CS 4120 *Natural Language Processing*

Time: MW 2:50-4:30
Location: Shillman Hall 135
Instructor: Terra Blevins
Canvas: https://northeastern.instructure.com/courses/239517

Description
Natural Language Processing is a subfield of Artificial Intelligence that uses methods from Computer Science, Computational Linguistics, Cognitive Science, Statistics, and Machine Learning to give computers the ability to automatically analyze, categorize, understand, and generate natural language. Unlike other kinds of language (say, programming languages) natural language is often ambiguous, nuanced and subjective. In this course, we will learn about: (i) the linguistic phenomena that make NLP hard for computers to approach; (ii) the main NLP problems and tasks, and strategies to address them; and (iii) the ethical considerations and potentials for bias in NLP systems.

The course will be very much hands-on with an emphasis on methods to solve key NLP tasks, including classification, structured prediction, language modeling, and generation. We will first become familiarized with classical statistical methods, which are both important baselines and lay the foundation for more sophisticated approaches. Then, we will learn how to implement and train neural networks starting from simple Multilayer Perceptrons and build our way up to state-of-the-art models based on pre-trained Transformers and Large Language Models.

Class Staff

Name	Office Hours	Location
Terra Blevins	M 12:30-2:30	Behrakis 204 (In-person)
William Seward	T 3:30-5:30	Snell Library 043 (In-person)
Parin Shah	W 11:00-1:00	Zoom (Online)
Chaitanya Agarwal	Th 2:00-4:00	Snell Library 007 (In-Person)
Akshitha Bhashetty	F 4:30-6:30	Zoom (Online)

Please contact the course staff via Piazza or during their office hours. If you have any questions about the course material, you should post a public question – we will redirect you there if you message us directly on topics that would be beneficial to the whole class.

Prerequisites
Courses: CS 3100/CS 3500/DS 3500
Skills: Students must be familiar with Python to complete the programming assignments. There are no other required prerequisite technical skills; however, modern NLP relies heavily on statistical methods, machine learning, and deep learning. Therefore, students must be comfortable with basic mathematical concepts from Linear Algebra, Probability, and Calculus. We will briefly review some of the concepts

needed for the models and algorithms that we will cover in the class. However, this will be a review and NOT a thorough and rigorous exposition of these subjects. Students are thus encouraged to proactively fill any gaps in their knowledge.

Readings
Readings will be from the following two books (both have freely available drafts online):

- Speech and Language Processing 3rd Edition (SLP), *Dan Jurafsky and James H. Martin* pdf
- Dive into Deep Learning (DDL), *Aston Zhang, Zachary C. Lipton, Mu Li and Alexander J. Smola* pdf

Another useful resource is Natural Language Processing (NLP), *Jacob Eisenstein* pdf. We will have one reading from this book (for Convolutional Neural Networks, 01/28), but it covers many other topics in the course if you'd like an additional perspective.

Coursework and Grading
Homework assignments: These involve implementing various models discussed in class and will be evaluated on standard benchmark datasets. The assignments are due Wednesdays at 11:59 pm. Each student will receive **3** late days to be used on Homework Assignments. If all late days have been used, Homework Assignments can be turned in for a 50% penalty within three days after the submission deadline. Assignments not submitted within three days of the deadline will receive a score of 0.

Notebooks: In some of the lectures, we will work through exercises in Jupyter notebooks — these must be submitted on the lecture day (there are no late submissions for notebooks). These will be posted on Canvas by the day of the appropriate lecture.

Midterm: The midterm will be written in-class and will consist of theoretical and practical questions (including the calculations needed for specific algorithms).

Final Project: The final project is an opportunity to explore a topic in depth (e.g., replicate a specific paper, thoroughly investigate a model or a task). For this project, you must work in groups of three to four (we suggest that you try to form groups as soon as possible). Instructions will be posted in Week 3.

Your final grade will be calculated as follows. Each assignment in *Homeworks* and *Notebooks* are equally weighted.

- Homework Assignments (40%)
    - HW1 - Linear Classifiers
    - HW2 - Feedforward NNs
    - HW3 - Transformers
- Notebooks (5%)
- Midterm (20%)
- Final project (35%)

Regrade Considerations: If you are confused or concerned about feedback on course assignments, you may submit any requests for grading reconsideration **within 7 days after the feedback was released.** For homeworks, the regrade requests must be submitted through Gradescope. You may contact the course staff via Piazza to request a regrade for the midterm exam.

Schedule

Weeks	Topic	Date	Lectures	Materials + Readings	Assignments Due
Week 1	**Introduction + Course Logistics**	01/07 (W)	1. Introduction to Natural Language Processing		
Week 2	**Linear Models for Text Classification**	01/12 (M)	2. Text Normalization and Classification	SLP 2.2-2.7 DDL 1, 2 (optional) Notebook 1	Notebook 1
		01/14 (W)	3. Naive Bayes	SLP 4 Notebook 2	Notebook 2
		01/19 (M)	No Class (Martin Luther King Jr. Day)		
Week 3	**ML Models for Text Classification**	01/21 (W)	4. Logistic Regression; Perceptron	SLP 5 DDL 4	
Week 4	**Neural Networks for Text Classification**	01/26 (M)	5. Multilayer Perceptron	SLP 7 DDL 5 Notebook 3	Notebook 3
		01/28 (W)	6. Neural Networks	NLP 3.4 DDL 7-8 Notebook 4	Notebook 4
Week 5	**Training Neural Networks** and **Word Embeddings**	02/02 (M)	7. Training Deep Neural Networks	DDL 12 Notebook 5	Notebook 5
		02/04 (W)	8. Word Embeddings	SLP 6.8-6.13 DDL 15.1-15.6 Mikolov+ (optional)	**HW1 - Linear Classifiers**
Week 6	**Word Embeddings** and **Evaluations**	02/09 (M)	9. Word Embeddings pt. 2	See 02/04 Notebook 6	Notebook 6
		02/11 (W)	10. Evaluation, Bias, and Fairness	SLP 4.7 Bolukbasi+	

		02/16 (M)	No Class (Presidents' Day)		
Week 7	**Language Models**	02/18 (W)	11. N-gram Language Models	SLP 3 Notebook 7	Notebook 7 **Final Project Proposal**
Week 8	**Language Models and Sequence Labeling**	02/23 (M)	12. Neural Language Models	SLP 7.6-7.7 DDL 9-10 Notebook 8	Notebook 8
		02/25 (W)	13. Sequence Labeling	SLP 17.1-17.3	**HW 2 - Feedforward NNs**
		03/02-03/04	No Class (Spring Break)		
Week 9	**Sequence Labeling**	03/09 (M)	14. Hidden Markov Models	17.4 Notebook 9	Notebook 9
		03/11 (W)	15. Attention	SLP 9.1 DDL 11.1-11.3; 11.5-11.6	
Week 10	**Midterm**	03/16 (M)	16. Midterm Review		
		03/18 (W)	Midterm Exam		
Week 11	**Large Language Models**	03/23 (M)	17. Transformer Language Models	SLP 10.1-10.2 DDL 11.7-11.9; 15.8-15.10 Illustrated transformer Vaswani+ (optional)	**Final Project Check-in Writeup**
		03/25 (W)	**No class:** Project Check-in Meeting		
Week 12	**Large Language Models**	03/30 (M)	18. Sequence 2 Sequence Models	SLP 8.7-8.8	
		04/01 (W)	19. Pretraining; In Context Learning	SLP 10.3; 11.1-11.4; 12.1 TBA+	**HW 3 - Transformers**
Week 13	**Special Topics in NLP**	04/06 (M)	20. Post-training	SLP 12.2-6	
		04/08 (W)	21. Interpretability		

Week 14	**Final Project**	04/13 (M)	Final Project Presentations	
		04/15 (W)	Final Project Presentations, cont.	
		04/20 (M)	No Class (Patriots' Day)	
Week 15	**Final Project**	04/22 (W)	Final Project Presentations, cont.	**Final Project Deliverables**

Course Policies

Academic Integrity: Please read the Northeastern Academic Integrity Policy. All students are required to adhere to this policy during the course. Note that while students are encouraged to discuss course materials, no plagiarism/copying is allowed. **Assignments and code that you turn in should be written entirely on your own**.

You are expected to use the internet as a place for online resources, such as documentation, not as a place to get solutions to your assignments. This includes the use of LLMs or other AI models: while you may use them as a resource, it is not acceptable to have them generate code or other assignment solutions. **If any chatbots, LLMs, AI coding agents, or other sources are used to complete the assignment, you must *fully* disclose where and how this used.** Failure to do so is a violation of the course's Academic Integrity policy.

Academic Integrity violations will result in a 0 on the assignment in question.

Classroom Environment: To create and preserve a classroom atmosphere that optimizes teaching and learning, all participants share the responsibility of creating a civil, non-disruptive forum for the discussion of ideas. Students are expected to conduct themselves at all times in a manner that is respectful towards all participants and does not disrupt teaching or learning. The instructor reserves the right to interrupt conversations that deviate from these expectations. Repeated unprofessional or disrespectful conduct may result in a lower grade or more severe consequences.

Title IX: Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, and transgender students, faculty, and staff.

If you or someone you know has been a survivor of a Prohibited Offense, *confidential* support and guidance can be found through **University Health and Counseling Services** staff and the **Center for Spiritual Dialogue and Service** clergy members. By law, those employees are not required to report allegations of sex or gender-based discrimination to the University.

Alleged violations can be reported non-confidentially to the Title IX Coordinator within **The Office for Gender Equity and Compliance** at titleix@northeastern.edu and/or through **NUPD** (Emergency

617.373.3333; Non-Emergency 617.373.2121). Reporting Prohibited Offenses to NUPD does **NOT** commit the victim/affected party to future legal action.

Faculty members are considered "mandatory reporters" at Northeastern University, meaning they are required to report all allegations of sex or gender-based discrimination to the Title IX Coordinator.

Please visit https://www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.

Students with Disabilities: Students who have disabilities who wish to receive academic services and/or accommodations should visit the Disability Resource Center at 20 Dodge Hall or call (617) 373-2675. If you have already done so, please provide your letter from the DRC to me early in the semester so that I can arrange those accommodations.