SI 630 Natural Language Processing: Algorithms and People

Winter 2020

Instructor: David Jurgens jurgens@umich.edu
Office hours: Tuesday 1-2pm via appointment @ NQ 3341

GSIs:
- Yulin Yu yulinyu@umich.edu
  - Office Hours: Friday 3:15-5:15 pm @ NQ1286
- Jiaqi Ma jiaqima@umich.edu
  - Office Hours: Tuesday 9:15-11:15 am @NQ1282

Natural Language Processing (NLP) is the study of the computational treatment of natural language--the words we use everyday. NLP draws on research in Linguistics, Theoretical Computer Science, Mathematics and Statistics, Psychology, Artificial Intelligence, Machine Learning, with broad applications to many other fields. This course introduces students to a variety of NLP methods available for reasoning about text in computational systems. We will focus on major algorithms used in NLP for various applications (e.g., part-of-speech tagging, parsing, machine translation), on the linguistic phenomena those algorithms attempt to model, and on the people who interpret and utter the language. Students will implement a variety of algorithms for different linguistic aspects (e.g., syntax, semantics) and also understand the creation of ground truth data through linguistically annotating data on which those algorithms depend.

Broadly speaking, the course has six major parts which are interleaved through most discussions:

1. Linguistic, mathematical, and computational background
2. Computational models of morphology, syntax, semantics, discourse, pragmatics
3. Core NLP technology: parsing, part of speech tagging, text generation, semantic analysis, etc.
4. Applications: text classification, sentiment analysis, text summarization, question answering, machine translation, information extraction, etc.
5. Human Factors in language understanding and generation, including sociolinguistics, style, and data annotation
6. Neural Networks and Deep Learning: autoencoders, recurrent NNs, LSTM, Transformers, GANs, etc.

The NLP course serves three major goals:

1. Learn the basic principles, algorithms, and theoretical issues underlying natural language processing
2. Learn computational techniques and tools used to develop practical, robust systems that can communicate with users in one or more
3. Gain insight into many open research problems in natural language

TEXTBOOKS AND OTHER MATERIALS:

The syllabus includes five textbooks, most of which are available online for free through the university. The inclusion of multiple textbooks is intended to allow students to have a complementary second (or third) description of an algorithm or topic. In general, we will use Dan Jurafsky and James Martin, Speech and Language Processing (3nd ed. draft) for most class activities and refer to the other materials only for specific topics; however students are encouraged to read other sources based on their interests.

- [SLP3] Dan Jurafsky and James Martin, Speech and Language Processing (3nd ed. draft) [Available here].
- [SLP2] Dan Jurafsky and James Martin, Speech and Language Processing (2nd ed., 2009)
- [PS] James Pustejovsky and Amber Stubbs, Natural Language Annotation for Machine Learning (2012) [Available for free on campus/VPN here]
- [BGHM] Jordan Boyd-Graber, Yuening Hu, and David Mimno, Applications of Topic Models (2017). [Available by courtesy of authors here].
- [MS] Foundations of Statistical Natural Language Processing (Chris Manning and Hinrich Schütze) ISBB: 0262133601 https://nlp.stanford.edu/fsnlp/ [Available for free on campus/VPN here]

In addition to the textbooks, each week will have an associated list of recent research papers, which are used five times during the semester for reading responses that are submitted as homework

GRADING:

42% - Homework Assignments

- 2% Intro assignment (Regular expressions)
- 8% Text Classification
- 8% Word Vectors
- 8% Parsing
- 8% Topic Modeling
- 8% Deep Learning

20% - Exams

- 20% Midterm

33% - Course project
- 4% Project Proposal
- 4% Halfway Update
- 5% Poster presentation at UMSI Expo
- 20% Final report

5% - Readings
- 1% for each weekly paper readings (must submit 5; can't submit more than 5)

Grading follows the traditional scale.
97-100: 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-
67-70: D+
63-67: D
60-63: D-
Below 60: F

There is no curve for this class. That said, usually the majority of students have received As. If you are concerned about your grade, please come see the instructors early (not the last week) while there is still time to improve your performance. No extra credit or make-up assignments are allowed (e.g., no retroactive work after the end of the semester to bump up grade).

CLASS COMMUNICATION AND PIAZZA POLICY:

All official course announcements will be sent through Canvas. Should any scheduling change (e.g., setting the date for the midterm), you will be responsible for monitoring this email. We will also announce the same things at the beginning of class. Please be on time to hear course announcements.

All other questions will use Piazza, accessed through Canvas. Questions will be replied to as quickly as possible, though we only guarantee a response time of 24 hours Monday-Friday and 48 hours on weekends. No questions should be submitted by email, unless it is concerning a personal matter.

Code should be kept to a minimum for questions on Piazza, or (preferably) avoided entirely. You are encouraged to discuss things at a high level or seek in-person guidance for code debugging. For

questions that require more than 10 sentences to explain, we encourage you to ask in the office hours.

SYLLABUS:

Each week will cover a set of related topics and have associated readings from different sources. The following is a week-by-week guide and are subject to change based on course scheduling. Any updates will be posted to Canvas.

Week 1: January 8th

- Topics: Welcome to NLP
    - Introduction
    - Regular Expressions
    - Text Classification (Part 1): Naive Bayes
- Readings:
    - SLP2 Chapter 1
    - SLP3 Section 2.1, SLP3 Chapter 6
    - MS Chapter 2 and 3

Week 2: January 15th

- Topics: Supervised Text Classification
    - Text Classification (Part 2)
    - Logistic Regression
    - Neural Networks
- Readings:
    - SLP3 Chapters 7
    - MS Chapter 16
    - (Optional) SLP3 Chapter 18
    - (Optional) G Chapter 2 (very math intensive)

Week 3: January 22th

- Topics: Language Models
    - Language Models
    - Recurrent Neural Networks
- Readings:
    - SLP3 Chapters 4, 8
    - MS Chapter 6

    - G Chapter 9, 14G

Week 4: January 29th

- Topics: Word Vectors
    - Latent Semantic Analysis
    - Word Vectors (word2vec + GloVe)
    - Morphology
- Readings:
    - SLP3 Chapters 15, 16
    - G Chapter 10

Week 5: February 5th

- Topics: Sequence Labeling
    - Hidden Markov Models
    - Part of Speech Tagging
    - LSTM Networks
- Readings:
    - SLP3 Chapters 9, 10
    - G Chapter 15
    - MS Chapter 9 and 10
    - (optional: G Chapter 19)

Week 6: February 12th

- Topics: Syntax
    - Context-free parsing
    - Dependency Parsing
- Readings:
    - SLP3 Chapters 11-14
    - MS Chapters 11 and 12
    - Michael Collins. Probabilistic context-free grammars, 2011. URL http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/pcfgs.pdf
    - Mark Steedman. A very short introduction to CCG, 1996. URL http://www.inf.ed.ac.uk/teaching/courses/nlg/readings/ccgintro.pdf

Week 7: February 19th

- Topics: Unsupervised Learning
    - Brown Clustering
    - EM Algorithm

    - Topic Modeling
- Readings:
    - SLP3 Section 16.4 (Brown clustering)
    - MS Section 14.2
    - Michael Collins. The naive Bayes model, maximum-likelihood estimation, and the EM algorithm, 2011. URL http://www.cs.columbia.edu/~mcollins/em.pdf
    - (optional) Kevin Knight. Bayes with Tears https://www.isi.edu/natural-language/people/bayes-with-tears.pdf
    - (optional) Resnik and Hardisty. Gibbs Sampling for the Uninitiated, 2010. URL http://www.cs.umd.edu/~hardisty/papers/gsfu.pdf (Great if you want to understand more of the intuition and math behind LDA starting from Naive Bayes!)

Week 8: February 26th

- Topics: Finishing Topic Modeling + new Neural Networks
    - Gibbs Sampling
    - Elmo, Bert, GPT-2
    - Machine translation
- Readings:
    - SLP2 Chapter 25

Week 9: March 5th (No class; spring break)

- Topics: 😎🌞🌊🏖️🎉💯

Week 10: March 11th

- Topics: Learning From Text
    - Information Extraction
    - Natural Language Inference
    - Question Answering
- Readings:
    - SLP3 Chapters 21, 28
    - SLP2 Chapter 22

Week 11: March 18th

- Topics: Discourse
    - Chat bots

    - Coreference Resolution
- Readings:
    - SLP3 Chapter 29, 30

Week 12 (regular class time): March 25th

- Topics: Lexical Semantics
    - Word Sense Disambiguation
    - Entity Linking
    - Semantic Roles + Frames
    - Diachronic Change
- Readings:
    - SLP3 Chapter 17, 22
    - MS Chapter 7

Week 13: April 1st

- Topics: Human Understanding of Language
    - Computational Sociolinguistics
    - Digital Humanities
    - Social NLP
- Readings:
    - All on Canvas

Week 14: April 8th

- **UMSI Winter Expo at Palmer Commons (time TBD, but during middle of the day)**
    - Your project's poster presentation (5% of your grade!)
    - Palmer Commons
    - Time TBD (will announce this as soon as I know)

Week 15: April 15th

- Topics: Data Acquisition
    - Annotation
    - Crowdsourcing
    - Construction of Truth
    - Hypothesis Testing
    - Algorithmic Bias
- Readings:
    - PS chapter 6

- Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Ng. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. URL http://www.anthology.aclweb.org/D/D08/D08-1027.pdf
- Jeff Bigham. My MTurk (half) Workday. URL: http://www.cs.cmu.edu/~jbigham/posts/2014/half-workday-as-turker.html
- Dirk Hovy and Shannon L. Spruit. The Social Impact of Natural Language Processing. URL http://www.dirkhovy.com/portfolio/papers/download/ethics.pdf
- R Artstein, M Poesio. Inter-coder agreement for computational linguistics. http://www.mitpressjournals.org/doi/pdfplus/10.1162/coli.07-034-R2

PREREQUISITES:

SI 330 or 507/508
Students should also have a good understanding of the following topics, though they can be learned throughout the duration of the course:

- Linear algebra: vectors and matrices
- Probabilities: random variables, discrete and continuous distributions, Bayes' theorem
- Programming: using a UNIX environment

You'll need to know how to write some programs on your own from scratch *and* how to extend an already-developed code-base to finish your assignment.

ASSIGNMENTS:

Students will have a programming-based assignment on a roughly bi-weekly basis. Assignments will be done in the python programming language, though students are allowed to use other languages with instructor permission. Python is strongly preferred for the NLP course due to the availability of deep learning libraries and other NLP-related packages. However, Java and R are also possible.

The full list of assignments and deliverables is as follows:

Week	Assigned	Due
1	HW0	
2	HW1	
3		HW0

4	HW2	HW1
5		Project Proposal
6		
7	HW3	HW2
8		
9	Relaxing (Spring Break)	Relaxing
10	HW4	HW3
11		Project Update
12	HW5	HW4
13		Midterm
14		Project Presentation
15		HW5
Finals Week		Project Report

READINGS:

Students will submit weekly responses to the readings. The first is a reading response to one of the research papers for that week. In this response, a student summarizes the paper and then comments on and critiques the paper's contributions. The responses are intended to (i) teach critical reading of academic papers (ii) expose students to the latest advancements in the field, and (iii) provide a model of the research process students can use in their own course projects. A detailed template for students to follow in these responses is provided through Canvas. All responses are expected to follow this template.

Students are expected to submit 8 of each response throughout the semester, for 1% of their grade each. Responses are graded as either a 0 or 1, with no regrades. The primary reasons for getting a 0 are not meeting the length requirement (it's one page) or writing overly vague responses. If for whatever reason you receive a 0, you can submit another response in a later week (if any are left) and get the points.

LATE POLICY:

Homework is due by the posted date and time. Students have three total late days they can use and no late homework after those days are used up is accepted without prior approval. For planned conflicts (e.g., conference travel, work interviews, etc.), students can request and extension by emailing the GSI at least 24 hours in advance of the deadline. Requests made under 24 hours will only be considered under exceptional circumstances (e.g., broken laptop, severe illness). Submitting at 12:05 will consume a whole late day, which is just as frustrating for us to see as it is for you. Please submit on time.

ACADEMIC HONESTY:

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. The instructors actively check for plagiarism and have zero tolerance. If caught, you will receive a zero for that portion of the assignment (or the whole assignment), depending on the scope of the plagiarism. All cases will be referred to the Office of Academic Integrity, regardless of their scope. Please (please) do not plagiarize; we hate finding these cases. If you are stressed and tempted to submit other's work, please come talk to us first.

Re-using worked examples of significant portions of other people's code from places like online tutorials, Kaggle examples, or Stack Overflow will be treated as plagiarism. If you are in doubt, please contact one of the instructors to check.

Any violation of the University's policies on Academic and Professional Integrity may result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program.

Violations of academic and professional integrity will be reported to Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed.

STUDENT MENTAL HEALTH AND WELLBEING:

The University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus.

You may also consult University Health Service (UHS) at (734) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a listing of other mental health resources available on and off campus, visit: http://umich.edu/~mhealth/

INFORMATION FOR STUDENTS WITH DISABILITIES:

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress.

As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations.

SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form.

I will treat any information that you provide in as confidential a manner as possible.