CS 6120/CS 4120: Natural Language Processing

Professor Lu Wang

Fall 2019

E-mail: luwang@ccs.neu.edu
Office Hours: M 1:30pm-2:30pm (or by appointment)
Office: 911, 177 Huntington Ave.
Web: http://www.ccs.neu.edu/home/luwang/
Class Hours: MW 2:50pm -4:30pm
Class Room: Snell Library 035

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.

Course and Discussion Websites

Course Website: http://www.ccs.neu.edu/home/luwang/courses/cs6120_fa2019/cs6120_fa2019.html
Discussion Forum: http://piazza.com/class/fall2019/cs6120/home

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. It would be beneficial if the students have prior knowledge on supervised machine learning.

Course Structure

Grading Policy

Each assignment or report is due by the end of day on the corresponding due date (i.e. 11:59pm, EST). Blackboard is used for electronic submission. Assignment or report turned in late will be charged 20 points (out of 100 points) off for each late day (i.e. every 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 two assignments, eight in-class quizzes, one course project, one open-book prelim, 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.
- **12%** of your grade will be determined by prelim.
- **25%** of your grade will be determined by final exam.
- **3%** 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. Please check the course website for most recent schedule. The learning goals below should be viewed as the key concepts you should grasp after each week.

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

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

Week 02, 09/16 - 09/20: Topic: Text Categorization and Evaluation, Naive Bayes, Part-of-Speech

- Reading: Ch5.1-5.5, Ch6.1-6.5

Week 03, 09/23 - 09/27: Topic: HMM, Word Sense Disambiguation

- Reading: Ch20

Week 04, 09/30 - 10/04: Topic: Machine Learning Basics (Maximum Entropy, Feedforward Neural Networks), Formal Grammars of English, Syntactic Parsing

- Reading: Ch6.6, Ch12.1-12.5
- Course project proposal due on Sep 30.

Week 05, 10/07 - 10/11: Topic: Parsing cont'd, Dependency Parsing, Semantics

- Reading: Ch12.1-12.5, Ch13.1-13.4.1
- Assignment 1 is due on Oct 8.

Week 06, 10/14 - 10/18: Topic: Vector-Space Lexical Semantics

- NO CLASS is planned on 10/14
- Reading: Ch20.7

Week 07, 10/21 - 10/25: Topic: Distributional Semantics

Week 08, 10/28 - 11/01: Topic: Neural Language Models, Sentiment Analysis

- Reading: Ch23

Week 09, 11/04 - 11/08: Topic: Prelim

- Prelim on Nov 6 (same time and location as the regular lectures), no lecture on Nov 4 (prelim preparation and extra office hour)

Week 10, 11/11 - 11/15: Topic: Information Extraction

- NO CLASS is planned on 11/11
- Reading: Ch22.1-22.2
- Assignment 2 is due on Nov 12.

Week 11, 11/18 - 11/22: Topic: Question Answering, project feedback

- Reading: Ch23.1-23.2
- Course project progress report due on Nov 18.

Week 12, 11/25 - 11/29: Topic: Machine Translation, Discourse Analysis

- NO CLASS is planned on 11/27
- Reading: Ch21, Ch24, Ch25.1-25.5

Week 13, 12/02 - 12/06: Topic: Project Presentation

- Project final report due on December 7.

Week 14, 12/09 - 12/13: Topic: Final Exam

- Pending confirmation.