Natural Language Processing

Introduction, course logistics.

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Welcome!

https://courses.cs.washington.edu/courses/cse447/24au/

Mon / Wed / Fri 3:30–4:20pm, CSE2 G20

CSE 447: Natural Language Processing, Fall 2024

MWF 3:30-4:20pm, CSE2 G20 (Gates, ground floor)

Instructor: Yulia Tsvetkov
yuliats@cs.washington.edu
OH: available on Zoom by appointment.

Teaching Assistant: Melanie Sclar
msclar@cs.washington.edu
OH: Fri 11:00-12:00pm, CSE1 220, Zoom

Teaching Assistant: Kabir Ahuja
kahuja@cs.washington.edu
OH: Tue 11:00-12:00pm, CSE1 220, Zoom

Teaching Assistant: Kavel Rao
kavelrao@cs.washington.edu
OH: Mon 4:30-5:30pm , CSE2 150, Zoom

Teaching Assistant: Khushi Khandelwal
khushik@cs.washington.edu
OH: Mon 11:00-12:00pm, CSE1 3rd Floor Breakout

Teaching Assistant: Melissa Mitchell
mcm08@cs.washington.edu
OH: Wed 2:30-3:30pm, CSE2 150, Zoom

Teaching Assistant: Riva Gore
rivagore@cs.washington.edu
OH: Thu 1:30-2:30pm, CSE2 150, Zoom

Instructors ❤️

Teaching Assistant: Melanie Sclar
msclar@cs.washington.edu
OH: Fri 11:00-12:00pm, CSE1 220, Zoom

Teaching Assistant: Kabir Ahuja
kahuja@cs.washington.edu
OH: Tue 11:00-12:00pm, CSE1 220, Zoom

Teaching Assistant: Kavel Rao
kavelrao@cs.washington.edu
OH: Mon 4:30-5:30pm , CSE2 150, Zoom

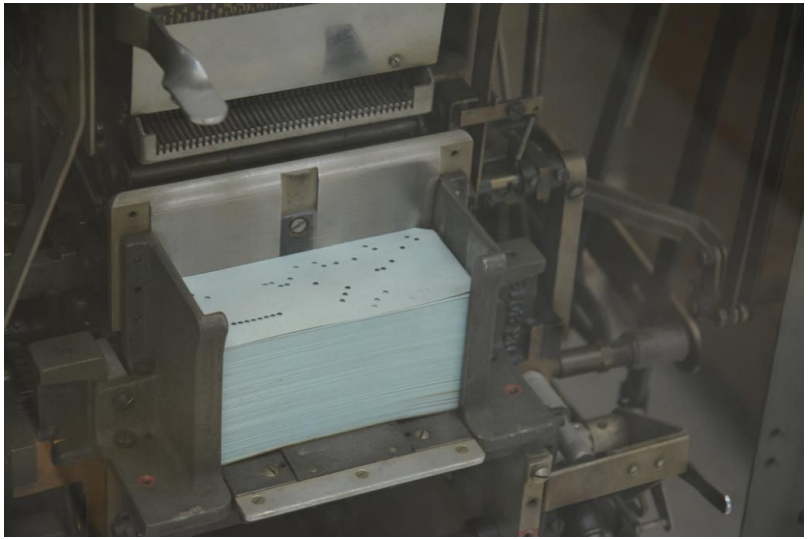
Teaching Assistant: Khushi Khandelwal
khushik@cs.washington.edu
OH: Mon 11:00-12:00pm, CSE1 3rd Floor Breakout

Teaching Assistant: Melissa Mitchell
mcm08@cs.washington.edu
OH: Wed 2:30-3:30pm, CSE2 150, Zoom

Teaching Assistant: Riva Gore
rivagore@cs.washington.edu
OH: Thu 1:30-2:30pm, CSE2 150, Zoom

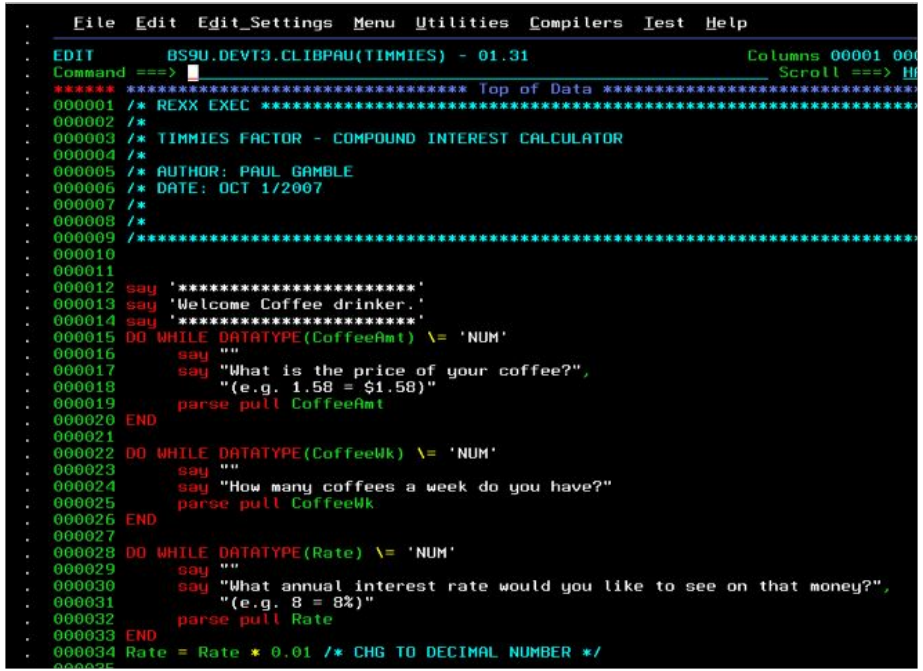
Communication with machines

- ~1950s-1970s



Communication with machines

- ~1980s

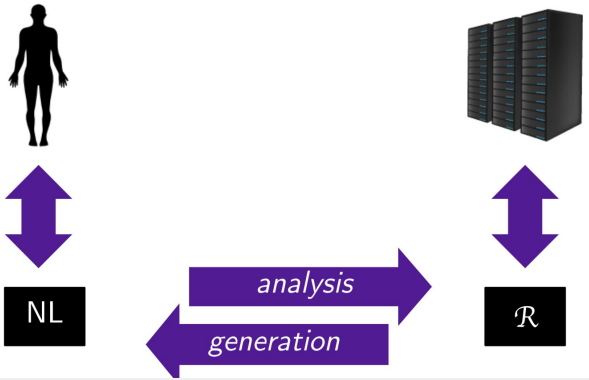


NLP: Communication with machines

- Today

What is Natural Language Processing (NLP)?

- NL ∈ {Mandarin Chinese, Hindi, Spanish, Arabic, English, … Inuktitut, Njerep}
- Automation of NLs:
  - analysis of ("understanding") what a text means, to some extent (NL → $\mathcal{R}$)
  - generation of fluent, meaningful, context-appropriate text ($\mathcal{R}$ → NL)
  - representation learning – acquisition of $\mathcal{R}$ from knowledge and data



Language technologies

What technologies are required to write such a program?

A conversational agent contains

-
-
-
-
-

Language Technologies

A conversational agent contains

- Speech recognition
- Language analysis
- Dialog processing
- Information retrieval
- Text to speech

Natural Language Processing

A conversational agent contains

- Speech recognition
- Language analysis
  - Language modelling, spelling correction
  - Syntactic analysis: part-of-speech tagging, syntactic parsing
  - Semantic analysis: named-entity recognition, event detection, word sense disambiguation, semantic role labelling
  - Longer range semantic analysis: coreference resolution, entity linking
  - Deeper semantic analysis: reasoning, knowledge, ect.
- Dialog processing
  - Discourse analysis, user adaptation, etc.
- Information retrieval
- Text to speech

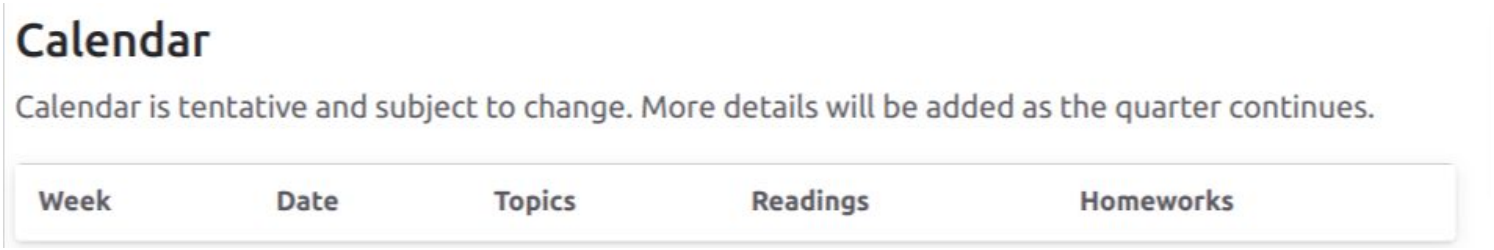
Syllabus https://courses.cs.washington.edu/courses/cse447/24au/

- Introduction
  - Overview of NLP as a field
- Modeling (ML fundamentals)
  - Text classification: linear models (perceptron, logistic regression), non-linear models (FF NNs, CNNs)
  - Language modeling: n-gram LMs, neural LMs, RNNs
  - Representation learning: word vectors, contextualized word embeddings, Transformers
- Linguistic structure and analysis (Algorithms, linguistic fundamentals)
  - Words, morphological analysis,
  - Sequences: part of speech tagging (POS), named entity recognition (NER)
  - Syntactic parsing (phrase structure, dependencies)
- Applications (Practical end-user solutions, research) - subject to change
  - Text classification: Sentiment analysis, toxicity detection
  - Language generation: Machine translation, summarization, QA
  - Ethics: bias, misinformation

Course structure

Will be updated on the course website's syllabus

https://courses.cs.washington.edu/courses/cse447/24au/

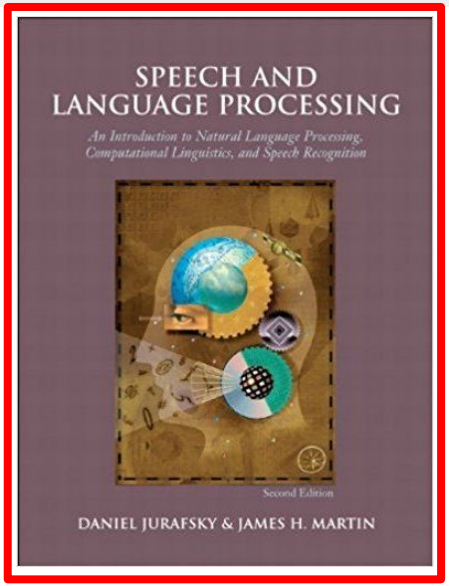


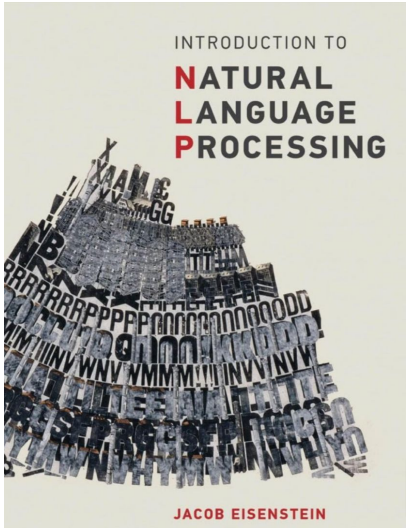
Calendar

Calendar is tentative and subject to change. More details will be added as the quarter continues.

Week	Date	Topics	Readings	Homeworks

Readings





- https://web.stanford.edu/~jurafsky/slp3/
- https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- +additional readings posted weekly

Course website

- https://courses.cs.washington.edu/courses/cse447/24au/
- Office hours, announcements, calendar, etc.
  - But check most recent announcements on Ed

Deliverables & grading

- **Homework projects** - 90%
  - 3 programming assignments, 30% each
  - "Semi-autograded" – Most of the grades (~70-80%) come from evaluating if the submission passes the hidden test cases. Sample test cases will also be provided for students to check their implementations. The rest of the grades would involve writeups on algorithm details, performance trends, and other conceptual questions.
  - We'll discuss the setup in detail when HW1 is released
- **Quizzes** - 10%
  - 8 simple quizzes weekly
  - 10 minutes at the beginning or end of the class
  - Starting from the 3rd week
  - 5 best quizzes, 2% each
- **Participation in course discussions** - 6% **bonus**
  - **A helpful response to HW questions** and discussions from your classmates on Ed
  - Contribute "insightful" discussions on Ed - 2% extra credit per response, 6% max

Homework assignments

- Project 0: [Optional] **Python and Pytorch Tutorial / Review**
- Project 1: **Text Classification and N-gram language models**
  - Implementing Naive Bayes and Logistic Regression for text classification
  - Training, evaluating, and sampling from n-gram Language Models
- Project 2: **Neural Text Classification and Neural Language Modeling***
  - Training feed-forward neural networks for text classification using word2vec and sentence transformers representations
  - Training a transformer-based language model from scratch
- Project 3: **Fine-tuning and Prompting Pre-trained Language Models***
  - Fine-tuning pre-trained model for text classification
  - Prompting LLMs for reasoning / QA. Will cover different prompting methods like In-context learning, CoT, and self-consistency as well as other tricks such as RAG

*Subject to change based on factors like class performance, compute feasibility, and topics covered during the course.

Late submissions

- **Late policy**
  - Each student will be granted **5 late days** to use over the duration of the quarter.
  - You can use a **maximum of 3 late days on any one project**.
  - Weekends and holidays are also counted as late days.
  - Late submissions are automatically considered as using late days.
  - Using late days will not affect your grade.
  - However, **projects submitted late after all late days have been used will receive no credit.** Be careful!
- Additional late days
  - We allocate an extra week for each homework assignment
    - E.g. if we believe that the homework will take you 2 weeks to complete, we set a deadline in 3 weeks
    - Start early!
- We will not grant any extensions beyond these

Communications with instructors

- You should be able to see yourselves be added to the Ed discussion board of CSE 447 / CSE M 547 24 au. **Please contact the staff if you are not.**
- **Discussion Board (EdSTEM)** will be used to answer questions related to lectures and assignments
  - We really encourage you to ask/discuss higher level questions on the discussion board.
  - We encourage that generic questions should be posted as “Public” so that other classmates would also got benefited from it.
  - Please do not post detail about your solutions (detail ideas, codes, etc.) on public threads. Private discussion should be used for these posts.
- For grading issues, please email the instructor team directly.

Class participation

- **In-person** instruction!
- Lectures and homework assignments complement each other
- Lecture materials are broader
- Homework assignments will go deeper into important topics
- Try to attend the lectures
- Quizzes are designed to encourage you to do so
- But if you miss a lecture – you can read assigned book chapters, read slides
- Participate in class discussions on Ed, 6% bonus is an incentive
  - But don't just provide code solutions to questions on homework projects– those are for individual work!
  - Provide insights, theoretical background, references to readings
- **Your questions are always welcome!**

Office hours

- Yulia – Thu 3:30 - 4:15pm CSE 566 only by appointment - email me if you're planning to come!
  - Questions about lectures, research, NLP in general

Questions about homework assignments, grading, course logistics:

- Mon: Kavel 4:30 – 5:30pm CSE2 150
- Mon: Khushi 11:00 – 12:00pm CSE1 3rd Floor Breakout
- Wed: Melissa – 2:30pm - 3:30pm CSE2 150
- Tue: Kabir – 11:00 – 12:00pm CSE1 220 (HW lead; only HW or NLP research questions, not grading/logistics)
- Thu: Riva – 1:30pm - 2:30pm CSE2 150
- Fri: Melanie – 11:00 - 12:00pm CSE1 220 (head TA; NLP research, outstanding personal issues, logistics, not grading)

- Teaching sections
  - We'll announce when we will have a teaching section
  - Not held by default

Quizzes

- 8 quizzes, students can drop 3
- Each quiz has ~5 simple multiple-choice questions, autograded
- Quizzes are on Canvas, open during the lecture time
- Quiz time - 10 minutes in the beginning of the class
- Starting from the 3rd week
- On Fridays unless we announce otherwise
- Grading on 5 best quizzes, 2% each
- Important: only Canvas window should be open during the quiz. We autograde the quiz but then check report from Canvas if you left the window during the quiz (e.g. switched to Chrome). We will zero-out all reported quizzes.

ChatGPT, Copilot, and other AI assistants

- Quizzes: canvas functionalities
- Homework assignments
  - You can “consult” with ChatGPT like you’d do with another student in the class
  - You cannot feed HW questions and paste solutions
  - We’ll run automated plagiarism checks
  - In the assignments you’ll be asked to clarify whether/how you used generative AI

ChatGPT Answers Programming Questions Incorrectly 52% of the Time: Study

To make matters worse, programmers in the study would often overlook the misinformation.

By **Matt Novak** Published May 24, 2024 | Comments (14)

Photo: Silas Stein/picture-alliance/dpa/AP (AP)

Artificial intelligence chatbots like OpenAI’s ChatGPT are being sold as revolutionary tools that can help workers become more efficient at their jobs, perhaps replacing those people entirely in the future. But a stunning new study has found ChatGPT answers computer programming questions incorrectly 52% of the time.

https://gizmodo.com/chatgpt-answers-wrong-programming-openai-52-study-1851499417

Session Information

Started at Thu Dec 10 2020 11:53:10 GMT-0700 (Mountain Star
Attempt 1

Action Log

00:03	Session started
00:33	Resumed.
00:48	Viewed (and possibly read) question #1
00:48	Viewed (and possibly read) question #2
00:48	Answered question: #1
00:53	Answered question: #2
01:03	Viewed (and possibly read) question #3
02:29	Answered question: #3
02:50	Answered question: #4
03:03	Viewed (and possibly read) question #4
04:18	Viewed (and possibly read) question #5
04:33	Answered question: #5
05:33	Viewed (and possibly read) question #6
05:49	Stopped viewing the Canvas quiz-taking page...
06:03	Resumed.

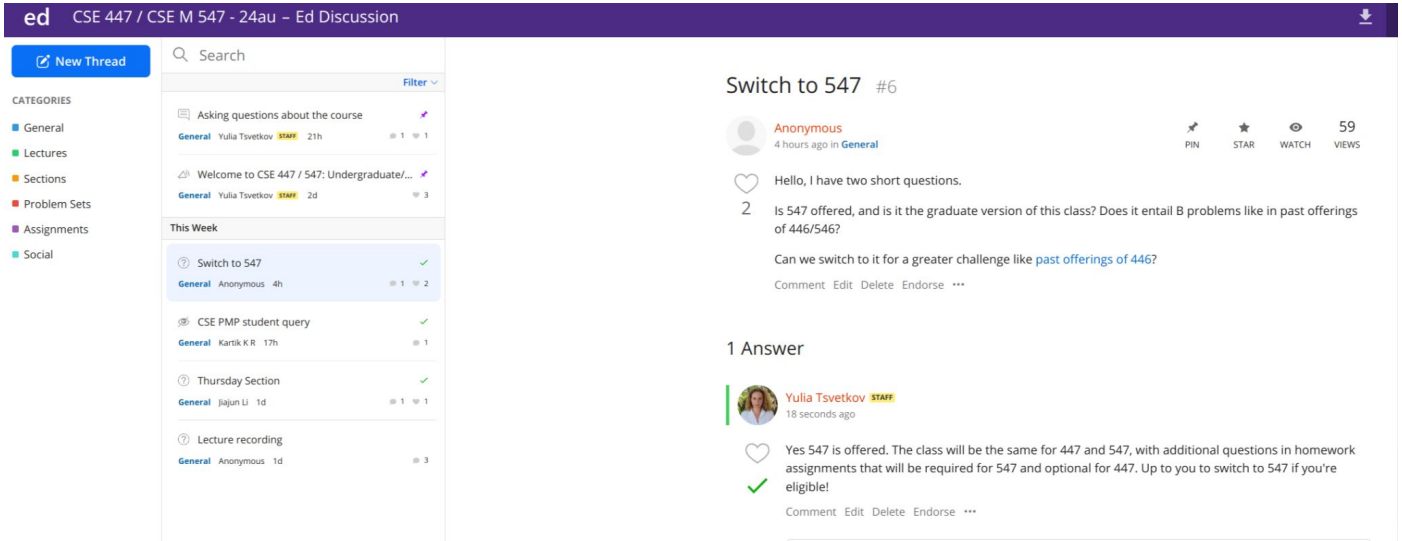
What background do I need to have?

- 447/547 prerequisite courses
- Python programming
- ML is not a prerequisite but we very strongly suggest to take the course only if you have some ML background
- Prior experience in linguistics or natural languages is helpful, but not required
- There will be a lot of algorithms and coding in this class, some statistics, probabilities, linear algebra

More course logistics

We care that you learn!

Your questions are always welcome.



447 vs 547

- Same course content
- Same quizzes
- Additional question in HW assignments for 547 that will be a bonus question for 447

Course registration

- The instructor cannot generate an Add Code
- ~~If you wish to register to the course and have completed prerequisite courses~~
  - ~~Fill out the 500 level course enrollment request form from (managed by the grad advisers)~~
    - ~~https://docs.google.com/forms/d/e/1FAIpQLSc9lbYwpg4KmbiCMmYSA7Ju11G8HZiSbnazwn9M4DNf1UGZOw/viewform~~
  - ~~Email Pim Lustig <pl@cs.washington.edu> and Ugrad Adviser <ugrad-adviser@cs.washington.edu> to request an Add Code~~
  - ~~Cc Yulia~~
- Unfortunately we cannot add new students to the course, sorry :(

Learning goals

At the end of this course, you will be able to:

- Build a supervised classifier to solve problems like sentiment classification
- Build a neural network and train it using stochastic gradient descent
- Build tools for extracting linguistic knowledge from raw text, e.g. names
- Learn ML fundamentals for text processings (including state-of-the-art methods)
- Learn important algorithms for text processings (that are useful also in other fields)
- Learn methodological tools (training/test sets, cross-validation)
- Build a toy “large” language model, fine-tune it for text classification, prompt it for question answering

- It's gentle (my goal is to explain everything) and broad (covering many many topics)
- Mastery independent learning, quizzes and programming homeworks
- No research project, but fun research-oriented lectures towards the end of the course

Questions?