CS 6120/CS 4120: Natural Language Processing

Professor Lu Wang

Spring 2018

E-mail: luwang@ccs.neu.edu
Office Hours: T 5:15pm-6:15pm
Office: 258 WVH

Web: http://www.ccs.neu.edu/home/luwang/
Class Hours: TF 3:25pm - 5:05pm
Class Room: West Village H 108

Course Description

This course aims to introduce fundamental tasks in natural language processing, and its recent advances based on machine learning algorithms (e.g., neural networks) and applications for interdisciplinary subjects (e.g., computational social science). The course materials are mostly delivered as lectures, and accompanied with reading materials. The students will be evaluated based on assignments, a research-driven course project, and an open-book final exam.

Required Materials

- Main textbooks
  - Dan Jurafsky and James H. Martin, "Speech and Language Processing, 2nd Edition", Prentice Hall, 2009.
  - Third edition draft is available at web.stanford.edu/~jurafsky/slp3/.
- Other Reference
  - Chris Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999
- Since many natural language processing problems are driven by machine learning techniques nowadays, we also highly encourage you to read machine learning textbooks:
  - Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.
  - Tom Mitchell, "Machine Learning", McGraw Hill, 1997.

Prerequisites/Corequisites

This course is designed for graduate students and senior undergraduate students majoring in computer science, linguistics, and other related areas. Students who take this course are expected to be able to write code in some programming languages (e.g., Python, Java, or C/C++) proficiently, and finish courses in algorithms (CS 5800 or CS 7800), multivariable calculus, probability, and statistics. Linear algebra is optional, but highly recommended.

Course Structure

Grading Policy

Each assignment or report, both electronic copy and hard copy, is due at the beginning of class on the corresponding due date. Blackboard is used for electronic submission. Hard copies are submitted in class. Assignment or report turned in late will be charged 20 points (out of 100 points) off for each late day (i.e. 24 hours). Each student has a budget of 5 days throughout the semester before a late penalty is applied. You may want to use it wisely, e.g. save for emergencies. Each 24 hours or part thereof that a submission is late uses up one full late day. Late days are not applicable to final presentation. Each group member is charged with the same number of late days, if any, for their submission.

Grades will be determined based on three assignments, ten in-class tests, one course project, one open-book exam, and participation:

- **30%** of your grade will be determined by two assignments, each of 15%.
- **5%** of your grade will be determined by eight quick in-class tests, each of 1%;three with lowest grades will be dropped, and no make-up.
- **25%** of your grade will be determined by your course project, including proposal (3%), progress report (7%), final report (10%), and presentation (5%) with 1% as bonus if selected as best project by peer student.
- **35%** of your grade will be determined by final exam.
- **5%** of your grade will be determined by your participation in the class.

Course Policies

Academic Integrity and Honesty

This course follows the Northeastern University Academic Integrity Policy. All students in this course are expected to abide by the Academic Integrity Policy. Any work submitted by a student in this course for academic credit should be the student's own work. Collaborations are allowed only if explicitly permitted. Violations of the rules (e.g. cheating, fabrication, plagiarism) will be reported.

Schedule and Weekly Learning Goals

The schedule is tentative and subject to change. The learning goals below should be viewed as the key concepts you should grasp after each week.

Week 01, 01/09 - 01/13: Topic: Introduction, Language Models

- Reading: Ch1, Ch4.1-4.9
- TODO: start thinking about projects and looking for teammates

Week 02, 01/16 - 01/20: Topic: LM cont'd, Text Categorization and Evaluation, Naive Bayes, Part-of-Speech Tagging, Sequence Labeling, Hidden Markov Models

- Reading: Ch5, Ch6.1-6.5, and http://web.stanford.edu/~jurafsky/slp3/6.pdf from 3rd edition

Week 03, 01/23 - 01/27: Topic: HMM cont'd, Word Sense Disambiguation

- Reading: Ch19-20, and http://web.stanford.edu/~jurafsky/slp3/17.pdf

Week 04, 01/30 - 02/03: Topic: Machine Learning Basics (Maximum Entropy, Feedforward Neural Networks), Formal Grammars of English, Syntactic Parsing

- Reading: Ch6.6, Ch12.1-12.5
- TODO: Course project proposal due

Week 05, 02/06 - 02/10: Topic: Parsing cont'd, Dependency Parsing

- Reading: Ch12.1-12.5, Ch13.1-13.4.1
- TODO: Assignment 1 is due

Week 06, 02/13 - 02/17: Topic: Vector-Space Lexical Semantics, Semantic Parsing, Combining Logical and Distributional Semantics

- Reading: Ch20.7, and http://web.stanford.edu/~jurafsky/slp3/15.pdf, http://web.stanford.edu/~jurafsky/slp3/16.pdf

Week 07, 02/20 - 02/24: Topic: Information Extraction

- Reading: Ch22.1-22.2

Week 08, 02/27 - 03/03: Topic: Question Answering

- Reading: Ch23

Week 09, 03/06 - 03/10: Topic: Spring Break!

Week 10, 03/13 - 03/17: Topic: Text Summarization, Sentiment Analysis, Recurrent Neural Networks

- Reading: http://web.stanford.edu/~jurafsky/slp3/6.pdf, http://web.stanford.edu/~jurafsky/slp3/18.pdf
- TODO: Assignment 2 is due

Week 11, 03/20 - 03/24: Topic: Sentiment Analysis, Opinion Mining, NLP and Social Media

- Reading: http://web.stanford.edu/~jurafsky/slp3/6.pdf, http://web.stanford.edu/~jurafsky/slp3/18.pdf
- TODO: Coure project progress report due

Week 12, 03/27 - 03/31: Topic: Dialog Systems and Chatbots

- Reading: Ch24, or http://web.stanford.edu/~jurafsky/slp3/29.pdf

Week 13, 04/03 - 04/07: Topic: Course Project Presentation

Week 14, 04/10 - 04/14: Topic: Course Project Presentation

Week 15, 04/17 - 04/21: Topic: Final Exam (Time and Location TBD)