Project Information

Project Description

- The project will count towards 35% of your final grade.
- This is a group project. Each group should consist of 5 to 6 members. It is highly recommended that you meet with your group members early and get to know each others' background and strengths so as to work effectively as a group.
- The objectives of the project is to apply the techniques that you have learned in this module to a real data set to perform some useful analyses.
- You are free to choose either your own data set or one of the data sets we provide. The details of the data sets we provide will be released in a separate document.
- You are required to perform at least two analysis tasks on the data set you have chosen. A list of possible tasks is given below. You need to clearly state what you plan to do in the project proposal. The instructor will give you feedback and approve the proposal before you start the project.

List of Analytics Tasks

1. Document Retrieval: For this task, you should first define a few keyword-based queries. For example, if you have chosen the SGNews data set, then you can aim to find the most relevant news articles about "housing" or about "transportation" from the data set. With the defined queries, you then need to use Python together with additional libraries to find the most relevant documents to each query. (Optional: You may consider writing an interactive Python program such that the end user can enter a query with a few keywords and the program will display the most relevant documents.)
2. Document Classification: For this task, you need a data set that already has labelled documents. First, define the labels you want to use to classify documents. This could be topic-based class labels, or this could be something else such as spam/non-spam. Next, train a classifier using labelled training documents. Finally, test the quality of the trained classifier by using the classifier to predict labels of unseen documents. To achieve this, you can first hold out a subset of your documents for testing purpose and exclude them from your training documents. (Optional: You can consider writing an interactive program that predicts the label for any text entered by the user. The classifier needs to be trained before hand and saved on disk.)
3. Document Clustering: For this task, you can use any data set. The goal here is to characterize the document set by grouping documents into clusters. You can play with different numbers of clusters. After clustering, you need to describe the meaning of each cluster of documents with the help of Python and other libraries. For example, you can use the most common words of each cluster to help you understand the common theme of that cluster.
4. Topic Analysis: For this task, you can use any data set. The goal is similar to document clustering, but instead of placing documents into "hard" clusters, topic analysis methods such as LDA find topics, which are represented by distributions of words. These topics give a summary of the major themes covered by the document set from which topics are extracted.

5. Information Extraction: For this task, you want to extract named entities from the data set you have chosen. You can perform some further analysis on the extracted named entities such as finding the most frequent named entities or finding pairs of named entities that co-occur frequently.
6. Sentiment Analysis: For this task, your goal is to find positive and negative expressions from the data set you have chosen. Not every data set contains many sentiment expressions so chose this task only if it makes sense on your data set.
7. You can propose any other text mining or NLP tasks and the corresponding details. Examples: Question answering, summarization, discourse analysis for relationship extraction, semantic analysis in social media data etc.,

Deliverables

There are three major deliverables for the group project.

Proposal

Each group is required to submit a proposal (up to 3 pages) to indicate the data set to be used and the intended tasks and analyses. Use some use cases to illustrate why you believe the analytics tasks you intend to perform are useful. You can add unlimited number of pages for appendix.

Note: The proposal counts towards 5% of the final grade.

Presentation

Each group will give a 15-minute presentation in week 13. All members should be present at the presentation, but you can nominate two or three members to give the presentation. All members should be prepared to answer questions from the instructor.

The presentation should include a brief introduction of the data set and the motivation for the analytics tasks, solution approaches, some details of how the analyses are done, evaluations, in particular any challenges faced, and some example output from the analyses. You need to also specify your analysis on why somethings didn't work and what can be done to improve them. The main content will be in the report. Choose key examples for the presentation.

Final Report

You are also required to submit a final project report. We will post a template to e-learn. The report should be include the following sections:

- Project title
- Group information: group members and each member's email address.
- Background and motivation: What is the data set used? What tasks are performed? Why are these tasks useful?
- Methodology: How have you performed the tasks? For each task, which processing steps have you used? It might be helpful to draw flow charts to show the steps.

- Results/findings: What are the analysis results? This could include the classification accuracy, the summaries of document clusters, the proportions of positive/negative expressions and sample sentiment expressions etc.
- Discussions: What have you learned from this project? Are there limitations with the techniques you have used? How do you think the analyses could have been done better if you had access to more advanced techniques? Gap analysis for failures.

Note: The presentation and final report counts towards 25% of the final grade.

Peer Evaluation

Details about peer evaluation will be handed out later.

Note: Peer evaluation counts towards 5% of the final grade.

Datasets

1. **Kaggle**:

   Kaggle contains a large number text datasets.

   Some examples are wine reviews, employee reviews, movie plots etc.

   a. https://www.kaggle.com/datasets?tags=14104-Text+Data
   b. https://www.kaggle.com/datasets?tags=13204-NLP
2. **Spam Datasets:**
   a. https://www.cs.bgu.ac.il/~elhadad/nlp16/spam_classifier.html
3. Also find other datasets from web search:
   a. TripAdvisor Data Set
   b. Apple Discussion
   c. Amazon or movie revies

4 Timeline

Table 1 shows the timeline for the project.

Date	Activity / Deliverable
Session 5	**Project proposal** due (submit by end of day of class)
Session 6	Project proposal feedback
Session 10	**Final project report and presentation slides** due (submit before class)
Session 10	**Presentation** (in class)

Table 1: Timeline for the group project.