEMENT PROJECT (30%)

CIS8045 Unstructured Data Management · MM2 · AY2023/2024 Spring

IMPORTANT DATES

- Topic Submission: March 20
- Project Proposal: March 29
- Term Project Presentation: April 24
- Project Report: April 26 (Submit to iCollege prior to 11: 59 PM)

OBJECTIVE

The term project involves designing and implementing an unstructured data analysis project for a topic of interest. You may work in teams, from 4 to 5 members. The project comprises a project proposal, a project presentation, and a final project report. Please remember that the final deliverables are at the group level and not at the individual level. Your individual project grade will be adjusted based on the peer evaluation within the group.

EXPLORING SOLUTIONS FOR A PROBLEM

The project requires building a data analytics project based on unstructured data for a real-world problem of your choosing. You will be responsible for collecting data for the problem of interest, pre-processing or/and analyzing the data with analytics tools, and presenting the findings for the problem in the suitable form of visuals. You may obtain a dataset from organizations or online sources.

Given that NLP is an important component of this course, the group project must involve certain forms of text preprocessing and text analysis.

TOPIC SUBMISSION

Each group will submit your project topic on the discussion forum in iCollege before **March 20**. Please note that no duplicate topics are allowed, so it is better to submit your topic as soon as you decide to ensure it is not taken. Your topic submission should include two aspects: 1) Your group number and 2) Two sentences describing your topic and the planned data source.

DELIVERABLES

Part A. Project Proposal

The project proposal is a *one-page* document that presents the progress and plan of your project. At a minimum, the proposal should include:

- A brief description of the intended problem(s) to be solved

- Plan of data collection, or a description of potential sources of data
- Plan of methods — intended unstructured data forms to be analyzed, intended NLP techniques and statistical methods to apply, intended relationships to find, how the techniques and methods relate to your proposed problem(s)

* *The proposal is expected to have one page, with 12 font size. Single-line spacing is recommended. Do not include a cover page. If your proposal document exceeds one page, only the first page will be graded. It is important to practice being succinct in communicating your ideas.*

Part B. Term Project Presentation

The project presentation demonstrates students' analytics skills for a problem of interest by the group. At a minimum, the presentation should cover the major process of developing the project: problems to be solved, data descriptions, applied NLP techniques, data analysis methods and tools, and summary of findings.

Part C. Project Report

The project report documents the process of developing the unstructured data analysis project. At a minimum, the report should include:

- Executive Summary (1 page)
- Main Body of the Report
    - Project background (context)
    - Description of the problem(s)
    - Data descriptions – sources of data, description of data fields
    - Data analysis – tools and methods used for analyzing data
    - Description of findings and visualized findings
    - Summary of the story
- Any relevant appendices

Submit the following files to iCollege prior to the due date:

- Project report – PDF Document
- Project presentation – MS PowerPoint (or PDF) Document
- A copy of the dataset
- Python code – Jupyter Notebook files (which should include the results output in the Notebook)

* *The report is expected to have 10-15 pages, with 12 font size. 1.5 line spacing is recommended. Do not include a cover page.*

ADDITIONAL REMARKS

The team should submit electronic copies of all the deliverables on iCollege prior to the due date. Create a zip archive of all the deliverables and submit the archive and not individual files. Also, please only submit 1 copy per team. Late submission will NOT be accepted unless prior arrangements have been made.

The group project constitutes 30% of your final grade, of which the project proposal will constitute 3%, final presentation will constitute 13%, and the project report will constitute 14%.

The key evaluation criteria for the project are on the importance of the investigated problem, the innovativeness of the project idea, the completeness of the project, the efforts involved in collecting and cleaning the unstructured data, the scope, and reasonableness of the NLP techniques, both theoretically and practically.
The group project relies on the efforts, discussion and coordination of ALL group members. If you have questions about your own duties on the project, please first discuss with your teammates before consulting the instructor. When consulting the instructor, please ask group-level questions and include all your team members in the conversation (in the email list or in the zoom meeting).

To ensure fair consideration of individual members' true contributions to the project, a peer evaluation will be conducted at the end of the course. The average peer evaluation score for a given team member will be used as an adjustment factor to compute his/her individual project grade.