CIS 9665 Term Project Instructions

Deliverables:

Component	Points	Due Date
Project Proposal	5	Oct 28
Final Presentation	5	Dec 7
Peer Evaluation	5	Dec 11
Final Report	15	Dec 11

Term Project Guidelines

Students will demonstrate their proficiency of NLP techniques through their work on the term project by using real world data. Projects must be submitted according to the specifications provided by the instructor on the due date.

When forming your teams, keep the following in mind:

- Students will be assigned randomly or self-organized into teams, and each team should have a similar number of students. To ensure that each team is of similar sizes, I may ask that some students to switch from one team to another but I will try to minimize this.
- Each team should nominate one contact person for the instructor.
- Teams should be formed no later than Sep 07, 2022. At that time, the team leader should submit their team roster to the instructor.
- There will be penalties for changing teams after the roster is sent to the instructor (10% deduction for all group effort assignments for the student who changed teams).

Milestones of projects:

- To assist you choosing a feasible problem and correct analytical approach, each team will have a short presentation about the scope of project, dataset, and intended analytical approaches on Oct 26, 2022, in class. The team needs to present around 5 minutes. Teams need to send their presentation slides to me by **emails** a day before their presentations (before 11:59 PM Oct 25).
- I will give you some feedback during the presentation regarding the scope of project, dataset, and intended analytical approaches. You might also receive valuable feedback from other students. Based on these feedback, you should revise and submit a project proposal to **Blackboard** by Oct 28, 2022.
- The final projects will be presented to the class on Dec 7, 2022. Presentations are to be no more than 15 minutes in length, followed by a brief question & answer period. Teams need to upload their presentation slides to **Blackboard** a day before their presentations (before 11:59 PM Dec 6).
- The team needs to submit a final report for your project on or before Dec 11, 2022. The report must contain an summary of your findings, as well as all other supporting material (e.g. dataset and Python code). Further information will be provided in the project description section.
- Please be sure to check the date of the presentation prior to planning any trips during that time. The absence of one of the group members during the presentation time will result in 10% penalty for the entire group.

PLEASE NOTE: While generally all members of a team receive the same points for team assignments, we cannot allow non-performing students to free ride on the teammates. Therefore, there are peer evaluations at the end of the semester. Peer evaluations must be submitted by Dec 11, 2022. Each evaluation

is to be on a 0-5 points scale. If you give evaluation score lower than 4 points, please provide a detailed explanation. If there are consensus and proof that a team member is not contributing, his or her grades on the assignment WILL be reduced proportionally, and therefore be different from the rest of the team. I strongly encourage that you incorporate such stipulations, or specific penalties, into your team contract. For your team contract, you are free to use different ways to ensure that members contribute fairly and that the team operates effectively.

Project Descriptions:

The goal of the project is to develop the appropriate NLP methods for text analysis using Python. During this project, you can demonstrate how well you have acquired the course material. In addition, you can gain practical experiences to be prepared for the future career.

1. Project Proposal (5 points): Due Oct 28, 2022 11:59 pm via Blackboard

You are encouraged to find a **text** data set that you are interested in. You can use your own data if you already have an ongoing project. Or you can use existing corpra (e.g. Movie Reviews) (from nltk.corpus import movie_reviews). Or you can search one from online data repository, e.g. Kaggle [1] (e.g. https://www.kaggle.com/tags/nlp). For example, one of the Kaggle NLP datasets is called "Fake and real news dataset" (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset).

Identify interesting problem(s) that can be addressed by that **text** data set (e.g. the data set contains text columns such as news, reviews, tweets, comments, etc.). Turn in a project proposal including the following items:

- Who are team members?
- Which text dataset you intend to analyze? Where is your data source? Please describe your dataset.
- Research question (i.e. what is the question you intend to address?)
- Which methods do you intend to preprocess the text? (e.g. cleaning the text by getting rid of punctuations and numbers or stopwords, tokenization, normalization (e.g. stemming or lemmatization) etc.)
- Which methods do you intend to analyze the text (e.g. Text Classification, Topic Modeling, Sentiment Analysis, etc.)
- What are the practical implications by examining this research question?

You do not need to include Python code, figures or tables in the proposal. But you could include tables or figures in the proposal presentation. The proposal should be **1-2 page** long (11-point font and single-spaced) excluding figures, tables and references.

2. In-Class Final Presentation (5 points): In Class Wed Dec 7, 2022

3. Final Report (15 points): Due Sunday Dec 11, 2022 11:59 pm via Blackboard

Submission Requirements:

- One final report (excluding figures, tables and references) should be about **4-5 pages** with 11-point font and single-spaced.
- Do not show Python code in the report. The report only conveys your work and results.
- Cite any references used in your project in the reference section.

[1] https://www.kaggle.com/datasets

- Turn in Python code to Blackboard as a separate file. The report will not be graded without code.
- Submit the used data set to Blackboard. All results in the report should be able to be reproduced using Python code and data.

Final Term Project Report should contain the following items:

- Introduction and background
- Motivation of your research question (e.g. Why do you think it is an important question to solve?)
- Text dataset description
- Text preprocessing description
- Text analytic method description and results (e.g. why do you choose this text analytic approach? How to use this method? What are the results?)
- Model evaluation if supervised learning method is used (e.g. show Accuracy, Prevision, Recall). Show model tuning process if unsupervised learning method is used (e.g. adjusting the hyperparameters)
- Make conclusions following logically from results and findings
- Practical implications (e.g. how do your results apply to the real world? or how the company or society or customers can benefit from your results?)