INLS 613: TEXT MINING

SYLLABUS, SUMMER II 2017

School of Information and Library Science, University of North Carolina at Chapel Hill

---

Instructor: Heejun Kim Email: heejunk@email.unc.edu

Time: Mon through Thurs 3-5pm

Place: Manning 208

Credits: 3 hours

Office Hours: Tues 11am-12pm/ Thurs 10-11am (or by appointment) at Manning 216

Description

As of late, we are exposed to a large amount of textual information from a variety of sources such as social media and the Web. The changes in technology and practices of electrical publishing have changed the paradigm of producing and sharing information. The huge volume of information from "big text" has spurred the development of a new field called text mining in the search for new knowledge and useful pattern from the text. For instance, there are huge demands in business intelligence such as predicting consumers' response on a specific product and analyzing market status. Through text mining, we want to turn text data into actionable knowledge or insight which can help our decision making. We also want to minimize human efforts in consuming text data in some sense.

The course is divided into three modules: basics, principles, and applications (see the course schedule below). The third part of the course will focus on several applications of text mining: methods for automatically organizing textual documents for sense-making and navigation (clustering and classification), methods for detecting opinion and bias, methods for detecting and resolving specific entities in text (information extraction and resolution), and methods for learning new relations between entities (relation extraction). Students will develop a deep understanding of one particular application through a course project. Throughout the course, a strong emphasis will be placed on evaluation of the application.

Objectives

The main goal of this course is to help students to understand the power of a large amount of text data, to learn underlying theories and techniques of text mining, and to make practices with real-world data and issues. Students will also learn how to develop and evaluate computer programs that detect useful trends and patterns from unstructured natural language text in pursuit of discovering knowledge.

The hand-on group project will help students to derive knowledge and insight from real-world data sets by utilizing a variety of latest tools and statistical methods. Latest use cases of text mining in a deluge of "big text" from social media and the Web will be introduced over lectures to help students to understand what happens in real-world.

By successfully completing this course, advanced undergraduate students and master students are expected to be able to

- Understand basic concepts and theories of text mining
- Learn and utilize computational tools and methods for text mining
- Complete an application pipeline from data collection and cleaning to model evaluation
- Comprehend data analysis, machine learning and evaluation metrics
- Be prepared as a text data analyst for an industrial and academic position.

Students will be expected to learn actively with passion and diligence to achieve the above learning objectives. Please be aware that this class will require students to invest a significant of time to succeed.

Prerequisite

Students should have a reasonable background in programming in a structured or object-oriented programming language, such as Java or C++. "Reasonable" means either coursework or equivalent practical experience. You need to be able to design, implement, debug and test small to medium sized programs. The reasonable background in programming is needed to successfully complete the term project which is 40% of the grade. If you are not certain whether you are confident for taking this course, please send me an email.

Required Textbook

Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition) Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pall. 2016. Morgan Kaufman. ISBN 978-0-12-804291-5. Available online or in the campus bookstore. This book is referenced as WFHP below.

If you have the 3rd edition (ISBN 978-0-12-374856-0), it will also work. Two books (3rd edition) are reserved at the SILS library and free e-book version for UNC students will be available online.

Additional Resources

Foundations of Statistical Natural Language Processing. C. Manning and H Schutze. 1999.

Introduction to Information Retrieval. C. Manning, P. Raghavan and H. Schutze. 2008.

Course Policies

Laptops

Usage of laptop computers is not recommended during class. While laptops are convenient for note-taking, they are also a source of distraction. There have been several observations that students who constantly have their laptops open do poorly in the course. Please keep in mind that participation is 10% of your grade. If your laptop is open and your mind is elsewhere, it will show.

Attendance

Students are expected to attend every class. Missing up to two classes during the semester due to important reasons (e.g., travel, sickness) is fine. However, if you expect to miss class more than twice during the semester, please notify the instructor prior to missing class. Of course, attendance factors into your participation grade. If you must miss a class, make sure to get lecture notes from one of your peers. Class discussions are an excellent source of exam questions.

Participation

Class participation is a key to your success in this course. Students are expected to come to be engaged (e.g., complete required reading before the class) and to participate in class discussion by asking questions and expressing their opinion. Sharing your view with others is an important part of your education. It will broaden your understanding of the material and help you build confidence in this new area. "Quality is better than quantity." Class participation is part of your grade for this course.

Honor Code

You are encouraged to learn from each other. However, all the work you hand in must be your own. All assignments and exams are expected to be completed individually. This means that you cannot look at another student's answer and copy or re-word it as your own. Students are expected to adhere to the UNC Honor Code which is strongly effective across assignments, exams, and projects.

Assignment

Assignments are due by before midnight of the due date (11:55 pm, EST) unless otherwise specified. Assignments are to be submitted using Sakai unless instructed otherwise. Late assignments will be penalized 10% for each day late up to a maximum of three days. If you have special circumstances, contact the instructor as soon as possible. The exceptions can be considered on a case-by-case basis. When considered appropriate, limited extensions may be granted.

Grading

The grade for this course will be determined by a combination of four distinct factors: class participation, exam, homework, and the final project. The approximate contributions of these four factors to the grade are as follows:

- 10% - Class participation
- 20% - Midterm exam
- 30% - Homework (10% each)
- 40% - Final project (5% project proposal, 25% project report, 10% project presentation)

The final grades will follow the standard UNC grading system as outlined in the explanation of grading system by the Office of the University Registrar. The grading scale will be curved, with the highest grades reserved (as outlined by the Registrar office) for those with "the highest level of attainment that can be expected."

Schedule

The schedule might subject to change according to the upcoming circumstances.

Lecture	Date	Topic	Events	Reading Due
1	June 26	Introduction to Text Mining: The Big picture + Course Overview		
2	June 27	Predictive Analysis of Text: Concepts, Features, and Instances I		WFHP Ch. 1, Mitchell '06, Hearst '99
3	June 28	Predictive Analysis of Text: Concepts, Features, and Instances II + Text Representation I	Homework #1 Out	WFHP Ch. 2, Dominigos '12
4	June 29	Text Representation II		
5	Jul 3	Machine Learning Algorithms: Naïve Bayes I	Homework #1 Due	WFHP Ch. 4.2, Mitchell Sections 1 and 2
6	Jul 4	Fourth of July		
7	Jul 5	LighSIDE Tutorial	Homework #2 Out	LightSIDE User's Manual (on Sakai)
8	Jul 6	An Example of Text Mining Application: Predicting Usefulness of Yelp Review	Project Proposal due	WFHP Ch. 4.7
9	Jul 10	Machine Learning Algorithms: Instance-based Classification		WFHP Ch. 4.7
10	Jul 11	Predictive Analysis: Experimentation and Evaluation I	Homework #2 Due	WFHP Ch. 5
11	Jul 12	Weka Tutorial + Middle Term review		WFHP Appendix B
12	Jul 13	Middle Term		
13	Jul 17	Open review for a term project I	Homework #3 Out	
14	Jul 18	Predictive Analysis: Experimentation and Evaluation II		Smucker et al., '07, Cross-Validation, Parameter Tuning and Overfitting
15	Jul 19	Exploratory Analysis: Clustering I		Manning Ch. 16
16	Jul 20	Exploratory Analysis: Clustering II + Predictive Analysis with Noisy Labelers + Sentiment Analysis		Pang and Lee, '08 (skip Section 5 and skim Section 6), Pang and Lee, '02, Sheng et al., '08
17	Jul 24	Detecting Viewpoint and Perspective + Text-based Forecasting	Homework #3 Due	Yano et al., '10, Somasundaran et al., '10, O'connor et al., '10, Lerman et al., '08
18	Jul 25	Information Extraction and Relation Learning + Summary		McCallum '05, Arguello '07, WFHP CH 13.5
19	Jul 26	Student Presentations		
20	Jul 27	Student Presentations		

21	Aug 1		Project Report Due (by 6pm)	

Acknowledgement

This syllabus and course materials draw heavily on resources provided by Jaime Arguello. My deepest appreciation to him for his invaluable assistance in planning this course.