CS 4120: Natural Language Processing, Spring 2023

Instructor:
Prof. Felix Muzny[1] (pronunciation: "Muse-knee"; pronouns: they/them and he/him)
Contact: f.muzny@northeastern.edu
Office: Meserve 307A

TAs:
Swati Agarwal (she/her, *sw-ah-tih*)
Grace Brown (she/her)
Ankit Ramakrishnan (he/him , *un-kith*)

Contact: we prefer that you make a post on piazza to contact course staff, as this will give you the shortest wait time for responses. Please post all general questions as follow-ups on piazza. For personal course content questions, make private piazza posts.

Location and credit:
Credit: 4 credits
Time: Tuesdays & Fridays 9:50 - 11:30am
Location: Shillman 325
Zoom link for synchronous remote lectures: go to “zoom meetings” on the left-hand side of the Canvas course
Remote attendance form: link

Office Hours: see course website

Lecture Format, Accessibility, Masking: Participating in synchronous course content (lectures, office hours) will be important to your overall learning and is expected. If you need to attend lecture remotely on any given day, you will need to fill out the form available on the course website and attend via Zoom. If you are attending remotely, do expect to be actively participating in all course content and collaborating with your peers in group activities.

As an on-ground class, it is expected by the university that the primary mode of instruction is in-person, on-ground. Please prepare yourself to caffeinate/hydrate/feed yourself as needed to be doing math and programming before noon :).

Prof. Felix will be masking for the entirety of the semester in accordance with their own health needs and to increase accessibility of their course. They request that all students both stay up-to-date with at-home covid tests (available for free for students in Forsyth Hall, further information here) and that all students mask for a minimum of the first two weeks of classes, following a similar request in response by Boston Public Schools to the current covid/flu/rsv/winter illness spike in the greater Boston area (read about the BPS policy here if you are interested).

[1] Call them "Felix", "Professor Muzny", or "Professor Felix"

Even though our pandemic-based world has changed dramatically since early 2020, there are a few lessons from this time that I'd like to emphasize:

- *life happens, communicate as proactively as you can and we will work together*
- *if you are sick in any way, do not come to class in person, even if you have received a negative covid test; attend class remotely if you are able to*
- *accessibility is important and providing accessible instruction is beneficial to everyone; if you are attending class remotely please make sure that you are set up for success in terms of focus, concentration, and engagement*

Course Overview

NLP is about getting computers to perform useful and interesting tasks involving spoken and written human language. NLP is sometimes referred to as Computational Linguistics to emphasize the fact that this subject involves the combination of CS methods with research insights from Linguistics (the study of human language). Practical applications of NLP include question answering, machine translation, information extraction, and interactive dialog systems (both written and spoken). Modern NLP systems rely heavily on methods involving probability, linear algebra, and calculus --- often in combination with machine learning methods.

We'll be exploring both applications and the computational methods behind them. You should be prepared to get your hands dirty in terms of the math, programming, and data that comprise the behind the scenes components of NLP systems.

Course Goals

1. Develop an understanding of the general problems that people who work on NLP study and the strategies they use to solve them.
2. Understand the role of data, machine learning, and neural networks in NLP systems.
3. Understand the ethical considerations and potentials for bias in NLP systems.
4. Be able to implement models to solve some "standard" NLP problems.
5. Be able to formulate potential starting points given a new problem with NLP elements.
6. Understand some of the motivating linguistic phenomena that make NLP problems hard and why these can be hard phenomena for computers to approach.

Topics

- Words, word counting, lexicons
- Probabilistic language modeling
- Ethics and bias in NLP
- Text classification with language models
- Text classification with single layer neural networks
- Vector semantics & word embeddings
- Part-of-speech tagging
- Viterbi algorithm (dynamic programming)
- Machine translation
- State of the art systems and transfer learning
- Information extraction*

- Question answering systems*
- Dependency parsing*

* if time allows

Textbook & Course Configuration

1. We'll be using draft chapters from the 3rd Edition of *Speech and Language Processing* by Dan Jurafsky and James H. Martin. You don't need to buy the current edition, draft pdfs of the new chapters are available from the textbook website. You can also download (and print, if you desire) the entire book from the website. We will also link a pdf of this text from the course website.
    a. We will be using the draft version from Jan 12, 2022 (link)
2. We will supplement this text occasionally with readings from:
    a. Eisenstein, Jacob. *Introduction to Natural Language Processing*. MIT Press, 2019.
    b. Kohn, Philipp. "Neural Machine Translation." arXiv preprint arXiv:1709.07809, 2017.

Websites & Technology

Make sure that you have access to all of the following websites and software:

- **Canvas**: We'll be using Canvas for some quizzes and links to homework submissions. Homework submissions will be done through Gradescope.
- **Gradescope**: Gradescope is where you will submit homework and some quizzes. You will also see your grades, feedback, and submit regrade requests via gradescope. You can find the link to Gradescope on Canvas.
- **Piazza**: This is our course discussion forum. This is where we will discuss relevant topics and answer your homework and content questions that come up outside of class. If you send us a content question via email, we'll likely ask you to post it to Piazza instead!
- **Python 3**: We'll be writing our homework coding assignments using python 3.
- **Jupyter Notebooks**: Many of the coding activities that we complete in class will be distributed as Jupyter Notebooks. You can install jupyter notebooks either by installing Anaconda or via the command line.
- **IDEs**: You can develop your code using whatever your preferred IDE is. When you have coding assignments, ***you'll submit your solutions as .py files, not jupyter notebooks*** (make sure to run any converted-from-jupyter-notebooks-.py-files again before you turn them in)! If you installed Anaconda, it comes with Spyder, which is an IDE that can be used to write and run .py files.
    - If you have less experience working with python and ***both*** Jupyter Notebooks and .py files, we *highly* encourage you to make time to come to office hours in the first few weeks of the course.

Classroom Environment & Expectations

- **Preparation:** When there are readings assigned, it is the expectation that you do them before the first class meeting in the following week. This course will be a great opportunity for those of you who are interested in NLP & research to start flexing those muscles, and the best way for us to go down those paths is for you to develop a solid foundation.
- **Attendance:** You are expected to attend lecture synchronously whenever possible. **You are expected to attend in person whenever this is a reasonable choice**. We will be doing interactive activities

during lecture as well as covering the material necessary for you to complete your homework and quizzes. If you are unable to attend lecture on a given day, it is your responsibility to attend office hours, consult course materials, and communicate with course staff as needed about what you have missed. You will receive credit for completing lecture activities (see grading rubric below), which are often easiest to complete during lecture when facilitated by course staff.

- **Classroom environment:** It is unusually common in Computer science classes for some students to ask questions that are not really questions so much as opportunities to demonstrate knowledge of vocabulary or facts beyond the topic at hand. This can have a discouraging effect on other students who are not familiar with those terms, causing them to worry that they are less prepared to do well in the class (this is rarely the case—knowing terms outside the scope of the course is not a good predictor of success). If you find yourself wanting to make such a question or comment, please come talk to me about the topic after class or during office hours—I'm always happy to discuss tangentially related topics at those times!
- **Accommodation letters:** If you have an accommodation letter, please bring it to me at your earliest convenience so that I can make sure this class is meeting your needs.
- **Name and pronouns:** If your name and pronouns are not in alignment with those listed on our class roster, please let me know either in person or via email so that I can ensure you are correctly addressed in this class.
    - If you wish to add, change, or update your pronouns in Canvas, go to "Account" > "Profile" > "Edit Profile", then add, change, or update your pronouns and display name.
    - If you wish to change or update your name here at Northeastern as a whole, find instructions with the registrar here.
- **Class expenses:** If obtaining any material for use in our class presents a financial hardship for you, please let me know and I will work with you to locate the resources that you need to succeed in this class.
- **Feedback:** Please don't hesitate to reach out to me if any aspect of this course or class community could be improved.
- **Illness:** If you are ill in any way, you are expected to attend class remotely, even if you have received negative covid tests. If you are unable to attend class remotely due to illness, you are expected to communicate with Felix and attend office hours as needed.

Late Policy

All homework should be turned in on time whenever possible. All homework may be turned in up to 2 days (48 hours) late for a 20% penalty. For example, if homework is due on Wednesday at 9pm, it may be turned in as late as Friday at 9pm.
If a student would have received a 95% had they turned their homework in on time, a late submission will earn them a 75% instead.

Once a semester, you may turn in your homework up to 48 hours late without penalty or explanation. This will be automatically applied the first time you turn in your homework late. This policy only applies to homeworks and does not apply to the final project or quizzes, which cannot be turned in late.

Quizzes may not be completed after the deadline. We've built some extra credit into the course (you'll find this on the homework), so if you miss a quiz, make sure to attempt the extra credit on the homework.

Extensions

Extensions for any work beyond the regular late policy will be given based on proactive communication with Prof. Felix. Whenever possible, this should occur at least 24 hours before the posted deadline. The sooner that you reach out, the easier this will be.

Email Felix (f.muzny@northeastern.edu) with the following information:

1) Which assignment are you requesting an extension on & why you are requesting an extension.
2) When are you requesting the extension until. (a specific time and date)
3) What is your plan for how this extension will impact the due dates for the other assignments in this course. (do you foresee there being cascading effects of this extension)

You don't need to write an essay, just be sure to include the above information. This extension policy is based on our mutual understanding that living during and recovering from a pandemic can be difficult, we're all doing our best, and the easiest way for you to succeed in this course is proactive communication. If a situation arises that makes it impossible to reach out 24 hours before the deadline, don't panic—send Prof. Felix an email as soon as you can and we'll discuss your options together.

Collaboration Policy

The work that you turn in should be your own. We encourage you to collaborate with your classmates, but remember that collaboration when working on individual assignments looks very different than working on a pair or group project.

Here are three big-picture points to remember when collaborating with your classmates:

- **Strategies:** You may talk with your classmates about *general strategies* but you may not talk about *specific solutions*.
- **Explaining concepts:** You may talk with your classmates about how certain techniques work *in general* but not how to write any part (or sub-part) of the solution needed for the homework.
- **A good rule of thumb:** don't show your assignments to other people; don't look at other people's assignments (this makes it very hard to come up with your own solution afterwards); don't write code together unless the assignment explicitly states that you may work in pairs. This includes working through solutions on whiteboards as well as telling your friend verbally what you have written.

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.

The finer-grained details:

- **Do not search for a solution online**: You may not actively search for a solution to the problem from the internet. This includes posting to sources like StackExchange, Reddit, Chegg, etc.
    - **StackExchange Clarification**: Searching for basic techniques in python is fine. If you want to post and ask "How do convert a float to an integer" that's fine. What you **cannot** do is post things like "Here's the function my prof gave me to write. I need to convert this temperature in celcius to farenheit".

- **Plagiarism:** assignments **and code** that you turn in should be written entirely on your own. Do not use AI Pair Programmers. You should not need to consult sources beyond the class notes, posted lecture notes, examples, and resources, and python and its associated libraries' documentation.
- **Tutors:** you should always consult the course instructional staff if you need extra help. They are here specifically to help you! You should never have anyone else write code for you. This includes tutors, friends, strangers, friends of friends, or anyone who is not you.
- **When in doubt, ask**: If you have doubts about this policy or would like to discuss specific cases, please ask the instructor.

Collaboration Policy violations will result in a 0 on the assignment in question.

The university's academic integrity policy discusses actions regarded as violations and consequences for students.

http://www.northeastern.edu/osccr/academic-integrity

Grading & Assignments

If you enrolled in this course after deadlines have passed **and** you contact Felix ASAP, we will work with you to adjust deadlines as needed.

Category	Due Dates & Points	Grade Percentage
Homework	Due on **Wednesdays at 9pm**. No HW grades will be dropped.	45%
Quizzes	Quizzes are due on **Fridays at 9pm** (most weeks when there is no homework due). It may be helpful for you to think of quizzes as mini take-home tests. Quizzes will typically focus on material covered in class in the preceding weeks, and may have a small number of questions based on readings and/or videos for the upcoming week.	20%
Class participation	Class participation is determined based on participation during lecture, doing lecture activities, participating on piazza (taking part in discussions and helping answer your peers' questions), and by attending office hours. You do not need to do all of these things to earn full credit in this category, but you should expect to do at least two of them fully. (e.g. participating in lecture and doing lecture activities or participating on piazza and coming to office hours). If you cannot attend lecture in person or	10%

	synchronously, you may still submit lecture activities. (And you are encouraged to do so! Do expect follow-up from course staff in this case.) You will receive this grade in about four installments: each about ¼ of the way through the semester. February 14th and 17th will have explicitly assigned points due to special activities that we'll be working on with guest lecturers on those days. If you are taking your wellness day on a Tuesday or Friday, we ask that you submit the activity assignment for the day—just write down that you took your wellness day on that date. (You may, of course, still do the class activities if you choose to).	
Final Project	Your final project will consist of four portions: 1) proposal (due Wednesday, March 29th; meet with Felix between March 27th & April 4th) 2) code & write-up (due Wednesday, April 19th @ 9pm) 3) presentation (due **in class** on Tuesday, April 18th) 4) reflection/extra credit (due Friday, April 21st @ 9pm) Final projects may be completed individually or with one or two partners. Final project details to be released in March.	25%

Final grades will be based on the following scale. All numbers will be rounded to the nearest integer.

Letter	**Range**
A	95 - 100
A-	90 - 94
B+	87 - 89
B	83 - 86
B-	80 - 82
C+	77 - 79
C	73 - 76
C-	70 - 72
D	60 - 69
F	< 60

Calendar

The course calendar will be (subject to change at the instructor's discretion):

Week	Topics	Reading	Homework (due Wednesdays @ 9pm)	Quiz (due Fridays @ 9pm)
1	Introduction to NLP, Vocabularies, normalization, python	SLP, Chp. 2		
2	n-grams, language models	SLP, Chp. 2, 3		Quiz 1
3	text classification, naive bayes	SLP, Chp. 4	Homework 1	
4	Naive bayes, logistic regression, SGD	SLP, Chp. 4, 5		Quiz 2
5	semantics, Word Embeddings	SLP, Chp. 5, 6	Homework 2	
6	Ethics + NLP	Supplementary Materials		
7	neural nets, neural language models, SotA systems	SLP, Chp. 7; supplementary materials	Homework 3	
8	academic papers, POS tagging, HMMs	SLP, Chp. 8; Appendix A; supplementary materials		Quiz 4
	SPRING BREAK			
9	Viterbi, SIGCSE (no class Friday, March 17th)	SLP, Chp. 8		Quiz 5
10	NER, RNNs	SLP, Chp. 9, 8	Homework 4	
11	LSTMs, Machine Translation	SLP Chp. 9, 10; Supplementary materials	Final project proposal due	
12	Machine Translation, bias + NLP, Transfer Learning	SLP Chp. 10, 11; Supplementary materials		Quiz 6

13	SoTA systems, "bonus" topics	SLP Chp. 11; Supplementary materials		
14	final project presentations		Final project due* (see grading table)	
15	finals week - NO FINAL EXAM			

Classroom Recording

This course, or parts of this course, may be recorded for educational purposes. These recordings will be made available only to students enrolled in the course, instructor of record, and any teaching assistants assigned to the course.

If you have objections or would like to opt-out of recordings, please contact the instructor.

Only students who have arranged an accommodation with the Disability Resource Center may use mechanical or electronic transcribing, recording, or communication devices in the classroom. Students with disabilities who believe they may need such an accommodation may contact the Disabilities Resource Center.

Accommodations

It is my job to create a classroom environment that is most conducive to you learning well. If you have accommodations from the Disability Resource Center, please provide your letter to me early in the semester so that I can arrange for these accommodations. If you wish to receive accommodations and do not have a letter, please visit the DRC at 20 Dodge Hall or call (617) 373-2675.

Student Names and Pronouns

We recognize that your legal information doesn't always align with how you identify. Students may update their first and middle names as well as gender marker with the registrar, even if they are not your legal names or gender marker. Those names and gender marker are what would appear publicly in most university systems. In the absence of such updates, what we see on most university systems by default are your legal name and gender marker.

Classroom Environment

To create and preserve a classroom atmosphere that optimizes teaching and learning, all participants share a responsibility in creating a civil and non-disruptive forum for the discussion of ideas. Students are expected to conduct themselves at all times in a manner that does not disrupt teaching or learning. Your comments to others should be constructive and free from harassing statements. You are encouraged to disagree with other students and the instructor, but such disagreements need to respectful and be based upon facts and documentation (rather than prejudices and personalities). 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.

Part of the learning process in this course is respectful engagement of ideas with others.

The Code of Student Conduct can be found on the OSCCR website.

Title IX

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

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, non-binary, 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.

Reports can be made 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 "responsible employees" at Northeastern University, meaning they are required to report all allegations of sex or gender-based discrimination to the Title IX Coordinator.

In case of an emergency, please call 911.

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

Religious Holidays

The course staff will make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required attendance. In this class, contact the course staff at least 7 days in advance of the conflicting date to reschedule a homework or quiz due date.