

# CSE398/498 NATURAL LANGUAGE PROCESSING

Fall 2017

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

**Catalog Description** Overview of modern natural language processing techniques: text normalization, language model, part-of-speech tagging, hidden Markov model, syntactic and dependency parsing, semantics, word sense, reference resolution, dialog agent, machine translation. Four 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/2017\\_fall\\_nlp/nlp.html](http://www.cse.lehigh.edu/~sxie/teaching/2017_fall_nlp/nlp.html)   Course website has the most up-to-date information (lecture notes, codes, datasets, references).

**Office Hours** Thursday 4:30-6:30pm, PL 329

## 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 4 projects are:

- A simple spelling-checker using language models.

- A POS-tagger using HMM.
- TBD (Probabilistic syntactic parser, or word collocation and phrase extraction).
- TBD (Chatbot or machine translation).

### Grading

Mid-term 15%, 4 coding projects (15% 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 be 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.

### Schedule

Week	Date	Contents	Deadline	Required reading
Week 1	8/29	Intro to the course; counting words and n-grams		SLP 6.1 - 6.2
	8/31	Smoothing and backoff	Project 1 release	SLP 6.3 - 6.6
Week 2	09/05	POS tagging and rule-based tagging		SLP 8.1 - 8.4, 8.6
	09/07	Intro to HMM		SLP 7.2 - 7.3
Week 3	9/12	HMM inference (Viterbi)		SLP 7.3
	9/14	HMM learning; HMM for POS tagging		SLP 8.5
Week 4	9/19	Intro to context-free grammar		SLP 9.1 - 9.6
	9/21	Intro to context-free grammar	Project 1 due	SLP 9.7 - 9.12
Week 5	9/26	Search-based parsing	Project 2 release	SLP 10.1 - 10.3
	9/28	Earley algorithm and finite-state parsing		SLP 10.4 - 10.5
Week 6	10/03	Probabilistic syntax parsing		SLP 12.1 - 12.2
	10/05	Probabilistic lexicalized CFG; dependency parsing		SLP 12.3 - 12.4
Week 7	10/10	First order predicate calculus		SLP 14.1 - 14.3
	10/12	Other semantic concepts	Project 2 due	SLP 14.4 - 14.6
Week 8	10/17	<b>Pacing Break</b>	Project 3 release	
	10/19	Mid-term exam		
Week 9	10/24	Syntax driven semantic analysis		SLP 15.1 - 15.2
	10/26	Earley parser with semantics		SLP 15.3 - 15.5
Week 10	10/31	Lexical semantics and WordNet		SLP 16.1 - 16.3
	11/02	Intro to discourse	Project 3 due	SLP 18.1
Week 11	11/07	Text coherence and discourse structure	Project 4 release	SLP 18.2 - 18.3
	11/9	Intro to dialogue and conversational agents		SLP 19.1 - 19.2
Week 12	11/14	dialogue Interpretation, structure and coherence		SLP 19.3 - 19.5
	11/16	Intro to text generation		SLP 20.1 - 20.2
Week 13	11/21	Surface realization		SLP 20.3
	11/23	<b>Thanksgiving Break</b>		
Week 14	11/28	Discourse planning		SLP 20.4
	11/30	Intro to Machine translation		SLP 21.1 - 21.2
Week 15	12/05	Interlingua; direct translation		SLP 21.3 - 21.4
	12/07	Statistical machine translation	Project 4 due	SLP 21.5
Week 16	12/13	Final exam		

**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 methods will 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.