CS 159: Natural Language Processing

or

A journey from **words** to **language** to **meaning**

Prof. Jonathan P. Chang - Fall 2025

A journey from **words** to **language** to **meaning**

Everyone thinks NLP is about these…

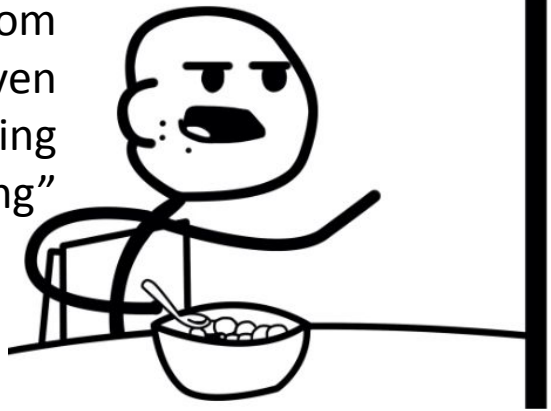
A journey from **words** to **language** to **meaning**

…but I think it's really about this!

A journey from **words** to **language** to **meaning**

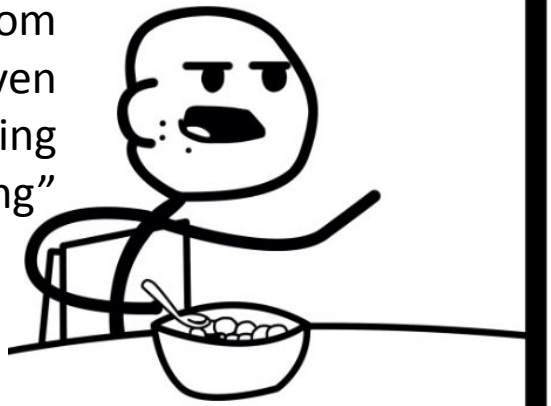
(Your Prof in 2020)

"NLP is still far from anything even approximating meaning"

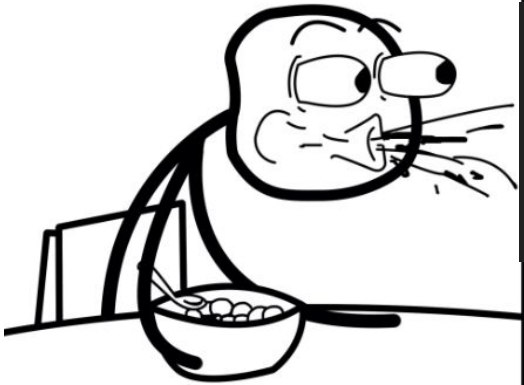


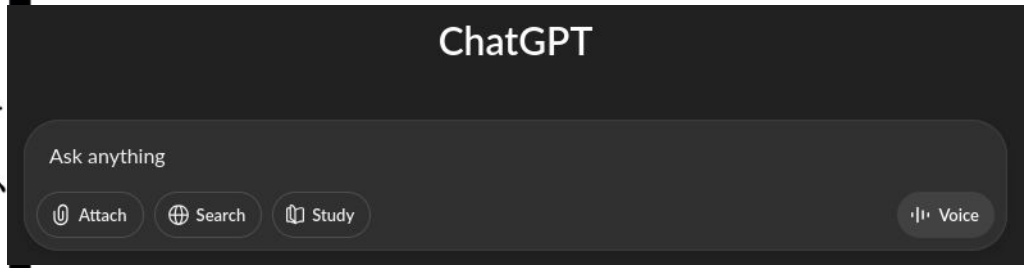
(Your Prof in 2020)

"NLP is still far from anything even approximating meaning"



(Your Prof in 2025)





Colin Fraser's "Dumb Monty Hall" example (probably outdated, but you get the point)

(Your Prof in 2025)

Monty hall offers you the opportunity to pick between three doors. Behind one of them is a new car, and behind the other two are goats. The doors are transparent and you're able to see clearly that the car is behind door number one. He asks you to choose a door, and you select door number one, since you want the car. He opens door number two and shows that there is a goat behind it. Then you have the opportunity to stay with your original choice, or switch doors. What should you do?

You should switch doors.
The probability of the car being behind door 1 is 1/3. After the host opens door number 2 and you know that a goat is there, the probability of the car being behind door 1 does not change, it is still 1/3. But now you know that the car is behind door 3 with probability of 2/3. So you will have a higher chance of winning the car by switching your door selection.

(this is complicated)

A journey from **words** to **language** to **meaning**

LIVE

BREAKING NEWS

ALIENS LAND IN CLAREMONT, CA

14:10
imgflip.com

SPACECRAFT REPORTEDLY TRAVELED 42 LIGHT-YEARS

Humans! What does this *mean*?

Humans! What does this *mean*?

RUMACK: You'd better tell the Captain we've got to land as soon as we can. This woman has to be gotten to a hospital.

ELAINE: A hospital? What is it?

RUMACK: It's a big building with patients, but that's not important right now.

From "Airplane!" (1980)

Activity, Part 1 (~15 mins)

- Form groups of 3-4
- Do some quick introductions (name, year, home college, why you're taking this class)
- Try to come up with some explanation you can give to the aliens to help them understand the movie scene you just watched
- Base assumptions:
  - The aliens don't speak any human languages (but your explanation will be appropriately translated to them using, IDK, telepathy or something I'm not a Hollywood writer)
  - The aliens also lack any notion of human culture; on their planet they do not have movies (but they understand the general concept of entertainment)
  - You can stay at a pretty high level (e.g., you can just say that in your explanation you need to define some specific words, but don't need to be particular about how the definition looks)
- Then, try to break up your explanation into chunks (e.g., "define what an airplane is") and write them down on post-it notes

Activity, Part 2 (~10 mins)

- Now it’s time to mingle! Go around and share your results with other groups (don’t forget to introduce yourselves!)
- If you and another group have a post-it note that is “similar”, put them up on the board together
- You can start by defining “similar” as “identical” (e.g., you both said “define what a hospital is”), but as post-it notes start going up we may start to see higher-level common themes (e.g., “define hospital” and “define patient” seem related), so feel free to move the post-it notes around and create bigger and bigger clusters

(this is complicated)

A journey from **words** to **language** to **meaning**

(that’s why we start from here)

A journey from **words** to **language** to **meaning**

Class Goals:

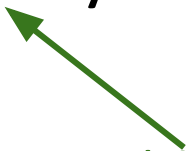
- Implement classic NLP algorithms
- Use appropriate metrics of evaluation
- Read primary literature
- Critique assumptions and design choices
- Practice scholarly skills

A journey from **words** to **language** to **meaning**

Class Goals:

- Implement classic NLP algorithms
- Use appropriate metrics of evaluation
- Read primary literature
- Critique assumptions and design choices
- Practice scholarly skills

A journey from **words** to **language** to **meaning**



(that's this!)

A journey from **words** to **language** to **meaning**

Theme of the next 2 weeks: this clip, but replace “birds” with “words”

What are words?

What are words? Attempt #1

A word is a string! Like “strawberry”.

What are words? Attempt #1

A word is a string! Like “strawberry”.

Can’t a string have multiple words? Like, the whole text of Moby Dick. But also: what’s a string?

What are words? Attempt #1

A word is a string! Like “strawberry”.

Can’t a string have multiple words? Like, the whole text of Moby Dick. But also: what’s a string?

A string is an array of characters. So “strawberry” is actually [‘s’, ‘t’, ‘r’, ‘a’, ‘w’, ‘b’, ‘e’, ‘r’, ‘r’, ‘y’].

What are words? Attempt #1

A word is a string! Like “strawberry”.

Can’t a string have multiple words? Like, the whole text of Moby Dick. But also: what’s a string?

A string is an array of characters. So “strawberry” is actually [‘s’, ‘t’, ‘r’, ‘a’, ‘w’, ‘b’, ‘e’, ‘r’, ‘r’, ‘y’].

Cool, cool. …And a character is what, exactly?

An aside on encodings

In general: an *encoding* is a mapping from A → B

In computers: a mapping from bits → data

For text: a mapping from bits → characters in a string

An aside on encodings

In general: an *encoding* is a mapping from A → B

In computers: a mapping from bits → data

For text: a mapping from bits → characters in a string

Someone needs to define this mapping!

ASCII

(American Standard Code for Information Interchange)

First published in 1963—as an evolution of telegraph codes!—by what eventually became the American National Standards Institute

01000001	01010011	01000011	01001001	01001001
A	S	C	I	I

ASCII

(American Standard Code for Information Interchange)

First published in 1963—as an evolution of telegraph codes!—by what eventually became the American National Standards Institute

01000001	01010011	01000011	01001001	01001001
A	S	C	I	I

code points:

65	83	67	73	73

Here’s ASCII. Literally all of it.

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[END OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Here's ASCII. Literally all of it.

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g

What do you think about this? Good enough for representing text in general?

Any limitations that stand out to you?

25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

An aside on encodings

In general: an *encoding* is a mapping from A → B

In computers: a mapping from bits → data

For text: a mapping from bits → characters in a string

Someone needs to define this mapping!

Any encoding is made up of decisions made by human beings.

Who gets to make those decisions? What assumptions or biases did they hold while making them?

How does that influence the final encoding, and what it can/can’t do?

Any ~~encoding~~ system is made up of decisions made by human beings.

Who gets to make those decisions? What assumptions or biases did they hold while making them?

How does that influence the final ~~encoding~~ system, and what it can/can't do?

This will come up repeatedly throughout this class!

Character sets (aka “charsets”)

- 1963: ASCII, a 128-character 7-bit standard
- 1987-2000: ISO 8859, a series of 256-character 8-bit standards
- Many languages need even more characters! So we get charsets like Big5 (2 bytes), GB 18030 and ISO 2022
- That’s a lot of charsets! Sounds like someone needs to get them all organized…

Unicode to the rescue (?)

Unicode is NOT an encoding!

It is a *listing* of code points (e.g., “65 corresponds to A”)

There are many different encodings that *implement* (subsets of) Unicode (e.g., UTF-32 is a fixed-length 32-bit encoding, UTF-16 is a variable-length encoding that uses *at least* 16 bits)

You’ll see UTF-8 a lot since it’s backwards-compatible with ASCII

Unicode makes your memes, texts, and TikTok captions possible

 “Waving Hand”: Unicode codepoint 1F44B

 “Red Heart”: Unicode codepoint 2764

 “Money With Wings”: Unicode codepoint 1F4B8

 “Bubble Tea”: Unicode codepoint 1F9CB

And more to come: Unicode is always updating!

You WILL encounter weird encoding-related errors/bugs when working with text data!

Common causes of trouble:

- Contextual characters (e.g. “a quote” for ")
- Newlines (LF, CR, or CRLF)
- Ligatures (Schofield)
- Combining characters (Z̈ä̈l̈g̈ö̈ text)

What did we learn today?

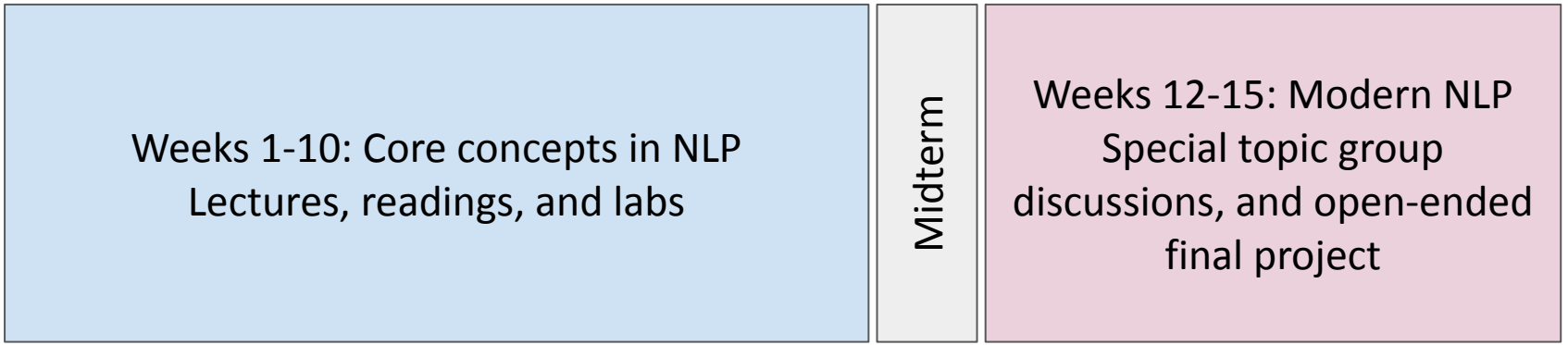
- We started with a seemingly-basic question (“what are words?”) and found ourselves down quite the rabbit hole!
- This rabbit hole is, in part, a result of *human systems* that were built to turn (messy) language into something computers can understand
- We’ll see similar rabbit holes throughout this class
- In fact, we’re not even done answering the question yet!

Ready for more?

Topics planned for the class:

- Basic NLP concepts: tokenization, segmentation evaluation
- Statistical modeling and classification of text
- Information retrieval
- Co-occurrence and vector semantics
- NLP and data ethics
- (briefly) Transformers, LLMs, and the future of NLP
- Special topics! (to be decided partly by you!)

Class structure (semester view)



Class structure (week view)

- Monday: This week’s info released on website
- Tuesday: In-class lecture; supplementary readings released (due before *next* Tuesday)
- Thursday: In-class lab (due before *next* Thursday)

Miscellany

- Lab assignments start *next* week (this Thursday we'll take syllabus questions and then do an in-class activity that's graded on completion).
- Nobody has yet been added to Gradescope or the course server. That will happen later this week or early next week, once adds/drops settle down.
- How do you want us to communicate with you? Please share your thoughts using the following form: