H6751 Text and Web Mining

Wee Kim Wee School of Communication and Information

Semester 2, Academic Year 2019-2020

Lecturer:

Dr. Zhao Rui, Dr. Chen Zhenghua
Email: rui.zhao@ntu.edu.sg, zhenghua.chen@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 is an introduction to text and web mining. It covers how to analyse unstructured data (i.e. text content) on the Web 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.

Students will learn following topics in the course:

1. Principles and concepts of text and web mining.
2. Various text mining techniques: Pre-processing for Text Mining, Text Categorization, Document Clustering, Information Extraction, and Opinion Mining & Sentiment Analysis.
3. Practical use of text mining to real world applications, such as Text Message Spam Detection, 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:

1. Appreciate the basics of text and web mining.
2. Understand the advantages and disadvantages of different text mining techniques.
3. 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.

Computer programming skill is required before you take this course. Python will be used for hands-on exercises in labs.

Method of Assessment :

Coursework (individual and group assignments): 45% (45 marks)
Class Participation (class interactions and attendance): 5% (5 marks)
Final examination: 50% (50 marks); **3 hours closed book exam**

The details could be found in the course website: https://h6751.github.io/

Course Web Site:

Materials for the course will be accessible from the following URL:
https://h6751.github.io/

Language and Communication :

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

Textbooks:

- 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).

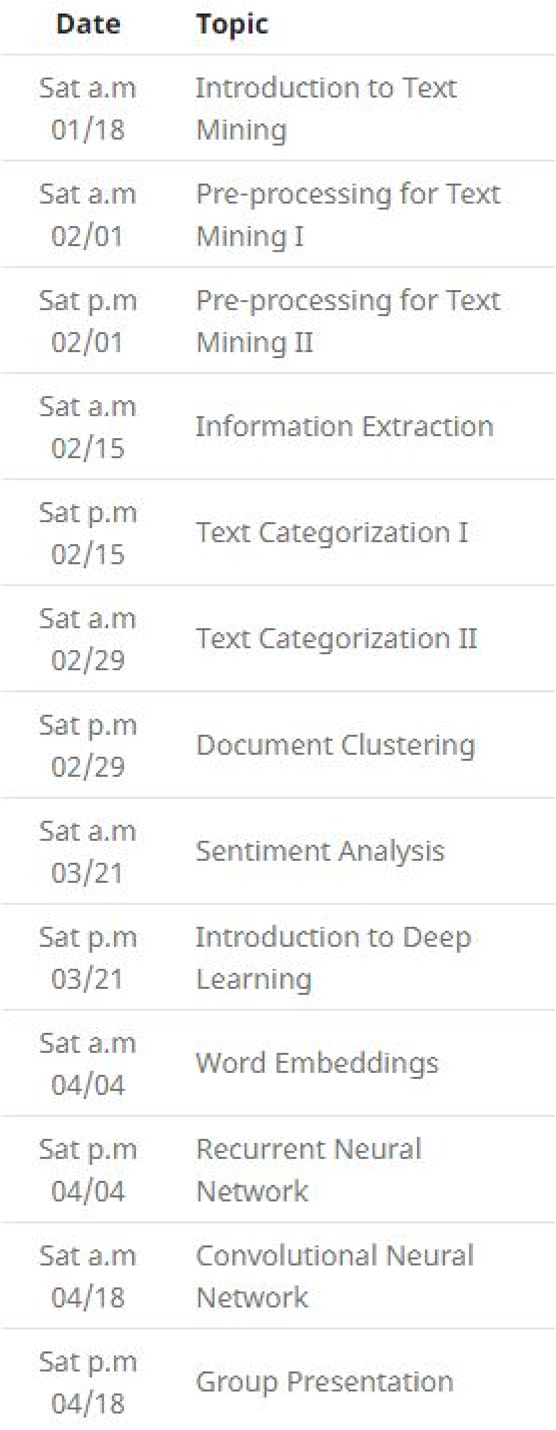
Acts of academic dishonesty include:

- 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:



Date	Topic
Sat a.m 01/18	Introduction to Text Mining
Sat a.m 02/01	Pre-processing for Text Mining I
Sat p.m 02/01	Pre-processing for Text Mining II
Sat a.m 02/15	Information Extraction
Sat p.m 02/15	Text Categorization I
Sat a.m 02/29	Text Categorization II
Sat p.m 02/29	Document Clustering
Sat a.m 03/21	Sentiment Analysis
Sat p.m 03/21	Introduction to Deep Learning
Sat a.m 04/04	Word Embeddings
Sat p.m 04/04	Recurrent Neural Network
Sat a.m 04/18	Convolutional Neural Network
Sat p.m 04/18	Group Presentation

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