CS 2731
Introduction to Natural Language Processing

Session 1: Course and NLP introduction

Michael Miller Yoder

August 25, 2025

School of Computing and Information

Overview: Course introduction and NLP basics

- Introductions
- What is NLP?
- Course logistics
- NLP applications and tasks

About Michael Miller Yoder

- You can call me "Michael"
- Teaching faculty, Pitt School of Computing and Information
- BA, Computer Science from Goshen College (2013)
- PhD, Language Technologies Institute at Carnegie Mellon University (2021)
- **Research interests:**
  - natural language processing (NLP)
  - computational social science
  - data science
  - ethics and bias in AI

Michael's office hours

- By appointment in person at Sennott Square 6309 or on Zoom
- Sign up for a slot here
  - Link also posted on course website
- Drop in to ask questions about the course or anything else

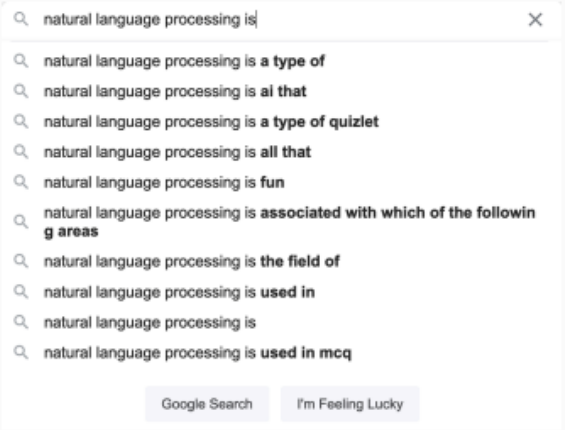
Introductions

1. What is your name?
2. What is your program/year/research interests?
3. What is a language or dialect other than English that you speak, or some your ancestors spoke?
4. [Optional] Is there anything that makes you interested in NLP or excited to take this class?

What is natural language processing (NLP)?

NLP is Everywhere

Did you ever wonder how web search engines work...



...or how Google can anticipate what you're searching for?

That's NLP!

Slide credit: David Mortensen

NLP is Everywhere

Did you ever wonder how ChatGPT generates language?

That’s NLP!

NLP is Everywhere

Did you ever wonder how digital assistants work?

That's NLP!

Slide credit: David Mortensen

NLP is Everywhere

Did you ever wonder how the government is spying on your every word?

That's also NLP!

Slide credit: David Mortensen

Natural language processing

- Also known as computational linguistics
- "Natural language" = human languages (not programming languages)
- Computational **analysis** and **synthesis** of language and speech
- Processing language with computers
- Intersects with
  - artificial intelligence (AI)
  - machine learning (ML)
  - large language models (LLMs)

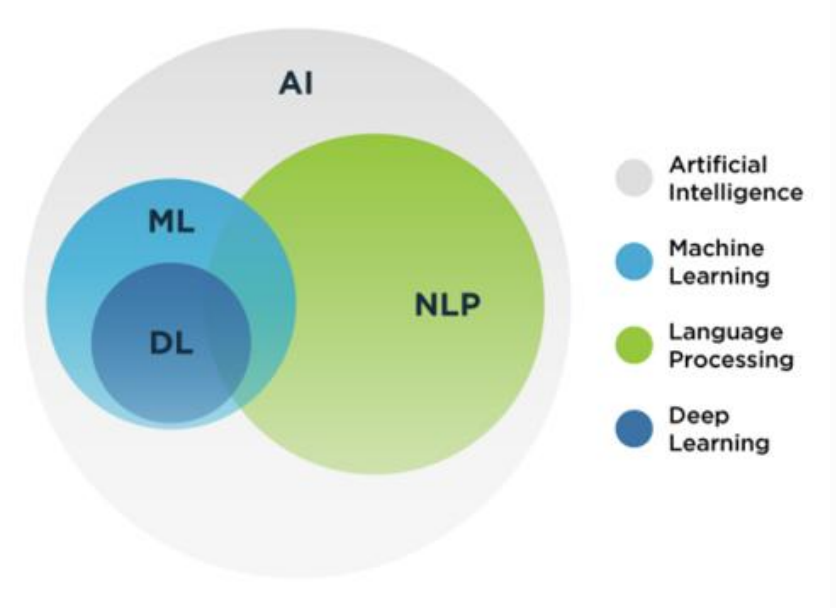


Image from slide by David Mortensen

A brief history of NLP

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

The Georgetown-IBM Experiment.
Credit: John Hutchins

- 1950s: **foundations**
  - Turing Test ("Computing Machinery and Intelligence" paper)
  - Georgetown-IBM Experiment translating Russian to English
- 1960s-1980s: **symbolic reasoning**
  - ELIZA, rule-based parsing, hand-built conceptual ontologies
- 1990s-2010s: **statistical NLP**
  - Learn patterns from large corpora (feature-based machine learning)
- 2000s-2020s: **neural NLP**
  - “Deep" layers of neural networks
- 2020s-today: **LLMs**
  - Transformer-based large pretrained models capable of impressive performance on many tasks

The other NLP 😂

Neuro-linguistic programming (pseudoscience)

Course objectives and overview

Learning objectives

At the end of this course, a student will be able to structure an NLP system to achieve a desired outcome from language data.

Learning objectives

When coming across a natural language problem, students will be able to:

- Recognize the class of tasks that a specific natural language task belongs to
- Explain the basics of language structure from linguistics (morphology, syntax, semantics, discourse) that are relevant to NLP
- Preprocess text into a machine-readable format
- Extract needed features from text for a variety of tasks
- Identify a suitable model to tackle the task
- Evaluate algorithms for that task
- Identify potential ethical pitfalls in an NLP system and how to potentially address them
- Communicate motivation, key components, and implications of an approach to NLP tasks in writing

Structure of this course

MODULE 1	Introduction and text processing	text normalization, machine learning, NLP tasks	
	Approaches	How text is represented	NLP tasks
MODULE 2	statistical machine learning	n-grams	language modeling text classification
MODULE 3	neural networks	static word vectors	text classification
MODULE 4	transformers and LLMs	contextual word vectors	language modeling text classification
MODULE 5	Sequence labeling and parsing	named entity recognition, dependency parsing	
MODULE 6	NLP applications and ethics	machine translation, chatbots, search engines, bias	

Resources

Textbook (free)

- Dan Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd edition draft, **2025-08-24**.
- **Available completely free online:** https://web.stanford.edu/~jurafsky/slp3/
- Why do the readings?
  - Learn better: get the information from readings and lectures
  - Spend class time more efficiently: come with questions
  - There will be in-class quizzes that cover content in the readings

Class sessions

- Cover the most important parts of the course content
- Students are expected to attend each class
- **Attendance will be taken via Top Hat at a number of random class sessions**
- **Bring a laptop or tablet for in-class coding exercises**
- Slides will be provided in advance of each lecture for note-taking
- There are no current plans for recording classes

Infrastructure: website

- How do I find the website?
  - https://michaelmilleryoder.github.io/cs2731
  - Or https://tinyurl.com/nlppitt
  - Link is in “Syllabus” on Canvas
  - In first Canvas announcement
  - From the ‘Teaching’ page of my website: https://michaelmilleryoder.github.io
- Up-to-date syllabus and schedule
- Class slides
- Homework assignment and project instructions

Infrastructure: Canvas

- Submit assignments
- Receive course announcements
- Post questions
- Check your grade

Programming languages and software

- **Python will be the expected programming language used in assignments**
- Python-based data science packages (numpy, pandas, jupyter, scikit-learn, pytorch) will be used and encouraged in both assignments and the project
- If you have zero familiarity with Python:
    - Check out the **Tutorials on Python and data science** section of the course website under **Learning resources**
- Let us know if you want to use other languages for assignments (it's probably fine)
- You can use whatever you want for the project

Assessments

Assessment overview

Assessment	Points	Percentage of grade
Homework assignments (4) total	430	43%
Project	410	41%
Participation	100	10%
Quizzes (6) total	60	6%

No exams

Homework assignments

- 4 total + an intro/setup homework
- ~10% of total course grade each
- Hands-on coding assignments in Python
- Due ~2 weeks after they are released
- Descriptions will be on the course website
- Submitted through Canvas

Homework 0 on setup of CRCD account will be released sometime early this week (when students receive CRCD accounts), will be due next Wed Sep 3

Project

NLP is inherently hands-on. The course project will demonstrate an ability to build a system that **makes a contribution** to NLP research or practice.

- Self-selected topic, type of research contribution, and idea
  - Can fit with your research interests outside of this class
  - Come up with your own idea or choose one of the example project ideas
- We will solicit ideas for the projects through a form, to be advertised to all students anonymously
- During class on project match day, you will find groups of 2-4 based on project idea interests
- There will be peer review of project teammates
- Types of contributions: new dataset analysis and/or annotations, new approach/application, new evaluation, new survey

Project components

Component	Points	Percentage of course grade	Due
Idea form	5	0.5%	09-11
Proposal	85	8.5%	10-16
Proposal presentation	*None*	*None*	10-20
Progress report	85	8.5%	11-13
Final presentation	*None*	*None*	12-08
Final report	235	23.5%	12-09

Quizzes

- Checks for comprehension of the main important ideas in preceding class sessions
- Designed to motivate you to keep up with the reading and come to class
- Auto-graded, generally multiple choice or short answer
- 6 total
- The lowest quiz scores will be dropped
- Only 6% of your course grade total
- If you will be gone that day, let me know and I will open up the quiz for you

Quizzes

- Will be completed in class on Canvas on Wednesdays (check the schedule)
- Allowed resources
  - Textbook
  - Your notes (on a computer or physical)
  - Course slides and website
- Resources not allowed
  - Generative AI
  - Internet searches
- First quiz will be **Wed Sep 10**

Participation grade

- Class interactions (activities, discussions) are better with more people in class
- Incentives to come to class and engage
- 10% participation grade
  - 6%: attendance on a random subset of class sessions, taken via Top Hat
  - 4%: engagement
    - Have you ever asked a question in class, afterward or over email?
    - Do you participate in in-class activities?
    - If yes to either, you will be fine

Policies

Grading scale

Range	Letter grade
92.5 – 100%	A
90.0 – <92.5%	A-
87.5 – <90.0%	B+
82.5 – <87.5%	B
80.0 – <82.5%	B-
77.5 – <80.0%	C+
72.5 – <77.5%	C
70.0 – <72.5%	C-
67.5 – <70.0%	D+
62.5 – <67.5%	D
60.0 – <62.5%	D-
< 60%	F

Late work

- Students are granted 5 total late days across all homework assignments without penalty.
- After those five late days, you will be penalized **10% for each day that your submission is late** except in extreme unforeseen circumstances, up to a maximum of 40% off.
- Group project work will be penalized 10% for each day late up to a maximum of 40% off. No late work will be accepted for the final project report.

Homework resubmissions

- If you are unsatisfied with your grade on an assignment and wish to resubmit work, talk with me
- If you completely miss parts of an assignment are missing (sections of the rubric are 0), a resubmission may be possible.
- Updated or added text in resubmitted reports must be highlighted in yellow.
- Resubmissions are subject to an automatic 10% deduction. Only 1 resubmission per homework assignment will be accepted.
- Resubmissions must be submitted by 11:59pm on the last day of class (Dec 8)

Academic integrity

- Students in this course will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity
- Discussing tools, concepts, and formalisms is acceptable collaboration
- Sharing code is prohibited. *You* knowing how to implement NLP systems is a key learning objective

Generative AI policy

- You are allowed to use generative AI (ChatGPT, DALL-E, GitHub Copilot, etc) in some circumstances
  - Exposes you to the current capabilities and limitations of such systems
- Allowed use:
  - **Use as an aid, not for a finished product.** Generating ideas, study guides, bibliographies (watch for hallucinations, though) is ok. Drafting entire homework assignments or project reports, even if you revise the draft, is not ok.
  - **Cite its use.** Citing the generative AI's tool contribution to your work is required. See the APA guidelines on how to cite ChatGPT.
  - **You are responsible for the work you turn in.** LLMs and other generative AI systems can and do generate biased, socially problematic language and assert unfounded claims.
- When in doubt, ask instructor if specific uses are ok. There will be no retaliation for asking

Disability rights

Many people have disabilities. We view disabilities as deficits not in disabled people but in the institutions and societies that are structured to disadvantage disabled people.

If you have a disability (visible or invisible), please let us know as soon as possible (you don't need to tell us the nature of the disability). You are encouraged to work with Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will work with you to determine reasonable accommodations for this course. This might include lecture materials that are usable by people with visual disabilities, sign language interpretation, captioning, flexible due dates, etc.

Adapted from policies by David Mortensen and Lori Levin

Maintaining scholarly discourse

In this course we will be discussing some complex issues. It is essential that we **approach this endeavor with our minds open** to evidence that may conflict with our presuppositions. Moreover, **it is vital that we treat each other's opinions and comments with courtesy even when they diverge and conflict with our own**. We must avoid personal attacks and the use of ad hominem arguments to invalidate each other's positions. Instead, we must develop a culture of civil argumentation, wherein all positions have the right to be defended and argued against in intellectually reasoned ways. It is this standard that everyone must accept in order to stay in this class; a standard that applies to all inquiry in the university, but whose observance is especially important in a course whose subject matter is so emotionally charged.

Questions?