COMP 4480/ 6480: Introduction to Natural Language Processing

Basic Information

Time & Place: Every Monday and Wednesday, 12:40 pm - 02:05 pm
Instructional mode: In-person, FedEx Institute of Technology 227
Instructor: Xiaolei Huang
Office Hours: Monday 4 pm – 5 pm, Remote & Zoom
https://memphis.zoom.us/j/86277366545?pwd=bHQ1TDN5T3VGbWdsVFBmNGp6V3ordz09&from=addon

Course Description

This course will cover fundamental concepts and techniques of statistical machine learning approaches to natural language processing. The course starts with primary concepts and methods for processing human language. Topics include necessary concepts of probability and statistics, language and classification model, syntax, parsing and semantics. Natural language processing is an interdisciplinary field that have many applications in the other fields, such as computational social science, psychology, health, cognitive science, etc. We will cover applications of neural models in several major applications of NLP techniques. Topics include topic models, information extraction, question answering, dialog and fairness.

This course will focus on hands-on assignments and projects. The final project will have three stages, initial team proposal, one-page midterm report and a 4-page project report. Students will also complete one in-class (or take-home) midterm exams (~45 minutes) that will test high-level understanding of concepts. Students in the 6992 session will finish additional challenges in homework / quizzes / exams. Specifically, each graduate student will finish one additional question in assignments, two additional questions in each exam, and 8-page final project report. And students in the 4992 session can finish the additional challenges for extra credits.

Prerequisite

COMP 2150 / COMP 4001

We will work extensively on the Python programming language. **It is assumed that you know how to program in Python and use Unix-like operating systems (Linux, OS X).**

Textbook (FREE)

- SLP: Speech and Language Processing, 3rd edition:
    - https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Schedule (subject to change)

Week 1: Introduction + Statistical Language Model-1

- Date: 08-28 & 08-30
- Reading:
    - SLP, Chapter 1 & 2.
- Homework 0:
    - Set up Python development environment.
    - Basic probability warmup

Week 2: Statistical Language Model-2 & Naïve Bayes Classification

- Date: 09-06
- Reading:
    - SLP, Chapter 3.
    - SLP, Chapter 4.
- Homework 1: start at 09-06

Week 3: Logistic Regression & Optimization

- Date: 09-11 & 09-13
- Reading:
    - SLP, Chapter 5.
- Homework 1: **due** at 09-17

Week 4: Feature Engineering & Classification Practice

- Date: 09-18 & 09-20
- Reading:
    - SLP, Chapter 8.1 - 8.3
- Homework 2: start at 09-18

Week 5: POS Tagging, HMMs, Viterbi. Lexical Semantics

- Date: 09-25 & 09-27
- Reading:
    - SLP, Chapter 8.4 – 8.8
    - SLP, Chapter 6.1 – 6.6.
    - https://web.stanford.edu/~jurafsky/slp3/A.pdf
- Homework 2: **due** at 10-01

Week 6: Deep Neural Networks & Word Embedding Intro

- Date: 10-02 & 10-04
- Reading:
    - PyTorch tutorials: https://pytorch.org/tutorials/
    - SLP, Chapter 6.8 – 6. 12

    - Distributed Representations of Words and Phrases and their Compositionality, Mikolov et al., NeuRIPS 2013
- Homework 3 starts at 10-04

Week 7: Word Embeddings & Final Project

- Date: 10-09 & 10-11
- Reading:
    - SLP, Chapter 7
- Homework 3 **due** at 10-15

Week 8: Embedding Practice & Review

- Date: 10-18
- Reading: ~No
- Midterm (take-home) Exam **due**: 10-22

Week 9: Recurrent Neural Networks

- Date: 10-23 & 10-25
- Reading:
    - SLP, Chapter 9
    - Sequence to Sequence Learning with Neural Networks, Sutskever et al., NeuRIPS 2014
- Project proposal **due** at 10~30

Week 10: Transfer Learning & Transformers

- Date: 10-30 & 11-01
- Reading:
    - SLP, Chapter 10
    - Attention Is All You Need, Vaswani et al., NeuRIPS 2017
- Homework 4: start at 11-01

Week 11: BART/T5, GPT-3 & Prompting Engineering

- Date: 11-06 & 11-08
- Reading:
    - SLP, Chapter 11
    - Scaling Instruction-Finetuned Language Models, Chung et al., 2022
    - PaLM: Scaling Language Modeling with Pathways, Chowdhery et al., 2022
- Homework 4: **due** at 11-12

Week 12: LLM Practice & Information Extraction

- Date: 11-13 & 11-15
- Reading:
    - SLP, Chapter 11
- Midterm Report: **due** at 11-19

Week 13: Question Answering and Dialogs

- Date: 11-20
- Reading:
    - SLP, Chapter 14
    - SLP, Chapter 15

Week 14: Machine Translation and NLP ethics

- Date: 11-27 & 11-29
- Reading:
    - SLP, Chapter 13
    - MT:
        - Papineni et al. (2002): https://www.aclweb.org/anthology/P02-1040.pdf
        - Collins' Notes on the IBM Models:
            - http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/ibm12.pdf
    - Ethics:
        - Hovy and Spruit (2016): https://www.aclweb.org/anthology/P16-2096.pdf
        - Shah et al. (2019): https://arxiv.org/pdf/1912.11078.pdf

Week 15: Final Presentation

- Date: 12-04 and 12-06
- Presentation submission **due**: 12-06

Week 16: Project Report

- Final Project Report **due**: 12-011

Evaluation

Grades will be computed as follows:

	6992	4992
Participation	20%	20%
Homework	40%	40% + 5% (extra credit available)
Exams	10%	10% + 5% (extra credit available)
Final Project	30%	30%
Total	100%	100% + 10% (extra credit available)

97+	A+
[93 - 97)	A
[90 - 93)	A-
[87 - 90)	B+
[83 - 87)	B
[80 - 83)	B-
[77 - 80)	C+
[73 - 77)	C
[70 - 73)	C-
[60 - 70)	D
< 60	F

"[" refers to "include", and ")" means "exclude".

It is possible to earn extra credits by going above and beyond the expectations of the assignments and exams for 4992. 6000-level students are expected to have extra challenges in the homework and exams.

Late Policy. There will be in total four no-penalty late days for students to freely distribute to the four homework assignments. The number of used late days should be clearly stated in the submitted homework. All late submissions are subject to a penalty of 20% per day for no more than three days. Late 1 day: 80%; Late 2 days: 60%; Late 3 days: 40%; Late > 3 days: 0%.

Exam. There will be take-home midterm and final exams. The midterm will cover material in the previous lectures, and you will be allowed to use one page of note.

Homework and Final Project Policy. Assignments and the final report are due at 11:59 pm Central Standard Time on Friday indicated on the schedule. Students are required to submit the **PDF** file of their project reports. Submissions of homework depend on homework requirements. A word processing software (e.g., LaTeX or Word) is recommended. You are allowed to collaborate with other peers, but copying and pasting from another student will be considered plagiarism. The final project will be a group research project on a topic of students' choices after consulting with the instructor. The final project will base on a short presentation and technical report (4 or more pages). The final project report will be due at the end of the final exam period. More information will be posted on a separated page for the final project.

Plagiarism

Plagiarism or cheating behavior in any form is unethical and detrimental to proper education and will not be tolerated. All work submitted by a student (projects, programming assignments, lab assignments, quizzes, tests, etc.) is expected to be a student's own work. The plagiarism is incurred when any part of anybody else's work is passed as your own (no proper credit is listed to the sources in your own work) so the reader is led to believe it is therefore your own effort. Students

are allowed and encouraged to discuss with each other and look up resources in the literature (including the internet) on their assignments, but appropriate references must be included for the materials consulted, and appropriate citations made when the material is taken verbatim.

If plagiarism or cheating occurs, the student will receive a failing grade on the assignment and (at the instructor's discretion) a failing grade in the course. The course instructor may also decide to forward the incident to the Office of Student Conduct for further disciplinary action. For further information on U of M code of student conduct and academic discipline procedures, please refer to https://www.memphis.edu/osa/students/academic-misconduct.php.

"Your written work may be submitted to Turnitin.com, or a similar electronic detection method, for an evaluation of the originality of your ideas and proper use and attribution of sources. As part of this process, you may be required to submit electronic as well as hard copies of your written work, or be given other instructions to follow. By taking this course, you agree that all assignments may undergo this review process and that the assignment may be included as a source document in Turnitin.com's restricted access database solely for the purpose of detecting plagiarism in such documents. Any assignment not submitted according to the procedures given by the instructor may be penalized or may not be accepted at all." (Office of Legal Counsel, October 17, 2005).

Accommodations

Any student who anticipates physical or academic barriers based on the impact of a disability is encouraged to speak with me privately. Students with disabilities should also contact Disability Resources for Students (DRS) at 110 Wilder Tower (901-678-2880). DRS coordinates access and accommodations for students with disabilities (http://www.memphis.edu/drs/).

If you are sick, in particular with an illness that may be contagious, notify me by email but do not come to class. If you are struggling with anxiety, stress or other mental health related concerns, please consider visiting the Counseling Center or calling 901.504.6442 or 901.468.3633.