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/cse447/22au/

Mon / Wed / Fri 3:30–4:20pm, CSE G01

CSE 447: Natural Language Processing, Autumn 2022

MWF 3:30-4:20pm, CSE2 G01

Instructor: Yulia Tsvetkov
yuliats@cs.washington.edu

Teaching Assistant: Daksh Sinha
daksh97@uw.edu

Teaching Assistant: Jacob Morrison
jacobm00@cs.washington.edu

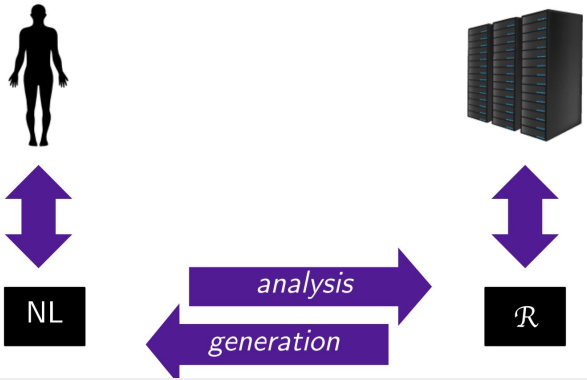
Teaching Assistant: Leo Liu
zeyuliu2@cs.washington.edu

Teaching Assistant: Leroy Wang
lryw@uw.edu

Teaching Assistant: Urmika Kasi
ukasi@uw.edu

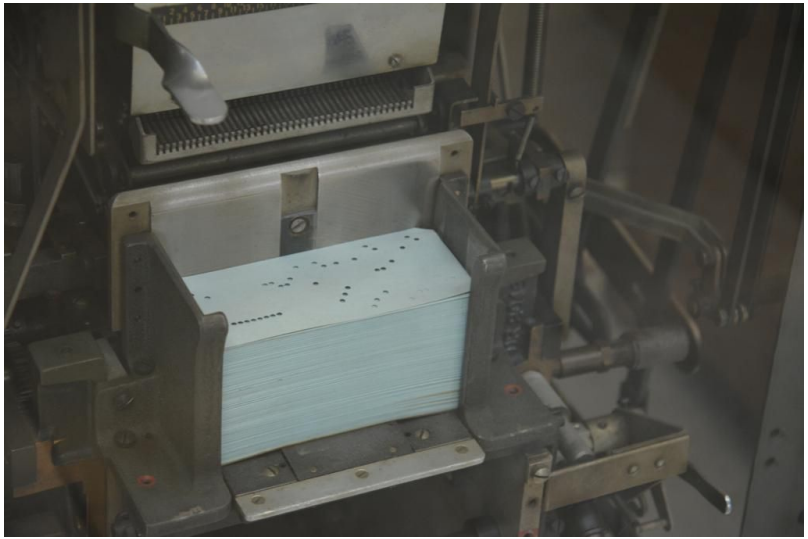
What is Natural Language Processing (NLP)?

- 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)
  - acquisition of $\mathcal{R}$ from knowledge and data



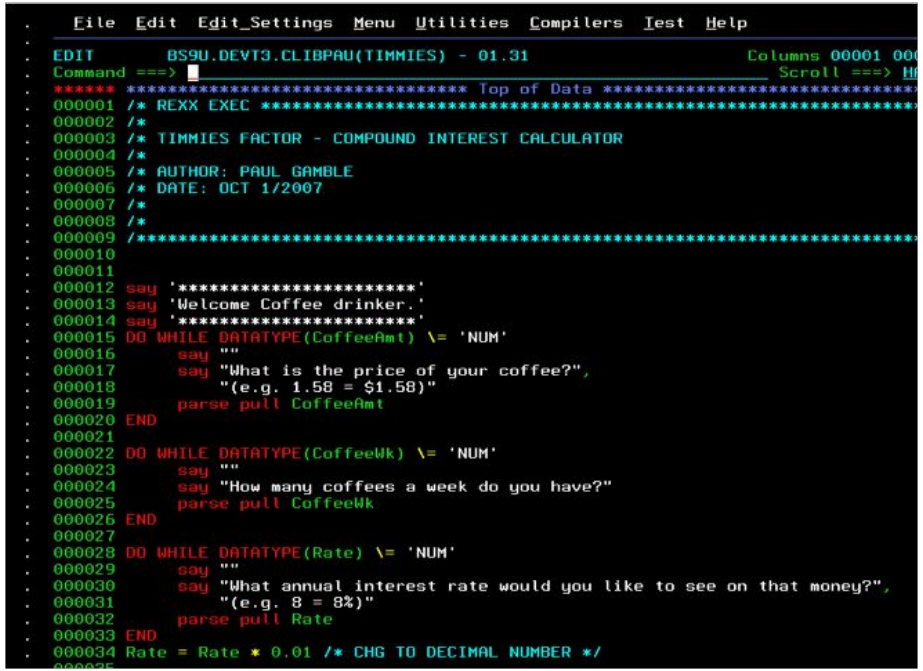
Communication with machines

- ~1950s-1970s



Communication with machines

- ~1980s



NLP: Communication with machines

- Today

Language technologies

What technologies are required to write such a program?

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
  - etc.
- Dialog processing
  - Discourse analysis, user adaptation, etc.
- Information retrieval
- Text to speech

Syllabus https://courses.cs.washington.edu/courses/cse447/22au/

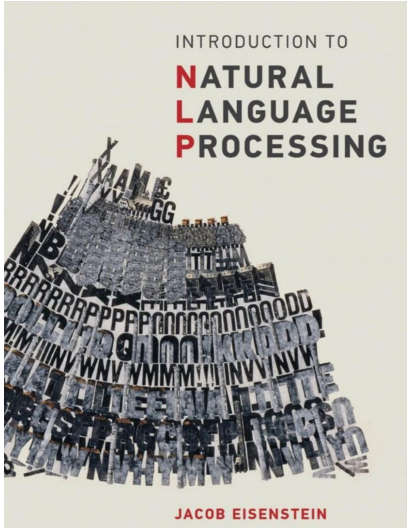
- Introduction
  - Overview of NLP as a field
- Modeling (ML fundamentals)
  - Text classification: linear models (perceptron, logistic regression), non-linear models (FF NNs, CNNs)
  - Language modeling: n-gram LMs, neural LMs, RNNs
  - Representation learning: word vectors, contextualized word embeddings, 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)
  - Sentiment analysis, toxicity detection
  - Machine translation, summarization
  - Computational social science
  - Interpretability
  - Fairness and bias

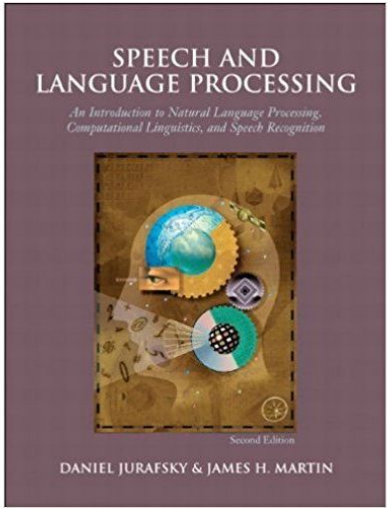
Course structure

please read the syllabus

https://courses.cs.washington.edu/courses/cse447/22au/

Readings





- https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- +additional readings posted weekly

Course website

- https://courses.cs.washington.edu/courses/cse447/22sp/
- Office hours, announcements, calendar, etc.

Deliverables & grading

- **Homework projects** - 90%
  - 3 programming assignments, 30% each
  - "Semi-autograded" – Most of the grades (~80%) come from replicating reference outputs in a given Jupyter notebook. You would usually know this part of your grades before submitting your assignments. The rest of the grades would involve things like write-ups, algorithm performance on hidden test sets, etc.
  - We'll discuss the setup in detail next week
- **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** - 10% **bonus**
  - **Respond to HW questions** and discussions from your classmates
  - Contribute "insightful" discussions on Ed - 5% extra credit per 3 responses (10% max)

Homework assignments

- Project 1: **Text classification**
  - We will build a system for automatically classifying song lyrics comments by era. Specifically, we build machine learning text classifiers, including both generative and discriminative models, and explore techniques to improve the models.
- Project 2: **Sequence labeling**
  - We focus on sequence labeling with Hidden Markov Models and some simple deep learning based models. Our task is part-of-speech tagging on English and Norwegian from the Universal Dependencies dataset. We will cover the Viterbi algorithm which could require a little bit prior knowledge of dynamic programming.
- Project 3: **Dependency parsing**
  - We will implement a transition-based dependency parser. The algorithm would be new and specific to the dependency parsing problem, but the underlying building blocks of the method are still some neural network modules covered in P1 and P2.

Homework submission

- **Submit via Gitlab**
  - We will pull your code for submission (with an assignment tag) and check the commit time.
  - A detailed grading rubric would be specified in the main Jupyter notebook of each assignment.

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 22 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 three 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
- Participate in class discussions, 10% 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 – Fri 2:30 - 3:15pm CSE 566 (preferably by appointment)
  - Questions about lectures, research, NLP in general, and course logistics

Questions about homework assignments:

- Mon: Urmika 12:00pm - 1:00pm
- Tues: Daksh 2:00pm - 3:00pm
- Wed: Leo - 2:00pm - 3:00pm
- Thu: Leroy - 12:30pm - 1:30pm
- Fri: Jacob - 2:00pm - 3:00pm

- 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
- Grading on 5 best quizzes, 2% each

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

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 statistics, algorithms, and coding in this class

More course logistics

We care that you learn!

Your questions are always welcome.

https://courses.cs.washington.edu/courses/cse447/22au/

Questions?