CSE398/498 Text Mining

Fall 2016

Instructor: Sihong Xie **Time:** TR 1:10-2:25pm **Email:** six316@lehigh.edu **Place:** PL 258.

Learning outcomes:

- Programming proficiency (to the extend of stand-alone ones).
- Basic linguistics for English.
- Text mining algorithms in real world applications;
- The ability to tackle more challenging NLP and text mining problems;
- Awareness of the values locked up in text data;
- Both computation- and data-driven thinking.

Communication:

- https://piazza.com/ for questions and answering about the course (like a forum).
- https://coursesite.lehigh.edu/ for posting grades and notifications.
- http://www.cse.lehigh.edu/~sxie/teaching/text_mining.html has the most up-to-date course information (lecture notes, codes, datasets, references).

Office Hours: R 4:30-6:30pm, PL 329

Required textbooks

- **FSNLP**: C. Manning and H. Schütze *Foundations of statistical natural language processing*, MIT Press, 2000.
- **IIR**: C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- **SAOM**: Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012.

Supplementary textbooks

- **SA**: Bing Liu, *Sentiment Analysis: mining opinions, sentiments, and emotions*, Cambridge University Press, 2015.
- **NLTK**: Bird, Steven, Edward Loper and Ewan Klein *Natural Language Processing with Python*, OReilly Media Inc, 2009.
- **PRML**: Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006.
- **ESL**: Trevor Hastie, Robert Tibshirani and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2013.

Prerequisites: Probability and statistics (MATH 231 or ECO 045) and programming experience (CSE 017).

Grading: Mid-term 25%, project 1 15%, project 2 50% (proposal 10%, presentation 10%, deliverables 30%), in-class quizzes 10%. See the course webpage for the most up-to-date exam time and deadlines.

Re-grading Requests must be made within 48 hours after the grades are released.

Accommodations for Students with Disabilities If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, Williams Hall, Suite 301 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

Principles of Our Equitable Community Lehigh University endorses The Principles of Our Equitable Community. We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.

Academic Integrity The work you submit must be entirely your own. While discussions of basic concepts covered in class with classmates are encouraged, plagiarism is never acceptable, and various methodswill be used to detect unreasonably similar copies of codes submitted. Such cases will be referred to the University Committee on Discipline and, if you are found guilty, you may be given the failing grade WF in the course. If you have questions about this policy at any point, ask me. It is far better to be safe than sorry when your academic career may be on the line.

Week	Date	Contents	Deadline
Week 1	8/30	Introduction to the course	
	9/1	Introduction to programming in Python, numpy, scipy and NTLK	
Week 2	09/06	Background probability and statistics	
	09/08	Background calculus and linear algebra	
Week 3	9/13	Text normalization, representations	
	9/15	Text classification, Naive Bayes, kNN, Logistic regression	
Week 4	9/20	Text clustering, k-means, hierarchical clustering	
	9/22	Topic model (LSI and SVD)	
Week 5	9/27	Word colocations	
	9/29	Foundation of NLP, syntatic parsing	
Week 6	10/04	N-gram model	
	10/06	Hidden Markov Model, POS tagging, Name entity recognition	
Week 7	10/11	Disambiguation, entity resolution	Project 2 proposal
	10/13	Neural network basics	
Week 8	10/18	**Pacing Break**	
	10/20	NN-based NLP, word embedding (Word2Vec, Glove)	Project 1 Due
Week 9	10/25	Introduction to sentiment analysis and opinion mining	
	10/27	Document level sentiment analysis	Project 2 milestone
Week 10	11/01	Sentence level sentiment analysis	
	11/03	Aspect-based sentiment analysis (aspect extraction)	
Week 11	11/08	Aspect-based sentiment analysis (aspect processing)	
	11/10	**Late mid-term**	
Week 12	11/15	Opinion lexicon generation	
	11/17	Opinion summarization	
Week 13	11/22	Relation detection	
	11/24	**Thanksgiving Break**	
Week 14	11/29	Crowdsourcing and active learning in text mining	
	12/01	Short text mining	Project 2 Due
Week 15	12/06	Trustworthiness issues, spam detection	
	12/08	Project Presentations	