IST 736 Text Mining
Course Project Instruction

1. Project Format

The objective of the project is to use the main skills taught in this class to solve a real text mining problem. Students should work individually or form a group of up to three students to finish the project.

2. Choose text mining problem and data set

For this project, you must choose your own dataset. It can be one that you created yourself or found from other resources.

Some rules/tips about choosing data sets:

a. Do not choose the data sets that we have analyzed in class, such as the Kaggle Sentiment data, movie review data, etc.
b. The data set should contain at least 100 examples.
c. Choose a data set that does not require excessive preprocessing.

3. Experiment design

Define a problem on the dataset as a classification and/or clustering problem, and describe it in terms of its real-world organizational or business application.

This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms or techniques, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if your problem is suitable, you may use more than one of the algorithms (Clustering + Classification, e.g.). Some explanation is needed to justify your choice of algorithms.

4. Project idea presentation

Post your project idea presentation to the Week 9 async forum. Describe the problem you are going to study, the data, and your initial strategies for data analysis. Your presentation should discuss the following information. Think of your presentation as a proposal that you will use to convince some funders to fund your project. DO NOT list each question and write answers to them.

(1) Description of the real world problem. Why does it matter?
(2) Problem modeling – if it is modeled as a classification problem, define the target categories; if a clustering problem, what types of clusters do you expect to get?
(3) What algorithm(s) or exploratory analysis methods do you plan to use? Why are they the best solution?
(4) How do you obtain the data? How long would it take? How many examples will you get? For classification problem, is the data set skewed or balanced?

(5) How are you going to evaluate your experiment result? Choose the evaluation method(s) and metric(s) and explain why they are the best choices.
(6) What challenge(s) do you foresee in this project?

Present your project idea in Week 9 Live Session. Presentation time will be evenly allocated to each team. The amount of time depends on the number of teams formed.

During the presentation we will discuss whether the problem modeling is valid, whether the project complexity is appropriate (if not, suggestion for adjustment), and whether the initial data analysis strategy is reasonable.

You are encouraged to read each other's presentation to learn from each other. It's OK to choose a problem that another student also chooses to work on, as long as your work is independent from each other. It's also OK to replicate famous experiments in published papers, and see whether you got the same result, and whether the paper provides necessary details for replicating the experiment.

5. Project result presentation

Week 10 Live Session will be the second round of presentation, focusing on the project results and challenges you are facing. The class will discuss and suggest solutions. Post your presentations to the Week 10 discussion forum before the live session.

6. Final project report

The final project report is due one week after the Week 10 Live Session, so that you will have time to incorporate the feedback that you receive.

Write the final report that conforms to general academic paper format. The grading rubrics will be similar to the previous rubric for homework assignments. (Pang, Lee, and Vaithyanathan, 2002) is a good writing example.

It is very important to cite and paraphrase relevant work appropriately.

Your report should be within 8 pages plus up to 2 more pages for references, 1 inch margin on all sides, and at least 12 point Arial or Times New Roman.

References:

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, 79-86. url: https://arxiv.org/pdf/cs/0205070.pdf