CSE398/498 Natural Language Processing

Fall 2018

Instructor: Sihong Xie **Email:** six316@lehigh.edu

Time: Mon/Wed 12:45-2:00 pm **Place:** BC 115.

Catalog Description Overview of modern natural language processing techniques: text normalization, language model, part-of-speech tagging, hidden Markov model, syntatic and dependency parsing, semantics, word sense, reference resolution, dialog agent, machine translation. Three projects to design, implement and evaluate classic NLP algorithms. Credit will not be given for both CSE 398 and CSE 498.

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

Learning outcomes

- Probabilistic ways to deal with natural languages.
- Computation/data-driven thinking.
- Programming proficiency (Java at the level of CSE 017).

Communication

- https://piazza.com/ For questions answering and notifications.
- https://coursesite.lehigh.edu/ For posting grades only.
- http://www.cse.lehigh.edu/~sxie/teaching/2018_fall_nlp/nlp.html Course website has the most up-to-date information (lecture notes, codes, datasets, references).

Office Hours Mon/Wed 11:30-12:30pm, BC 326

Textbooks

- SLP = *Speech and Language Processing.* Dan Jurafsky and James H. Martin. 2nd and 3rd editions.
- FSNLP = *Foundations of statistical natural language processing.* C. Manning and H. Schütze. MIT Press, 2000.

Projects

We will provide datasets, code sketches, third-party packages for each project. Since the evaluation (running time, error rate) of your projects will be part of your final grades, we strictly require using Java and the given training and test datasets. Your programs will be tested on Sunlab machines. The three projects are:

- A simple spelling-checker using language models.
- A POS-tagger using HMM.
- Probabilistic syntactic parser.

Grading

For graduates Mid-term 15%, 3 coding projects (20% each), final exam (25%). Late submissions will be penalized 20% per late day (24 hours or part thereof) and no assignment will be accepted more than four days after its due date. The projects will graded partly based on your programs' performance in terms of metrics defined in individual projects. Requests for re-grading must be made within 48 hours after the grades are released.

For undergraduates Only projects 1 and 2 needs to be submitted (30% each). The mid-term and final will be the same but you will be required to answer some of the questions only. The other policies are the same as graduates. **Schedule**

Week	Date	Contents	Deadline	Required reading
Week 1	8/27	Intro to the course; counting words and n-grams		SLP 4.1 - 4.2
	8/29	Smoothing and backoff		SLP 4.3 - 4.6
Week 2	09/03	POS tagging; intro to HMM for POS tagging	Project 1 release	SLP 5.1 - 5.3, 5.5 - 5.6
	09/05	HMM (forward)		SLP 6.3 - 6.5
Week 3	9/10	HMM (forward)		SLP 6.3 - 6.5
	9/12	HMM (backward)		SLP 6.6 - 6.7
Week 4	9/17	HMM (backward)		SLP 6.6 - 6.7
	9/19	HMM (EM training)		SLP 6.8
Week 5	9/24	HMM (EM training)	Project 1 due	SLP 6.8
	9/26	HMM (EM training)	Project 2 release	SLP 6.8
Week 6	10/01	Intro to context-free grammar		SLP 12.2
	10/03	CFG (CKY)		SLP 13.4.1
Week 7	10/8	CFG (CKY)		SLP 13.4.1
	10/10	CFG (Earley)	Project 2 phase I due	SLP 13.4.2
Week 8	10/15	**Pacing Break**		
	10/17	**Mid-term exam**		
Week 9	10/22	PCFG	Project 2 phase II due	FSNLP 11.1 - 11.2
	10/24	PCFG (Inside algorithm)		FSNLP 11.3
Week 10	10/29	PCFG (Outside algorithm)		FSNLP 11.3
	10/31	PCFG (EM of Inside-Outside)		FSNLP 11.3
Week 11	11/05	PCFG (EM of Inside-Outside)		FSNLP 11.3
	11/7	Probabilistic lexicalized CFG	Project 3 release	SLP 14.6
Week 12	11/12	Lexicon semantics		SLP 19.1 - 19.2
	11/14	Word similarity (dictionary and corpus based)		SLP 20.6 - 20.7
Week 13	11/19	Word similarity (distributional)		SLP 20.6 - 20.7
	11/21	**Thanksgiving Break**		
Week 14	11/26	Word sense disambiguation (Lesk)		SLP 20.1 - 20.5
	11/28	Word sense disambiguation	Project 3 due	SLP 20.1 - 20.5
Week 15	12/03	Information retrieval (Brian Davison)		
	12/05	Machine Translation		
Week 16	12/10	**Final exam**		All your notes taken
	12/12		Project 2 final due	

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.