Natural Language Processing

Introduction, course logistics

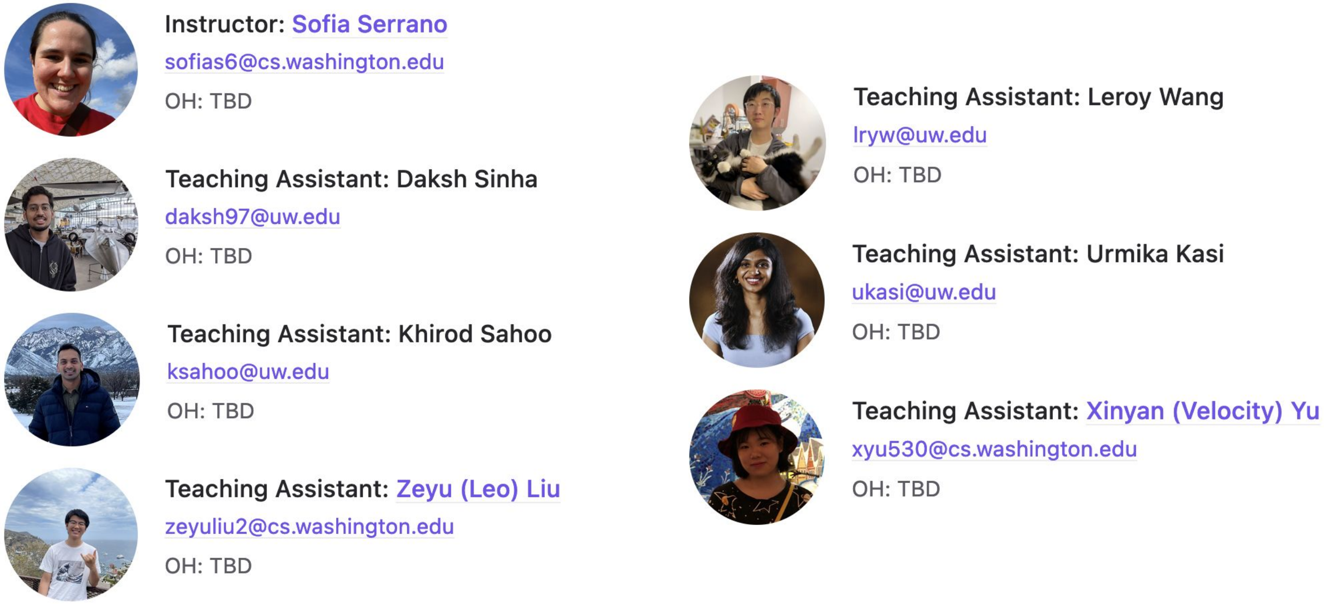
Sofia Serrano
sofias6@cs.washington.edu

Credit to Yulia Tsvetkov and Noah Smith for slides

Welcome!

https://courses.cs.washington.edu/courses/cse447/23wi/

Mon / Wed / Fri 1:30–2:20pm, CSE2 G01



Instructor: Sofia Serrano
sofias6@cs.washington.edu
OH: TBD

Teaching Assistant: Daksh Sinha
daksh97@uw.edu
OH: TBD

Teaching Assistant: Khirod Sahoo
ksahoo@uw.edu
OH: TBD

Teaching Assistant: Zeyu (Leo) Liu
zeyuliu2@cs.washington.edu
OH: TBD

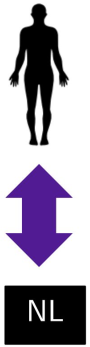
Teaching Assistant: Leroy Wang
lryw@uw.edu
OH: TBD

Teaching Assistant: Urmika Kasi
ukasi@uw.edu
OH: TBD

Teaching Assistant: Xinyan (Velocity) Yu
xyu530@cs.washington.edu
OH: TBD

What is Natural Language Processing (NLP)?

- NL ∈ {Mandarin Chinese, Hindi, Spanish, Arabic, English, American Sign Language… Inuktitut, Njerep}



What is Natural Language Processing (NLP)?

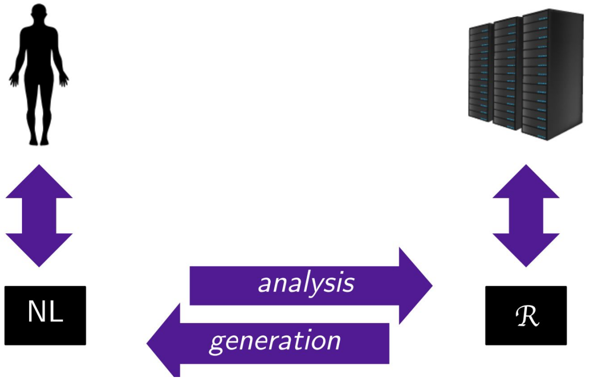
- NL∈ {Mandarin Chinese, Hindi, Spanish, Arabic, English, American Sign Language… Inuktitut, Njerep}

NL

$\mathcal{R}$

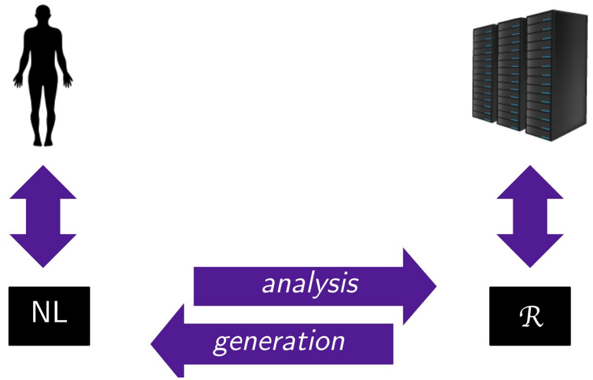
What is Natural Language Processing (NLP)?

- NL∈ {Mandarin Chinese, Hindi, Spanish, Arabic, English, American Sign Language… Inuktitut, Njerep}



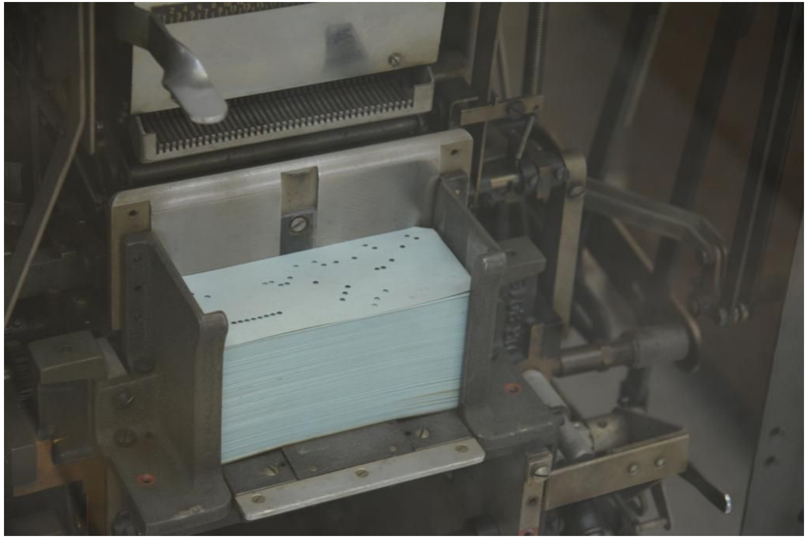
What is Natural Language Processing (NLP)?

- NL ∈ {Mandarin Chinese, Hindi, Spanish, Arabic, English, American Sign Language… 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



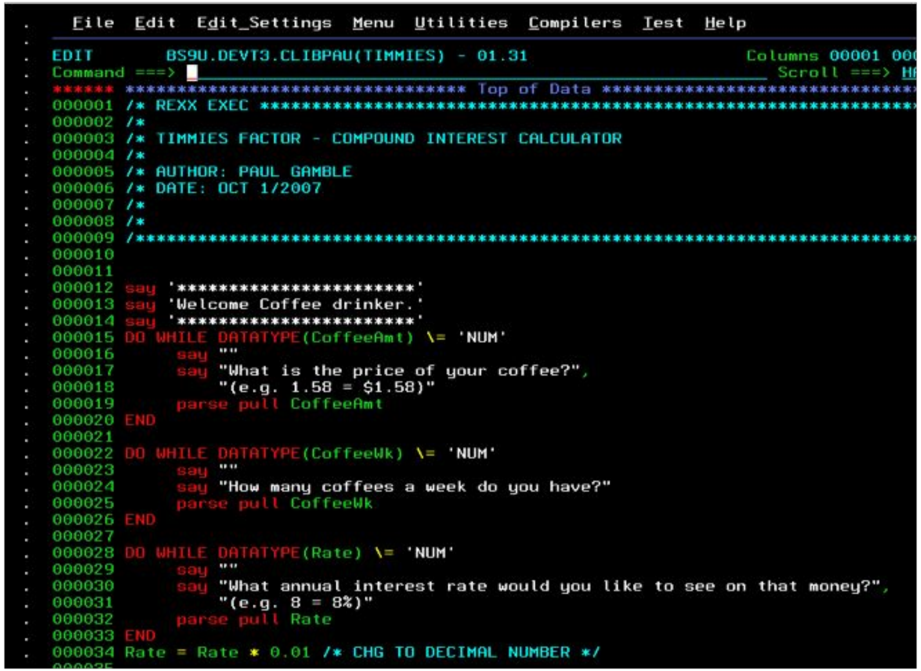
Communication with machines

- ~1950s-1970s



Communication with machines

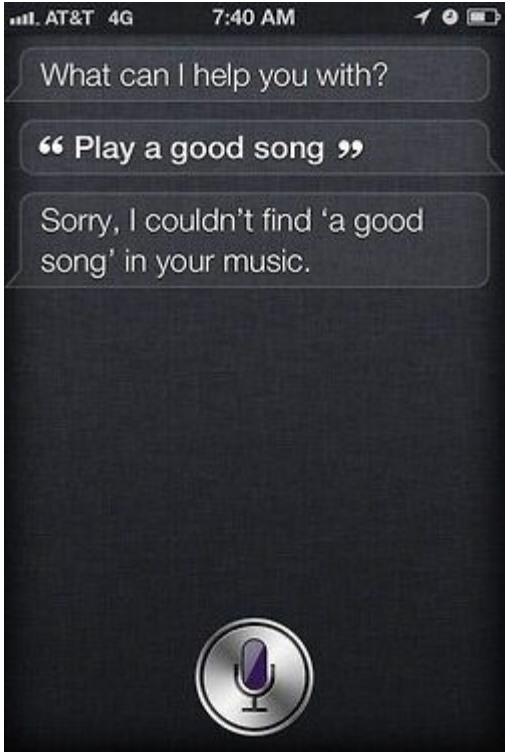
- ~1980s



```
EDIT       BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31                    Columns 00001 000
Command ===>                                                           Scroll ===> HA
****** ********************************* Top of Data ******************************
/* REXX EXEC ***********************************************************************
/*
/* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
/*
/* AUTHOR: PAUL GAMBLE
/* DATE: OCT 1/2007
/*
/*
/**********************************************************************************


say '************************'
say 'Welcome Coffee drinker.'
say '************************'
DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
     say ""
     say "What is the price of your coffee?",
         "(e.g. 1.58 = $1.58)"
     parse pull CoffeeAmt
END

DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
     say ""
     say "How many coffees a week do you have?"
     parse pull CoffeeWk
END

DO WHILE DATATYPE(Rate) \= 'NUM'
     say ""
     say "What annual interest rate would you like to see on that money?",
         "(e.g. 8 = 8%)"
     parse pull Rate
END
Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

NLP: Communication with machines

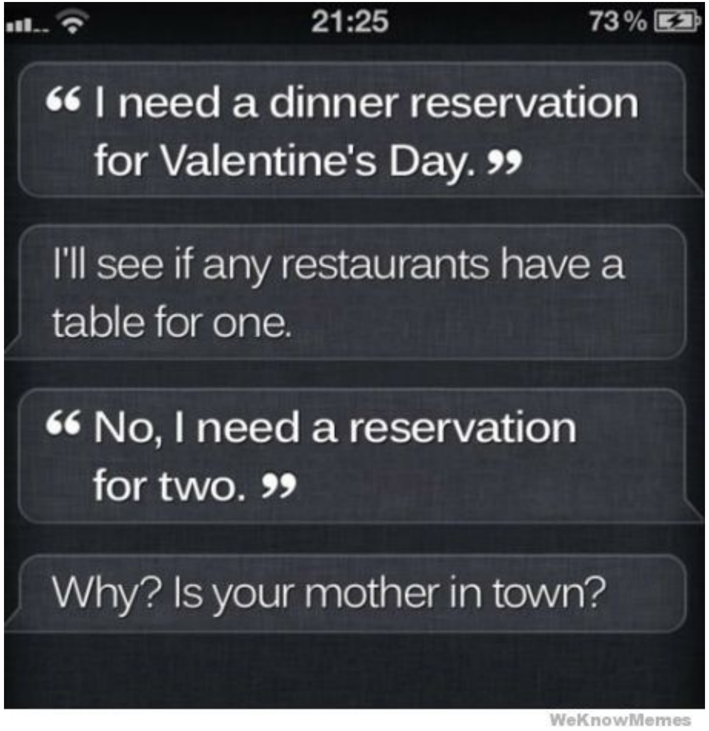
- Today



What can I help you with?

“ Play a good song ”

Sorry, I couldn’t find ‘a good song’ in your music.



“ I need a dinner reservation for Valentine's Day. ”

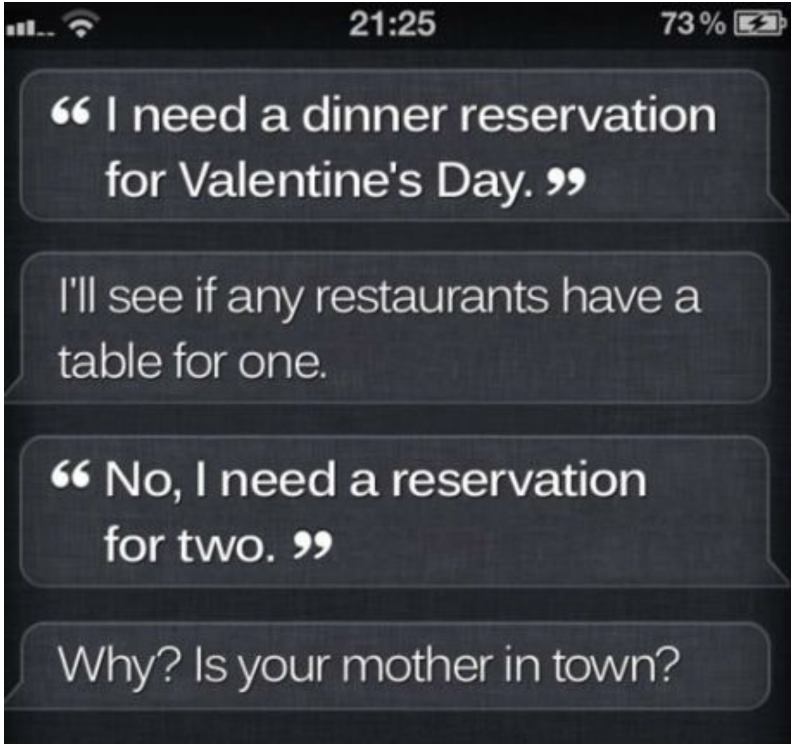
I'll see if any restaurants have a table for one.

“ No, I need a reservation for two. ”

Why? Is your mother in town?

Language technologies

What technologies are required to write such a program?



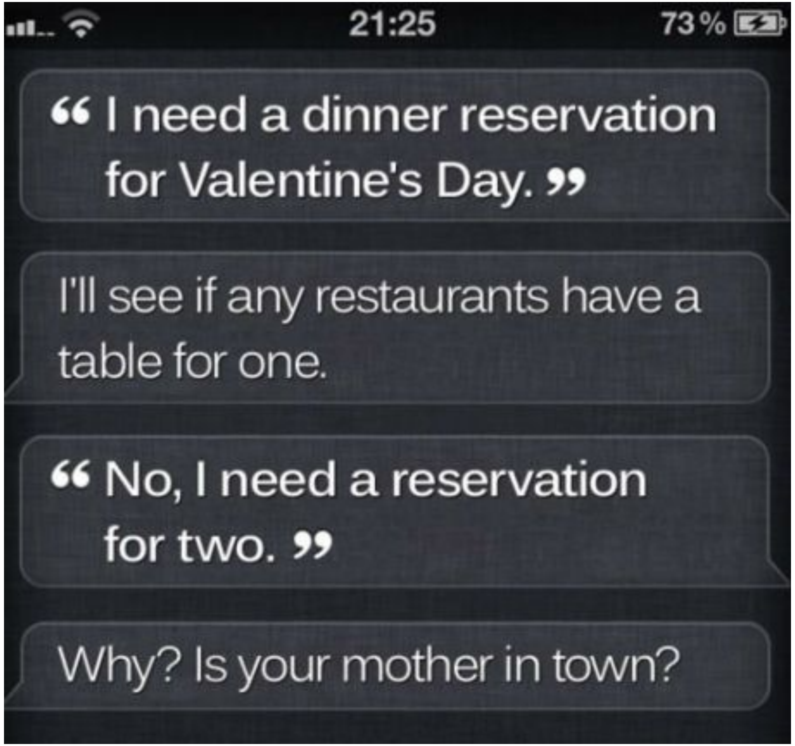
“I need a dinner reservation for Valentine's Day.”

I'll see if any restaurants have a table for one.

“No, I need a reservation for two.”

Why? Is your mother in town?

Language technologies



“ I need a dinner reservation for Valentine's Day. ”

I'll see if any restaurants have a table for one.

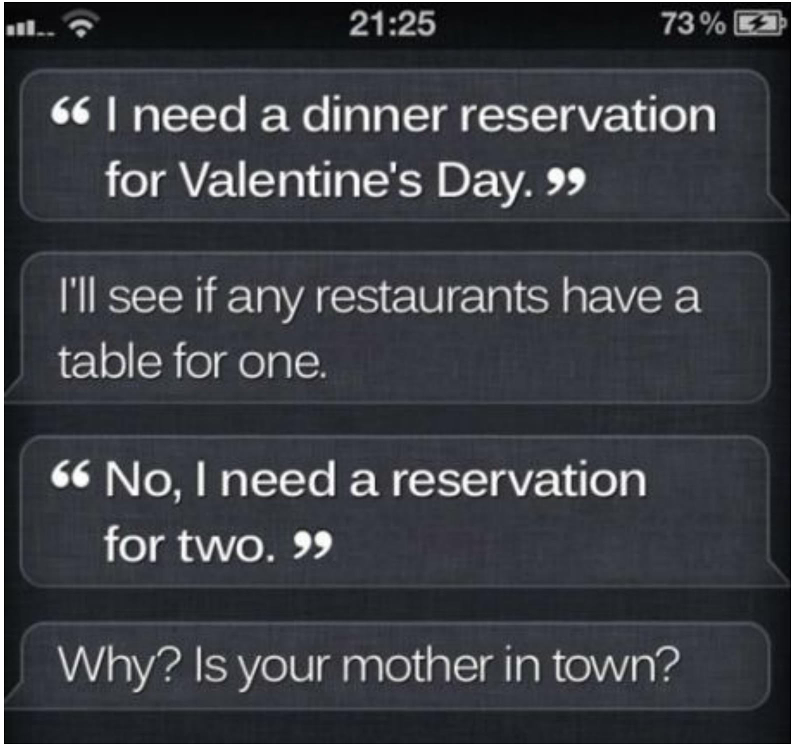
“ No, I need a reservation for two. ”

Why? Is your mother in town?

A conversational agent contains

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

Natural Language Processing



“ I need a dinner reservation for Valentine's Day. ”

I'll see if any restaurants have a table for one.

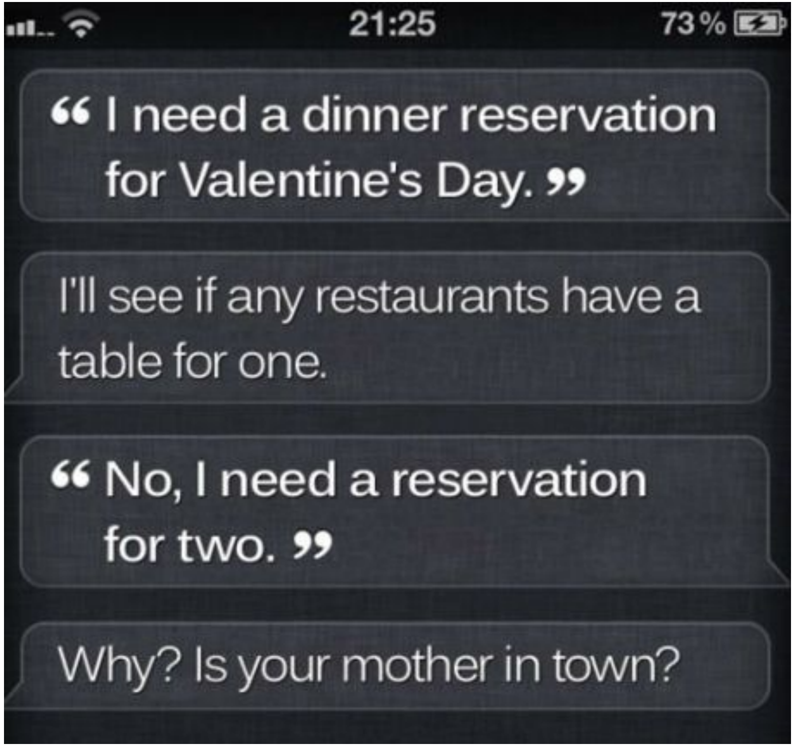
“ No, I need a reservation for two. ”

Why? Is your mother in town?

A conversational agent contains

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

Natural Language Processing



"I need a dinner reservation for Valentine's Day."

I'll see if any restaurants have a table for one.

"No, I need a reservation for two."

Why? Is your mother in town?

A conversational agent contains

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

What makes this difficult?

> Every fifteen minutes a woman in this country gives birth.
>
> Our job is to find this woman, and stop her!
>
> – Groucho Marx

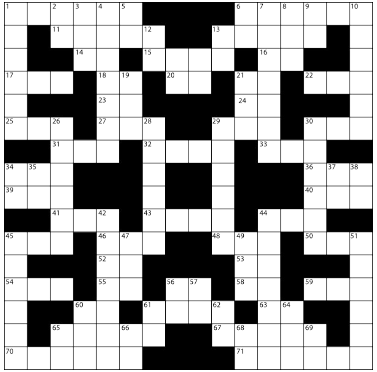
What makes this difficult?

A ship-shipping ship, shipping shipping-ships.

What makes this difficult?

Ambiguity… but also

- Richness: any meaning can be expressed in many ways
- Linguistic diversity across languages, dialects, genres, styles…
- Appropriateness of a representation depends on the context



102 Down: Say “..., say,” say

Say X, say,

where X is “..., say”

Connections between NLP and other fields

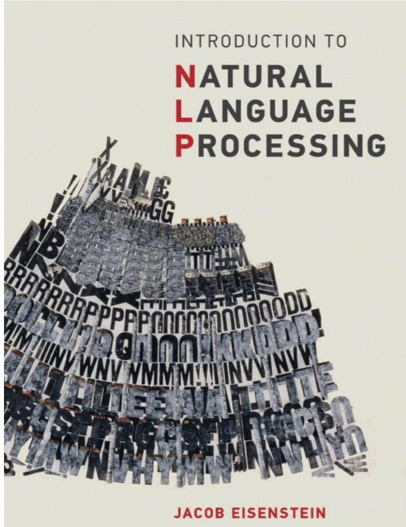
- Machine Learning
    - is about building programs from examples
    - Many of the tools we use are drawn from machine learning
- Linguistics
    - is about how language works
    - NLP must contend with natural language data as found in the world
- Artificial Intelligence
    - aims to automate human mental capacities
    - Language is a fundamental part of human mental function

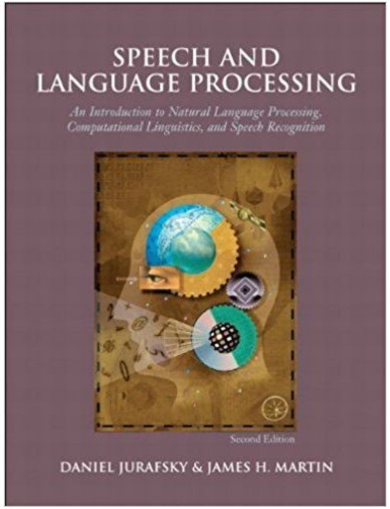
Logistics

Syllabus https://courses.cs.washington.edu/courses/cse447/23wi/

- 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

Readings




SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

- 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 on website

Course website

- https://courses.cs.washington.edu/courses/cse447/23wi/
- Office hours (TBA), announcements, calendar, etc.
    - Posted office hours will start next week

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 long, released at the end of class on Wednesdays and available for 12 hours
  - 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 of 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 A1 and A2.

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!
- The schedule builds in an extra week for each homework assignment
  - Homework is due every three weeks, but each assignment is designed to be completed in two
  - Start early!
- We will not grant any extensions beyond these

Communication with instructors

- You should be able to see yourselves be added to the Ed discussion board of CSE 447 - 23 wi. **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 also benefit 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 course staff directly.

Class participation

- **In-person** instruction!
  - (unless I’m sick or something)
- Lectures and homework assignments complement each other
  - Homework assignments will go deeper into three important topics
- Try to attend the lectures
  - But if you miss a lecture – you can read assigned book chapters and watch the recordings
- 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

- Sofia
  - Questions about lectures, research, NLP in general, and course logistics

Questions about homework assignments:

- Urmika, Daksh, Leo, Leroy, Velocity, Khirod
- (Likely) at least some office hours every weekday (TBD)

Section

There will be no section on Thursdays (or any other day of the week)!

Instead: occasional video recordings giving more details about how to do something (e.g., access course hardware)

Quizzes

- 8 quizzes, students can drop 3
- Each quiz has ~5 simple multiple-choice questions, autograded
- Quizzes are on Canvas, open immediately following lecture time for 12 hours
- Quiz time - should take about 10 minutes
- Starting from the 3rd week
- Grading on 5 best quizzes, 2% each

Academic Integrity

- To make sure that we’re all on the same page about what constitutes cheating in the context of this course, we have an Academic Integrity Form (see Canvas)
    - Your job: read and sign it
- Please let me know if you have any questions about this!

Course registration

- The instructor cannot generate an Add Code
- If you wish to register for the course and have completed prerequisite courses (or equivalents)
  - Fill out the petition form linked on https://www.cs.washington.edu/students/ugrad/non-major-registration that applies to you
  - Email Pim Lustig <pl@cs.washington.edu> and Ugrad Adviser <ugrad-adviser@cs.washington.edu> to request an Add Code
  - Cc Sofia

What background do I need to have?

- 447 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
- Not sure about your specific case?
    - We'll be releasing assignment 1 on Friday so that you can take a look through it early and get a sense of what's required for this course
    - Ask me!

More course logistics

We care that you learn!

Your questions are always welcome.

https://courses.cs.washington.edu/courses/cse447/23wi/

We also have an anonymous google form (linked on the course site) for submitting feedback.

Your To-Dos:

- Academic integrity form (https://canvas.uw.edu/courses/1610941/assignments/7962742)
  - Read it and ask me any questions you may have about it
  - Sign it and upload it to Canvas by next Friday (January 13)
- Make a GitLab account for yourself (if you don’t already have one)
  - gitlab.cs.washington.edu
- Let us know if you still need access to Ed or Canvas
  - https://edstem.org/us/courses/32306/
  - https://canvas.uw.edu/courses/1610941

Questions?