IS6751 Text and Web Mining

Wee Kim Wee School of Communication and Information
Semester 2, Academic Year 2024-2025

Lecturer:

Associate Professor Jin-Cheon Na
Office: 03-50, WKW School of Communication & Information
Email: tjcna@ntu.edu.sg

Course Description:

Nowadays, with the popularity of the Internet, there is a massive amount of text content available on the Web, and it becomes an important resource for mining useful knowledge. From a business and government point of view, there is an increasing need to interpret and act upon the large-volume text information. Therefore, text mining (or text analytics) is getting more attention to analyze text content on the Web. For instance, opinion mining and sentiment analysis is one of text mining techniques to analyze user-generated content on social media platforms.

This course covers how to analyse unstructured data (i.e., text content) using text mining techniques. Students will learn various text mining techniques and tools both through lectures and hands-on exercises in labs. The course will also explore various usages of text mining techniques to real world applications. This course focuses on Web content mining, but not on Web structure and usage mining.

Since deep learning has outperformed traditional machine learning techniques recently, this course will mainly use deep learning techniques for text mining and natural language processing.

Students will learn following topics in the course:

- Principles and concepts of text mining and natural language processing.
- Various text mining techniques: Pre-processing for Text Mining, Text Categorization, Sentiment Analysis, Question Answering, and Machine Translation.
- Practical use of text mining to real world applications, such as News classification, and Sentiment Analysis Systems analyzing public opinion towards various subjects, such as electronic gadgets, movies, stocks, etc., using social media content.

Course Objectives:

At the end of this course, students should be able to:

- Appreciate the basics of text mining and natural language processing.
- Understand the advantages and disadvantages of different text mining techniques.
- Work on practical problems that can be solved using text mining techniques.

Prerequisites:

A student should take this course only if

- The student has some aptitude for low-level logical thinking since lectures and labs will focus on technical aspects of Text and Web Mining.
- **The student has good computer programming skills since Python and PyTorch will be used for hands-on exercises in labs and assignments.**

If you are new to computer programming, especially Python, it is advisable to take foundational courses like IS6752 Data Extraction Techniques in an earlier semester and enroll in this course in a subsequent semester. Additionally, a basic understanding of data mining or machine learning would be helpful.

Method of Assessment:

- Lab Assignments (individual assignments): 10%
- Take Home Projects (individual assignments): 50%
- Group Project (group assignment): 30%
- Class Participation (class interactions and attendance): 10%

Lab Assignments (10%)

Due date: see the class schedule

These are **individual** lab assignments. For the lab assignments, you will be asked to submit lab reports after certain lab sessions. Note that lab assignment reports handed in after the due date/time will not be marked. These assignments will account for 10% of the overall grade.

Take Home Projects (50%)

Due date: see the class schedule

These are **individual** assignments: Take Home Project 1 (25%) and Take Home Project 2 (25%). For each Take Home Project, you will be asked to build deep learning models, such as sentiment classifier and surname classifier, with provided datasets and specified algorithms, such as Multi-Layer Neural Networks, RNN, CNN, and BERT. In general, you need to explore various text preprocessing techniques, hyperparameters, and any other relevant techniques to get the best accuracy. Note that for each Take Home Project, reports and required files/documents handed in after the due date will be marked down by **10%** per day.

Group Project (30%)

Project title selection due: see the class schedule
Project due date: see the class schedule

This is a group project (the size of each group will be announced later). All team members will receive the same grade. This project will account for 30% of the overall grade. **Please e-mail the lecturer your project title by the project title selection due date.** Include the names of team members in the message.

This is a text mining project where you collect your own sample text dataset or use an existing dataset, and using text mining techniques and tools, build an interesting model that mines knowledge/information from the text dataset. Generally, the project scope is entirely up to you, but I suggest that you build a useful and interesting model. Then, write a project report explaining your methodology and presenting the results. **Note that you must use PyTorch in your group project since PyTorch is covered in this course.**

You may conduct investigative analysis on any one of the following topics related to text mining:

- Text classification
- Opining mining and sentiment analysis
- Question Answering
- Machine Translation
- Information Extraction
- Other topics are possible with consent from the instructor.

The report should contain **up to 3,000 words**. The report **must** include a bibliography listing all references (including URLs, if any) cited, as well as a cover page where you should **indicate the total number of words excluding references**. The soft copy of the report should be submitted through **Turnitin** by the due date. Also submit a Flash Drive containing the report, presentation slides, dataset, and programming code files (such as Python files) used by the project to the instructor by the due date; or email the instructor a URL of any web drive where he can download the files. You are encouraged to produce high quality work as all the reports will be made available on the course web site for the benefit of the class.

For each project, each team will conduct a 15/20-minute presentation on their project work. This will be followed by a 5-minute question-and-answer session to allow for clarification by students and the lecturer. Schedules for the presentation will be announced later.

The assignment is due at the beginning of the class period on due date. Note that reports and required files/documents handed in after the due date will be marked down by **10% per day**.

Course Web Site:

Materials for the course will be accessible from the following URL:
http://ntulearn.ntu.edu.sg/

Language and Communication:

The language of instruction and communication is strictly English. Incomprehensible answers provided in the assignments will not be awarded any marks.

Textbooks:

- Natural Language Processing with Pytorch. (2019). Delip Rao and Brian McMahan. O'Reilly

References:

- Neural Network and Deep Learning, Andrew Ng, https://www.coursera.org/learn/neural-networks-deep-learning
- Natural Language Processing with Deep Learning, http://web.stanford.edu/class/cs224n/
- Deep Learning. (2016). Ian Goodfellow and Yoshua Bengio and Aaron Courville. MIT Press.
- Fundamentals of Predictive Text Mining. (2015). Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. Second Edition. Springer.

Academic Honesty & Plagiarism:

The work that you submit for assessment in this course must be your own individual work (or the work of your group members, in the case of group projects). The NTU Academic Integrity Policy (http://academicintegrity.ntu.edu.sg/) applies to this course. It is your responsibility to familiarise yourself with the Policy and to uphold the values of academic integrity in all academic undertakings. As a matriculated student, you are committed to uphold the NTU Honour Code (http://www.ntu.edu.sg/sao/Pages/HonourCode.aspx).

Acts of academic dishonesty include:

- *Plagiarism*: using or passing off as one's own, writings or ideas of someone else, without acknowledging or crediting the source. This includes
  - Using words, images, diagrams, graphs or ideas derived from books, journals, magazines, visual media, and the internet without proper acknowledgement;
  - Copying work from the Internet or other sources and presenting as one's own;
  - Direct quoting without quotation marks, even though the source is cited;
  - Submitting the same piece of work to different courses or to different publications.
- *Academic fraud*: cheating, lying and stealing. This includes:

- Cheating – bringing or having access to unauthorised books or materials during an examination or assessment;
- Collusion – copying the work of another student, having another person write one's assignments, or allowing another student to borrow one's work;
- Falsification of data – fabricating or altering data to mislead such as changing data to get better experiment results;
- False citation – citing a source that was never utilised or attributing work to a source from which the referenced material was not obtained.

- *Facilitating academic dishonesty*: allowing another student to copy an assignment that is supposed to be done individually, allowing another student to copy answers during an examination/assessment, and taking an examination/assessment or doing an assignment for another student.

Disciplinary actions against academic dishonesty range from a grade mark-down, failing a course to expulsion. Your work should not be copied without appropriate citation from any source, including the Internet. This policy applies to all work submitted, either through oral presentation, or written work, including outlines, briefings, group projects, self-evaluations, etc. You are encouraged to consult us if you have questions concerning the meaning of plagiarism or whether a particular use of sources constitutes plagiarism.

Schedule:

Week	Topic
1	Lecture: Class Information; Introduction to Text Mining Lab: PyTorch Basics (chapter1)
2	Lecture: Pre-processing for Text Mining and Traditional Machine Learning Lab: Pre-processing & Text Classification using Naïve Bayes Classifier
3	No Class: Chinese New Year
4	Lecture: Neural Networks: Logistic Regression Lab: Text Classification using Logistic Regression
5	Lecture: Neural Networks: Multi-layer Neural Networks Lab: Classifying Sentiment of Restaurant Reviews (chapter3) LabAssignment#1 (due on Wednesday, WK7)
6	Lecture: Improving Deep Neural Networks Lab: Surname Classification (chapter4)
7	Lecture: Word Embedding Lab: The CBOW Classifier Model (chapter5); Word Embeddings with Gensim TakeHomeProject#1 (due on Wednesday, WK9) **Group Project Title Selection Due (February 26, 2025)**
	Recess Week
8	Lecture: Sequence Models Lab: Classifying Surname Nationality Using a Character RNN (chapter6) LabAssignment#2 (due on Wednesday, WK11)
9	Lecture: Convolutional Neural Networks for NLP Lab: Surname Classification with a CNN (chapter4); The News Classifier Model (chapter5)
10	Lecture: Machine Translation, Seq2Seq and Attention Lab: Neural Machine Translation (chapter8)
11	Lecture: Transformers and Self-Attention Lab: Sentiment Classification (use Bert)
12	Lecture: Question Answering and Generative Language Models Lab: Question Answering (use Bert)
13	Student presentation (Group Assignment) **Group Project Submission Due (a soft copy of the report by April 13, 2025; the rest by April 16, 2025)** TakeHomeProject#2 (due on Wednesday, WK15)

The schedule may be subject to change depending on the pace of the course.