COS 484: Natural Language Processing

L1: Introduction to NLP

Spring 2024
Logistics and course structure
Course staff

Instructor
Karthik Narasimhan

Graduate TAs
Anika Maskara
Colin Wang
Howard Yen
Yihan Wang
Logistics

Course webpage: https://nlp.cs.princeton.edu/cos484/

- All readings are optional but can help strengthen your understanding.
- All assignments will be due on Friday at 12pm EST before the lecture.

Schedule

Lecture schedule is tentative and subject to change. All assignments are due at 12pm EST before Friday lectures.

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topics</th>
<th>Readings</th>
<th>Assignments</th>
</tr>
</thead>
</table>
| 1    | Fri (2/2) | Introduction to NLP    | 1. Advances in natural language processing  
2. Human Language Understanding & Reasoning | A1 out       |
|      |        | n-gram language models | J & M 3.1-3.5                                             |              |
| 2    | Fri (2/9) | Text classification | Naive Bayes: J & M 4.1-4.6                                |              |
Logistics

- **Lectures:** Friday 1:30pm - 4:20pm in CS 104 (10-15min break in between)

- **Precepts:** 1-hour precept every week taught by TAs (optional)
  - Tuesdays, 3:30-4:30pm, location will be announced on Ed
  - Refresher for basic maths, review course materials, Colab/Pytorch tutorial etc
Logistics

• Sign up for Ed and Gradescope today
  • We will use iClicker for polls in the class.

• Ed will be the main (and only) forum for all class-related questions and discussion.
  • You can use private Ed posts for personal matters.
  • No email please

• We would like to help in our office hours!
  • All the information will be posted this week on the website
  • We will also have a number of undergraduate TAs’ office hours
Assignments (40%): 4 total

- **A1 (10%)** will be released TODAY and due in two weeks
- **A2, A3, A4**: 10% each - each assignment has 2 or 3 weeks
- Every assignment has a **written** component and a **programming** component based on Colab (A4 requires access to GPUs).
- You will need to know how to program in PyTorch/Numpy.

You have **96 free late hours** for all assignments; After that, 10% penalty for each late day (up to a maximum of 3 days beyond which submissions will not be accepted)
Course structure

Midterm (25%)
- A 3-hour timed exam on Gradescope in a 27-hour window (March 7-8)
- All the topics up to March 1 will be covered
- No final exam
Course structure

Final project (35%)

- Complete in a team of 3
- Two options: (a) reproducing a state-of-the-art NLP paper (ACL/EMNLP/NAACL 2021-2022) and coming up with your analysis, ablations or innovations (**strongly encouraged**) (b) competing a research project
- Proposal (0%) due before - date will be announced soon
- Poster presentation tentatively scheduled on May 3 (during reading period)
- Final report due on May 7

Extra bonus (5%) - participation in class and Ed discussions

*No pre-determined cut-offs for final grades*, will be decided at the end taking into account the performance of the entire class and will be fairly assigned to measure your level of understanding of the subject.
Textbooks
(NLP is a rapid-moving field…)

https://web.stanford.edu/~jurafsky/slp3/

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin

Here’s our Jan 7, 2023 draft! This draft is mostly a bug-fixing and restructuring release, there are no r
the applications section earlier, reflecting how we and others tend to teach NLP, and combines the linguisti

A good way to learn about state-of-the-art NLP concepts is through research papers and blog posts

2nd edition ✗
Course goals

• Gain an understanding of the **fundamentals** of different sub-fields within NLP

• Understand **theoretical** concepts and algorithms

• **Hands on experience** building statistical models for language processing

• Carry out an **independent research project** at the end
This is an advanced class

Image: https://matt.might.net/articles/phd-school-in-pictures/
This is a very advanced class

Image: https://matt.might.net/articles/phd-school-in-pictures/
Prerequisites

• **Required: COS324**, knowledge of probability, linear algebra, calculus (A0 will give you a sense)
• Be ready to pick up new ML concepts
• Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Q. Why is COS324 a prerequisite?

We assume you have learned the following concepts already:
• Language models
• Logistic regression w/ regularization
• Unsupervised vs supervised learning
• Feedforward neural networks, convolutional neural networks
• PyTorch programming
• (A little bit of reinforcement learning)
What is NLP and what is the course about
Natural Language processing

• NLP = building **computer programs** to analyze, understand and generate **human language** - either spoken or written (informal)

• NLP is an interdisciplinary field
Natural Language processing

• NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)

Communication with humans (ex. personal assistants, customer service)  
Access the wealth of information about the world — crucial for AI systems

Banking assistant

**ONLINE**

**OFFLINE**
MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization.

**Meta’s New AI Ranked in the Top 10% at the Game ‘Diplomacy’—and Human Players Were None the Wiser**

By Edd Gent > November 28, 2022

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

Ability to understand and generate language ~ intelligence
How it started

"Within three or five years, machine translation will be a solved problem"

Georgetown experiment

1954

How it’s going
The man shot the elephant while wearing his pajamas

आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

Man shoots elephant wearing his pajamas
> Pick up the red block on top of a green one
> OK.

- THREE OF THEM

> How many red blocks are there?

- THRESHOLD

- Rule-based, requiring extensive programming
- Limited domain

SHRDLU, 1968
Statistical learning

• Use of machine learning techniques in NLP
• Increase in computational capabilities
• Availability of electronic corpora

Unsupervised vs. supervised?
Statistical learning

IBM translation models

Speech recognition

Anytime a linguist leaves the group the (speech) recognition rate goes up
- Fred Jelinek 1998
The era of deep learning

- Significant advances in core NLP technologies
- **Essential ingredient**: large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**
The era of pre-training / LLMs

- BERT, ELMo, ERNIE...
  - Leverages a lot of unlabeled text
  - Model size increased by $10^3 - 10^5$ in parameters
ChatGPT examples

What NLP topics should I teach in 2023?

Here are some NLP topics that are popular and relevant in 2023:

1. Transformers (BERT, GPT, etc.)
2. Transfer Learning in NLP
3. Multi-lingual NLP
4. Attention Mechanisms
5. Adversarial Training in NLP
6. NLP for Question Answering
7. Sentiment Analysis
8. Text Generation (summarization, translation, etc.)
9. Named Entity Recognition
10. Dialogue Systems

Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.
ChatGPT examples

what do you know about the Princeton NLP group?

The Princeton NLP group is a research group at Princeton University that focuses on the study of natural language processing (NLP) and its applications. The group is led by Professor Christopher Manning, who is a leading expert in the field of NLP. The group has published numerous research papers and has developed several widely-used NLP tools, including the Stanford Parser and Stanford Named Entity Recognizer. The group also collaborates with other research groups and organizations, such as the Allen Institute for Artificial Intelligence and Google, on NLP research projects.

Why do you hallucinate so much?

As an AI language model, I don't have the capacity to hallucinate as I don't possess emotions, sensations, or consciousness. I can only generate responses based on patterns in the text I was trained on.
NLP subfields

- Machine translation
- Question answering
- Information extraction
- Text summarization
- Dialogue
- Semantic parsing
- Syntactic parsing
- Language grounding
- ....
Topics we will cover in this class

• We will focus on fundamentals, important concepts rather than concrete applications
• The class will be a mix of statistical and neural NLP approaches (>70% neural)

- N-gram language models
- Text classification
- Word embeddings $\times 2$
- Sequence models $\times 2$
- Neural networks for NLP
- Recurrent neural networks $\times 2$

(Many of the above topics can be studied in multiple courses and papers)

Before midterm

- (neural) machine translation $\times 2$
- Transformers $\times 2$
- BERT, ELMo, …
- Large language models

+ guest lectures!

After midterm
Why is language difficult to understand?
Why is language difficult to understand?

- Ambiguous
- Dialects
- Accents
- listener has to infer - pragmatics
- humor, sarcasm, irony
- context, dependencies
Lexical ambiguity

The fisherman went to the *bank*.

*bank*

/ˈbaŋk/   (noun)

plural noun: banks

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe. More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company.

One word can mean several different things
Lexical ambiguity

The fisherman went to the *bank*. He deposited some money.

**bank**

/ˈbaŋk/ (in British English)

noun

plural noun: banks

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe. More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company.

Word sense disambiguation
Lexical variations

According to the thesaurus, “they’re humid, prepossessing homosapiens with full sized aortic pumps” means “they’re warm, nice people with big hearts.”

Several words can mean the same thing!
Distributed representations

Project words onto a continuous vector space

Similar words closer to each other

\[ v(\text{king}) - v(\text{man}) + v(\text{woman}) = v(\text{queen}) \]
Comprehending word sequences

- My brother went to the park near my sister’s house
- Park my went house near to sister’s my brother the
- “My brother went park near sister’s house”? 
- The old man the boat
- Fat people eat accumulates

Garden Path sentence

• Implicit structure in all languages
• Coarse-to-fine levels (recursive)
• What are some good data structures to represent this?
Syntactic ambiguity

I shot an elephant in my pajamas

Human language is full of such examples!
Syntactic parsing

Parse tree

Sentence

His duties will be assumed by John Smith who has been elected deputy chairman

Online tools: http://nlp.stanford.edu:8080/corenlp/
Discourse ambiguity

- The man couldn’t lift his son because he was so heavy.
- The man couldn’t lift his son because he was so weak.

What does “he” refer to?

- The city councilmen refused the demonstrators a permit because they feared violence.
- The city councilmen refused the demonstrators a permit because they advocated violence.

What does “they” refer to?

Anaphora resolution
Next: n-gram language models