CS 333: Natural Language Processing

Fall 2022

Prof. Carolyn Anderson

Wellesley College

My Work in NLP

Large Language Models

Solving and Generating NPR Sunday Puzzles with Large Language Models
J Zhao, CJ Anderson
arXiv preprint arXiv:2306.12255

Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis) information in a Large Language Model
S Nguyen, CJ Anderson
Proceedings of the Society for Computation in Linguistics 6 (1), 400-402

ProSPer: Probing human and neural network language model understanding of spatial perspective
T Masis, C Anderson
Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting …

Text-to-Code Models

Knowledge Transfer from High-Resource to Low-Resource Programming Languages for Code LLMs
F Cassano, J Gouwar, F Lucchetti, C Schlesinger, CJ Anderson, ...
arXiv preprint arXiv:2308.09895

StarCoder: may the source be with you!
R Li, LB Allal, Y Zi, N Muennighoff, D Kocetkov, C Mou, M Marone, C Akiki, ...
arXiv preprint arXiv:2305.06161

StudentEval: A Benchmark of Student-Written Prompts for Large Language Models of Code
HML Babe, S Nguyen, Y Zi, A Guha, MQ Feldman, CJ Anderson
arXiv preprint arXiv:2306.04556

SantaCoder: don't reach for the stars!
LB Allal, R Li, D Kocetkov, C Mou, C Akiki, CM Ferrandis, N Muennighoff, ...
arXiv preprint arXiv:2301.03988

MultiPL-E: a scalable and polyglot approach to benchmarking neural code generation
F Cassano, J Gouwar, D Nguyen, S Nguyen, L Phipps-Costin, D Pinckney, ...
IEEE Transactions on Software Engineering

Computational Linguistics

Guess who's coming (and who's going): Bringing perspective to the rational speech acts framework
CJ Anderson, BW Dillon
Proceedings of the Society for Computation in Linguistics 2 (1), 185-194

Tell me everything you know: a conversation update system for the rational speech acts framework
CJ Anderson
Proceedings of the Society for Computation in Linguistics 2021, 244-253

Natural Language Processing

Human language

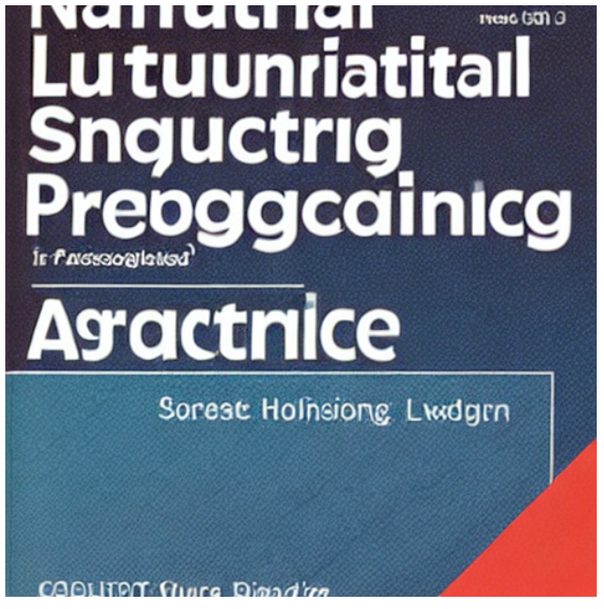
Natural Language Processing

Doing stuff with language data

Is NLP AI?

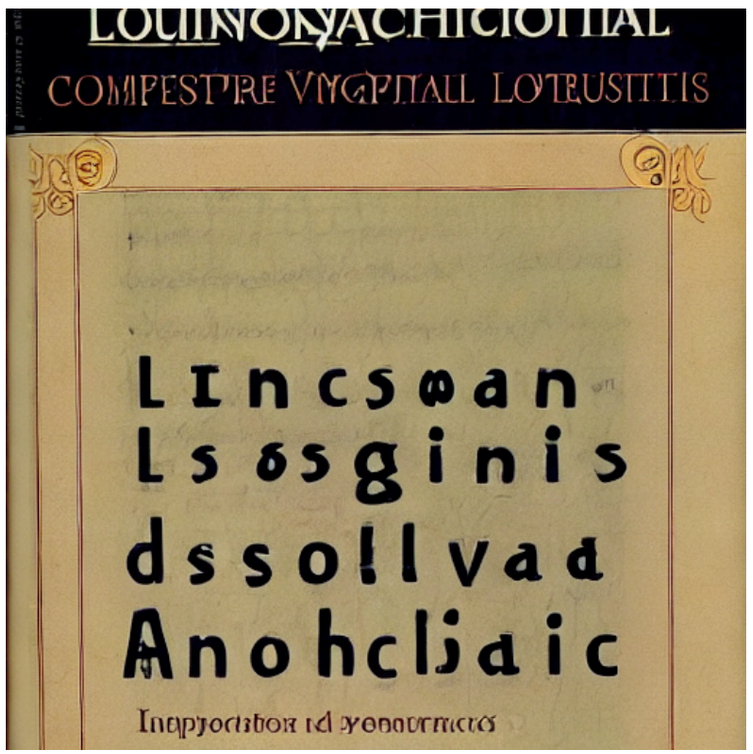
Artificial Intelligence

Natural Language Processing

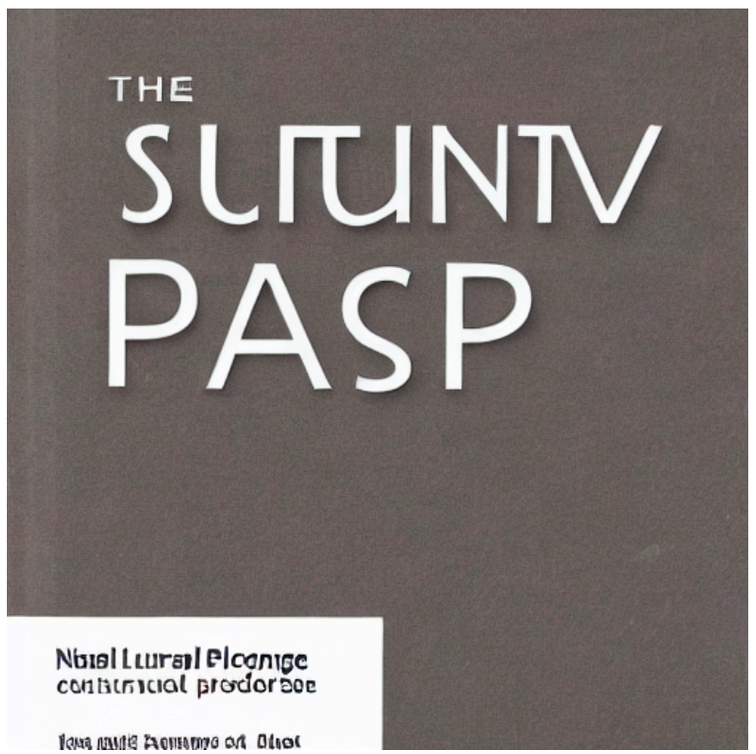


Is NLP Computational Linguistics?

Computational Linguistics

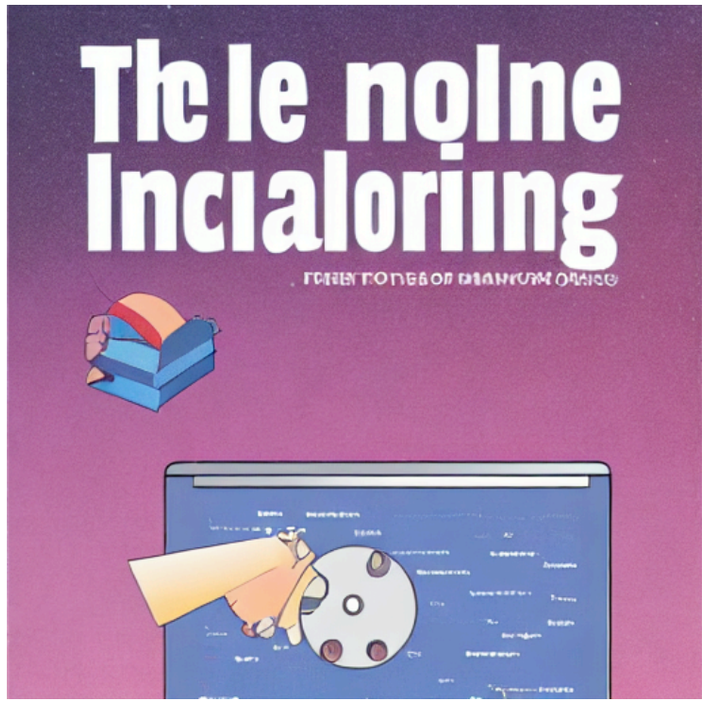


Natural Language Processing

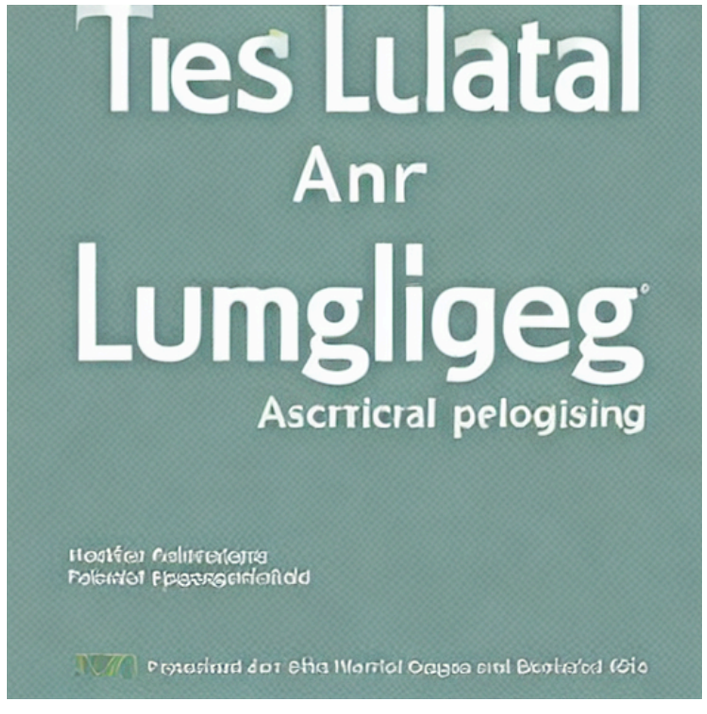


Is NLP Machine Learning?

Machine Learning



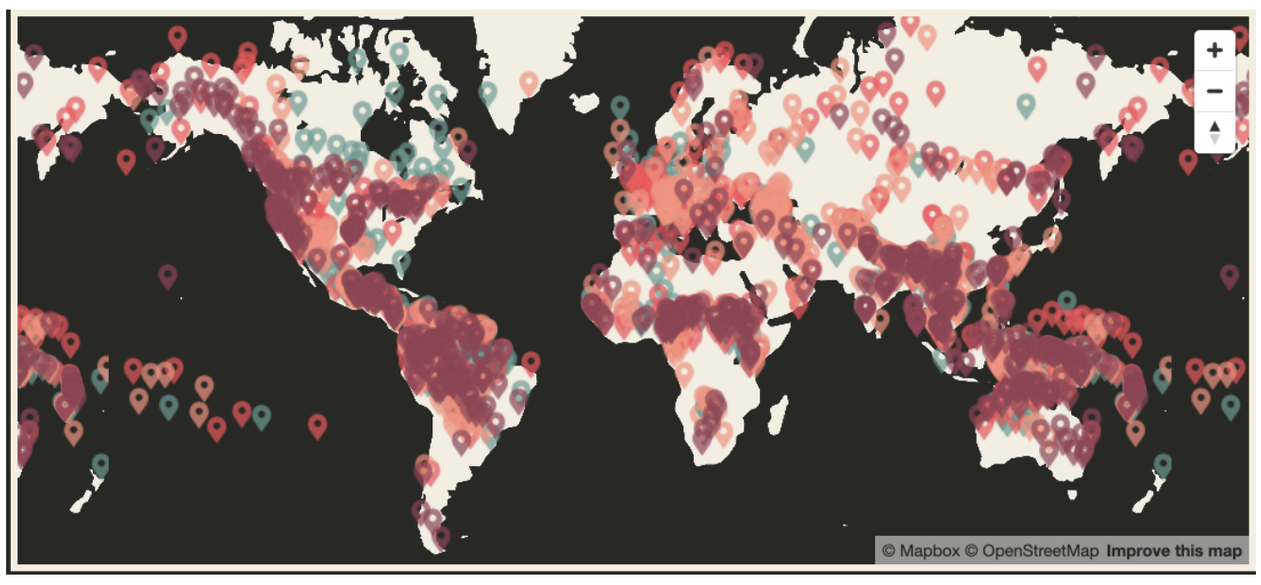
Natural Language Processing



Natural Language

Natural Language

- There are around 7000 human languages
- 50% of the world's languages are **endangered**
- Languages can be spoken or signed



https://interactive.howwegettonext.com/endangeredlanguages/

Layers of Linguistic Representation

Up Top: Pragmatics

B asks a yes-no question, but A does not respond with yes or no.

Up Top: Pragmatics

A literally says: it is 7 in the morning.

Up Top: Pragmatics

A implies: it's way too early for ice cream.

Up Top: Pragmatics

Pragmatics: the meaning of sequences of sentences.

Next Layer: Semantics

How do we know what B's question means?

Next Layer: Semantics

What does the sequence of words *wanna go get ice cream* mean?

Next Layer: Semantics

Too complicated to explain here-- go take semantics!

Next Layer: Semantics

What does the sequence of words *it's 7 am* mean?

Next Layer: Semantics

[[it's 7am]] = NOW(7am)

Next Layer: Semantics

$[[\text{it's 7am}]]^{c,w} = \text{True if } w_t == \text{7am else False}$

Next Layer: Semantics

Basically: *it's 7am* is a **function** that takes a **world** and returns true for some worlds and false for others.

Next Layer: Semantics

Semantics: the **meaning** of a sentence is its **truth conditions** (the conditions under which it is true).

Middle Layer: Syntax

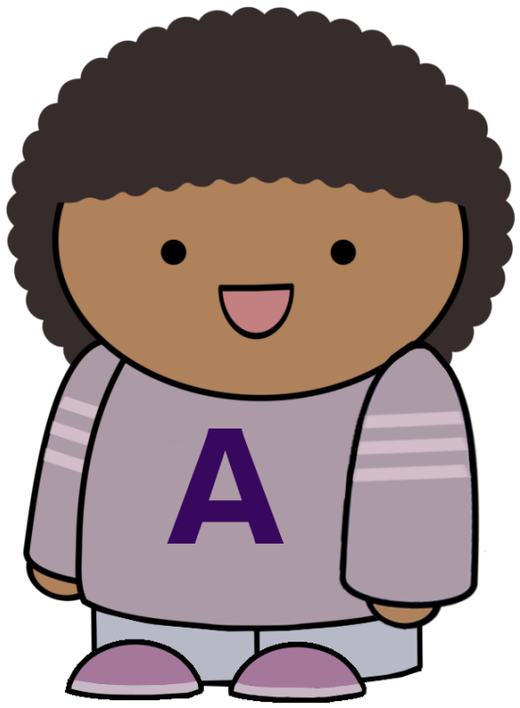
How do we determine the order of the words?

Middle Layer: Syntax

How do we determine the order of the words?

Middle Layer: Syntax

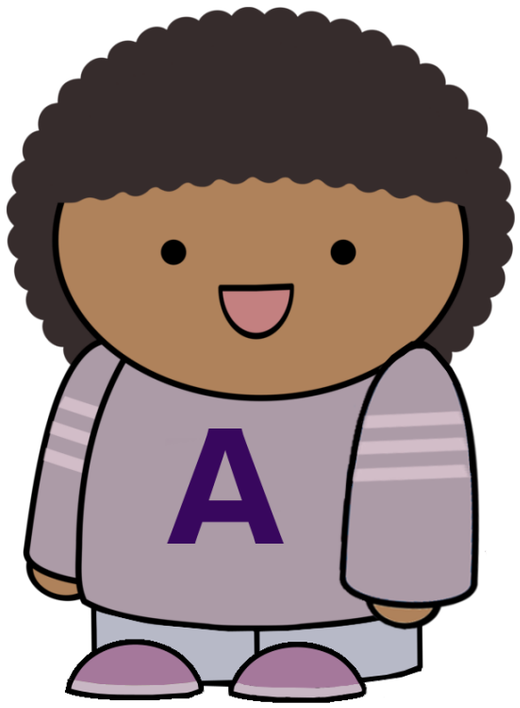
Must be German...



How do we determine the order of the words?

Middle Layer: Syntax

Must be German...



Syntax: the structure of a sentence is determined by a set of language-specific syntactic rules.

Glue Layer: Morphology

Morphology: the rules that determine how words are formed.

Lower Layer: Phonology

Phonology: the rules that determine how the sounds/signs of a language are organized

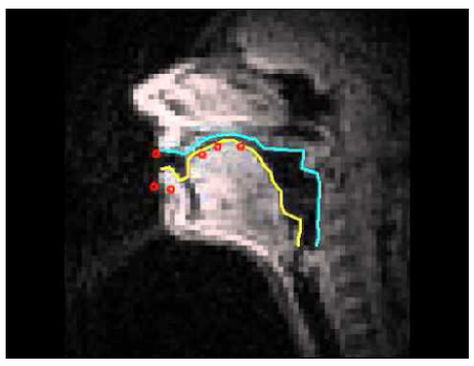
Lower Layer: Phonology

Phonology: the rules that determine how the sounds/signs of a language are organized

Foundation: Phonology

Phonetics: how do language users produce the building blocks of language?

Layers of Linguistic Abstraction

PRAGMATICS	It's 7am ⟹ it's a weird time for ice cream so I don't know how to respond.
SEMANTICS	[[It's 7am]] → True if now(w) == 7am else False
SYNTAX	{7, it's, am} → It's 7am.
MORPHOLOGY	{7 → 7, [it is] → it's, am → am}
PHONOLOGY	ɪts ˈsɛvən ə ɛm.
PHONETICS	

Natural and Artificial Language Learning

How do people learn language?

Humans learn language instinctively:

- Language has a **critical acquisition period**
- Language acquisition begins before birth and follows predictable developmental stages
- Humans can't decide **not** to learn language
- Language acquisition does not seem to correlate with intelligence
- All human cultures have language; no other species do
- All human languages are equally expressive

Example: Child Language Acquisition

Example 1

cj and ember manning liked

Gareth Roberts @garicgymro · 45m

Just overheard from two of my kids:
Osian (5;1): Look how I catched Mickey!
Eirwen (8;2): Do you mean caught?
Osian: … yeah.
Eirwen: But you can keep saying catched!
Osian: Look how I catched him!

Example 2

Learning Language

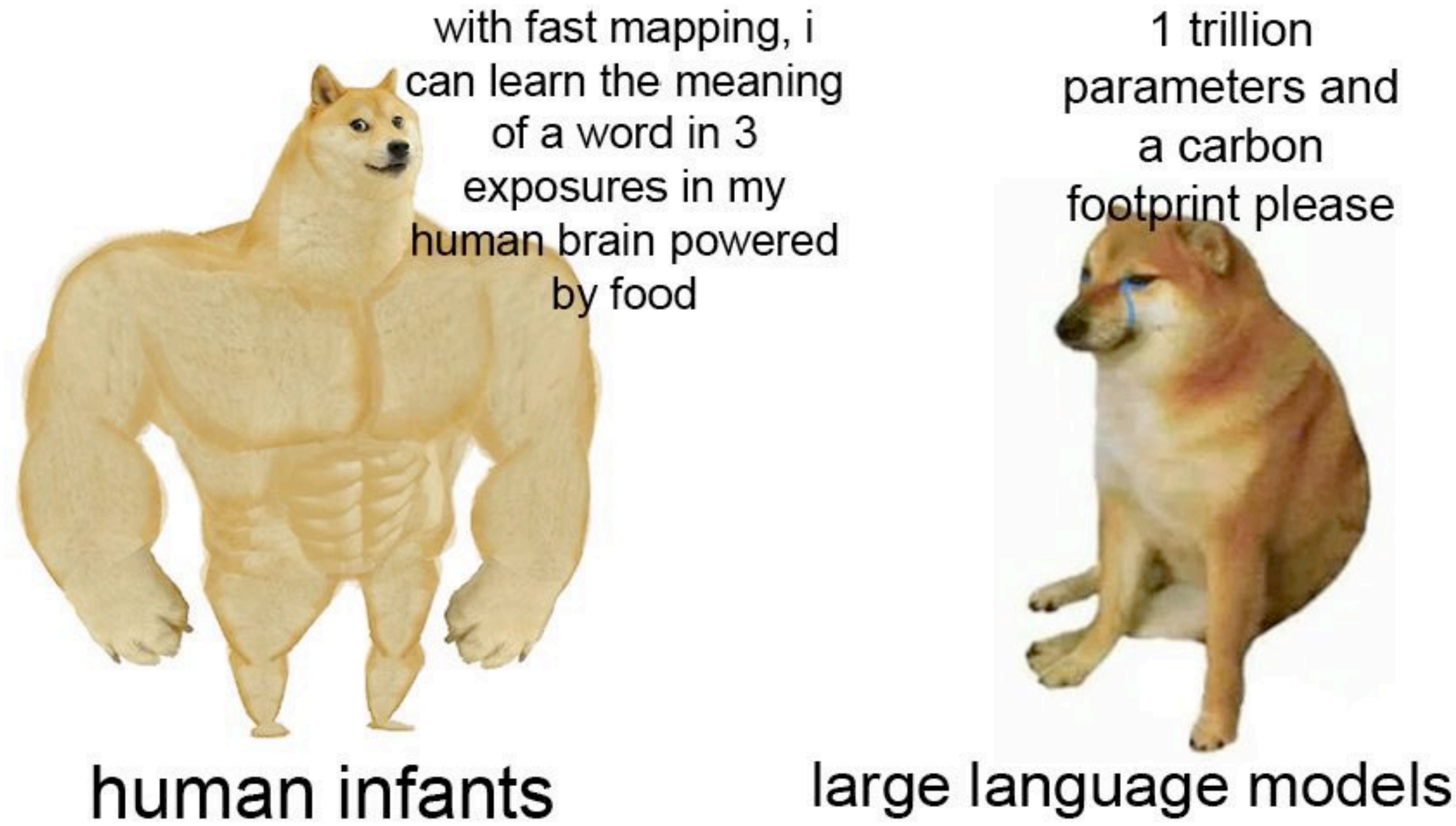


photo credit: Josef Fruehwald

Practicalities

Schedule

- Room: SCI L039
- Lecture: 8:30-9:45 on Tuesdays and Fridays
- Assignments are due on **Thursdays at 10 PM**

Help Hours

- Mondays 4-5:30pm
- Thursdays 4:30-5:30pm
- By appointment (schedule using online calendar)

Come to my help hours to …

- Get help with CS333
- Talk about NLP

Cynthia Wang

Tutor

Readings

This course has required weekly readings. Most are from the course textbook: *Speech and Language Processing* by Jurfasky & Martin. The third edition is free online.

All readings are listed on the schedule.

Please finish each week's required reading before coming to class on Tuesday.

Quizzes

There will be a **quiz** in class **every Tuesday** to test your understanding of the assigned reading.

+ open note, closed computer
+ timed but very brief (1-2 questions)

Homework will be in Python

I recommend setting up a Python 3.8 virtual environment.

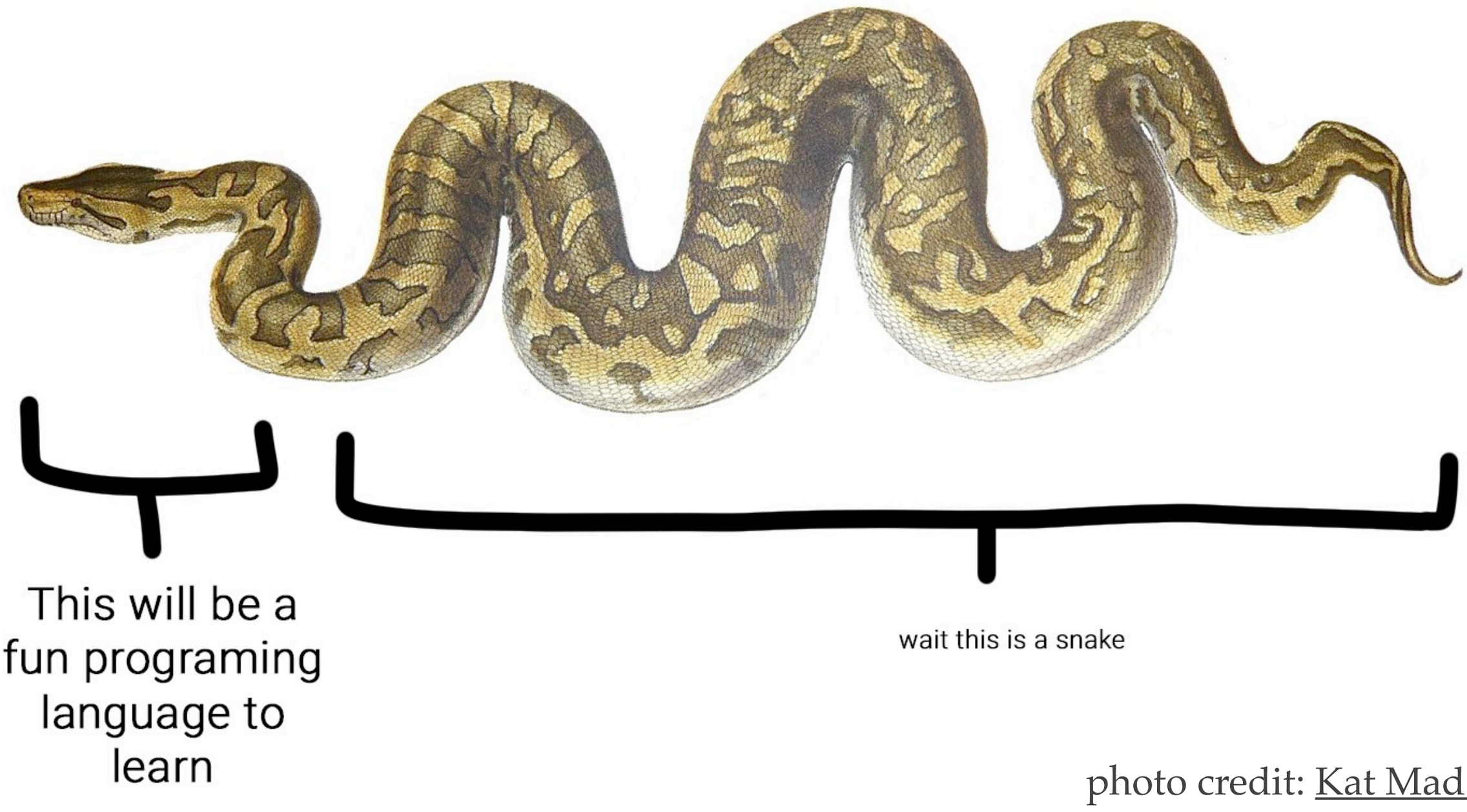


photo credit: Kat Maddox

Assignments

- Assignments are due on **Thursdays at 10 PM**
- Homework submission will be through Gradescope.
- There are 10 weekly assignments.
- **HW 0** is due this Thursday.

Late Policy

You have **5 late days** for the semester, which you can use all at once, or spread across assignments. **I will not accept late work beyond these days**.

Important: I will not answer questions on late work during help hours.

If you have a prolonged illness or unexpected circumstance, let me know and **we'll work together** to make a **custom plan**.

Collaboration policy

In this class, you can talk at a high-level with other students about assignments, but **you cannot show them your code**.

If you discuss a homework problem with another student, **please note this** on your assignment when you submit it.

You may not use ChatGPT, Bard, Codex or any other AI system unless explicitly stated in the homework assignment.

Midterms and Final Paper

- **Midterm 1**: in-class programming exam on Oct. 20th
- **Midterm 2**: in-class paper exam on Nov. 17th
- **Final paper** on a research topic of your choice due at the end of term

Interested in Note-Taking?

There is a student in our class who requires the service of a note taker. If you take accurate and legible notes, please apply for this position at https://shasta.accessiblelearning.com/wellesley/.

This is a paid position.

Course Goal

To make you into a skilled NLP practioner who can:

- Understand and implement core NLP algorithms and models.
- Explore the challenges posed by different aspects of human language.
- Analyze ethical concerns about language technology.
- Complete a series of projects to implement and improve NLP models.

FALL 2023 WELLESLEY COLLEGE THEATRE MAINSTAGE PRODUCTION

R.U.R.

by Karel Čapek

Adapted by
Marta Rainer

Auditions

WEDNESDAY 9/6 | 6-8:30PM
THURSDAY 9/7 | 6-8:30PM
RUTH NAGEL JONES THEATRE

SEEKING 10 ENSEMBLE MEMBERS: WE WILL NEED YOU, BRAVE HUMANS OF WELLESLEY, TO PERFORM - AND POTENTIALLY SAVE HUMANITY. OPEN TO ALL!

MAKE THEATRE, COLLABORATE AND BOND WITH YOUR PEERS AND EARN 1.0 ACADEMIC CREDIT - WITH THE SUPPORT OF A TEAM OF INDUSTRY PROFESSIONALS!

THST 345 REHEARSES M/T/W 6:30-9:30
& TECH & PERFORMANCES DEC 7-10, 2023

Next class: text processing

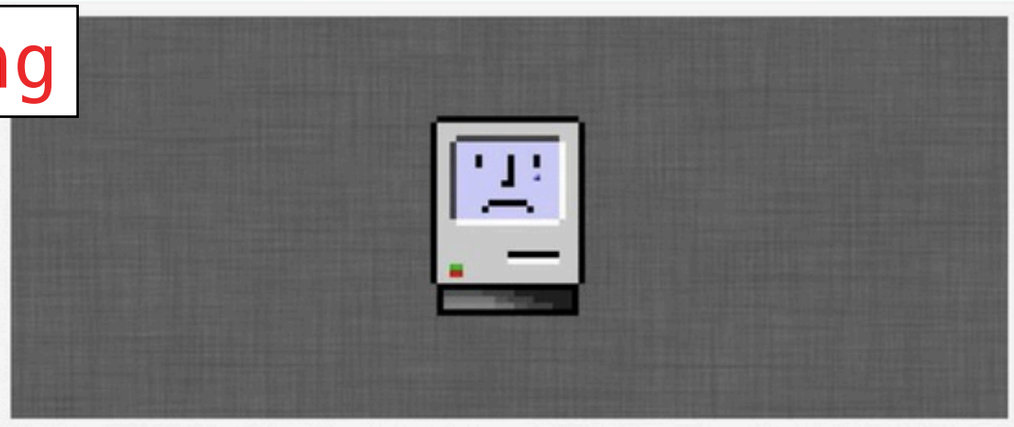
REMOTE on Zoom
(I'll be at a conference)

Natural Language Processing

is mostly

DATA WRANGLING

Wrangling data is frustrating



Can we make it fun?

No.

OK but can we eliminate frustration?

Also no.

Sorry.