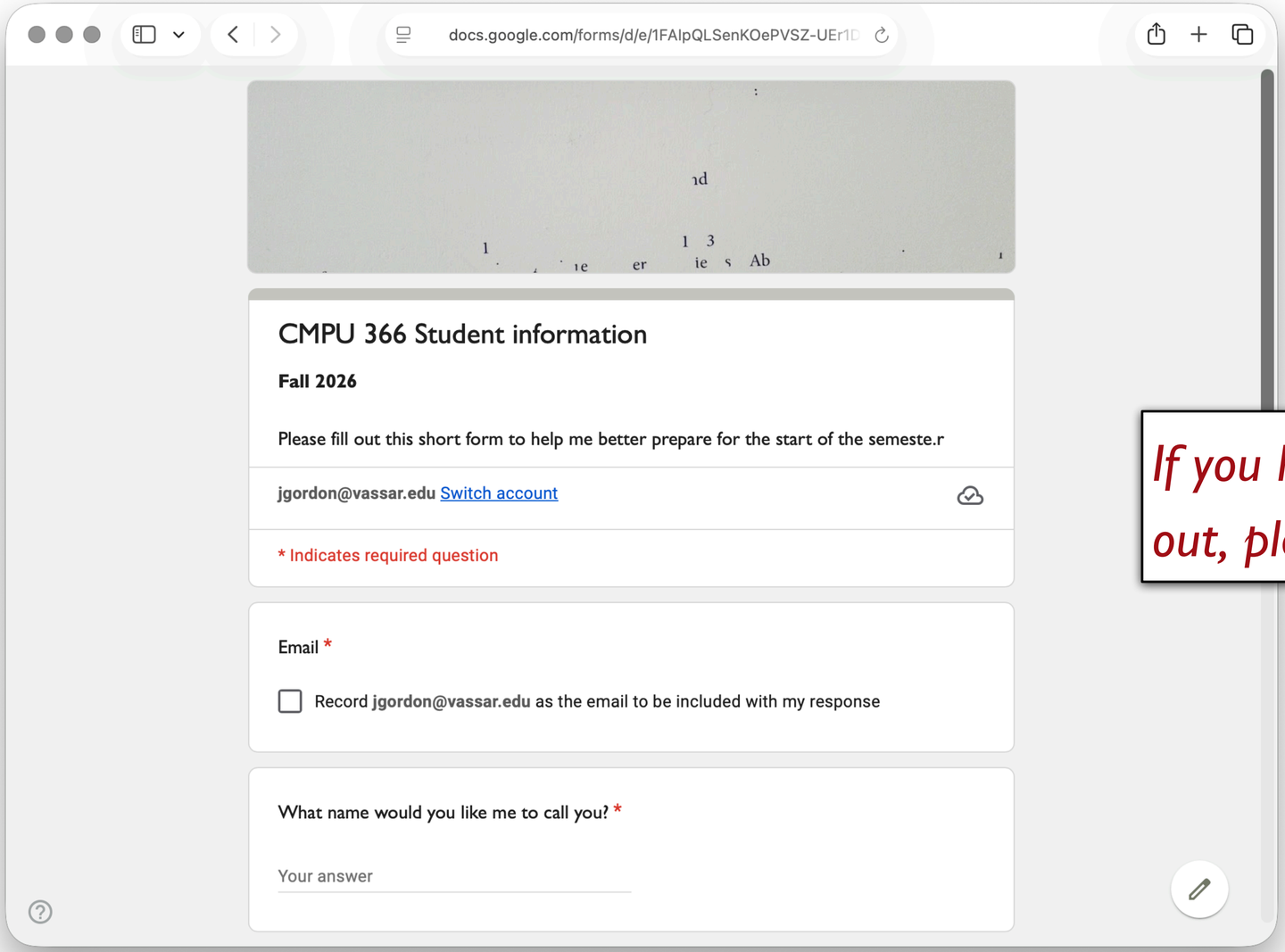
CMPU 366 · Natural Language Processing

Text as Data?

3 September 2026



If you haven't filled this out, please do

forms.gle/WLo5Q9Y9ekRFBdVn7

www.cs.vassar.edu/~cs366/

Thursday 1:30–2:45 p.m.

Sanders Classroom 006

Professor Gordon

↓ Current

1 *Introduction*

Sep. 1 Computers and language

Sep. 3 Text as data?

Assignment 0

Core reading:

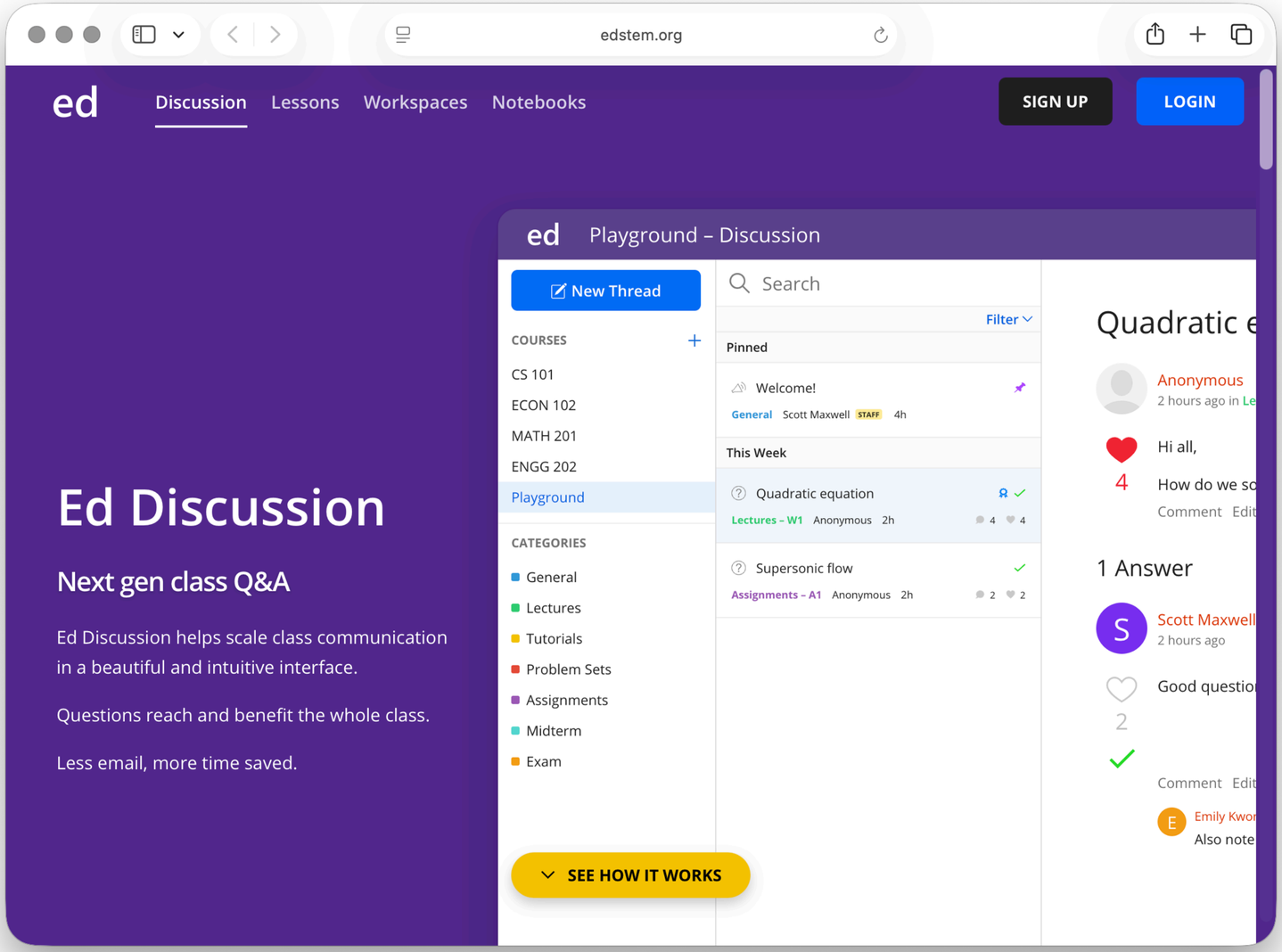
- Syllabus
- J&M 1: Introduction

Further exploration:

- Weizenbaum, 1966: ELIZA – a computer program for the study of natural language communication between man and machine
- Zentgraf, 2015: What every programmer absolutely, positively needs to know about encodings and character sets to work with text
- Erad, 2017: How the appetite for emojis complicates the effort to standardize the world's alphabets

Readings!

Assignment 0

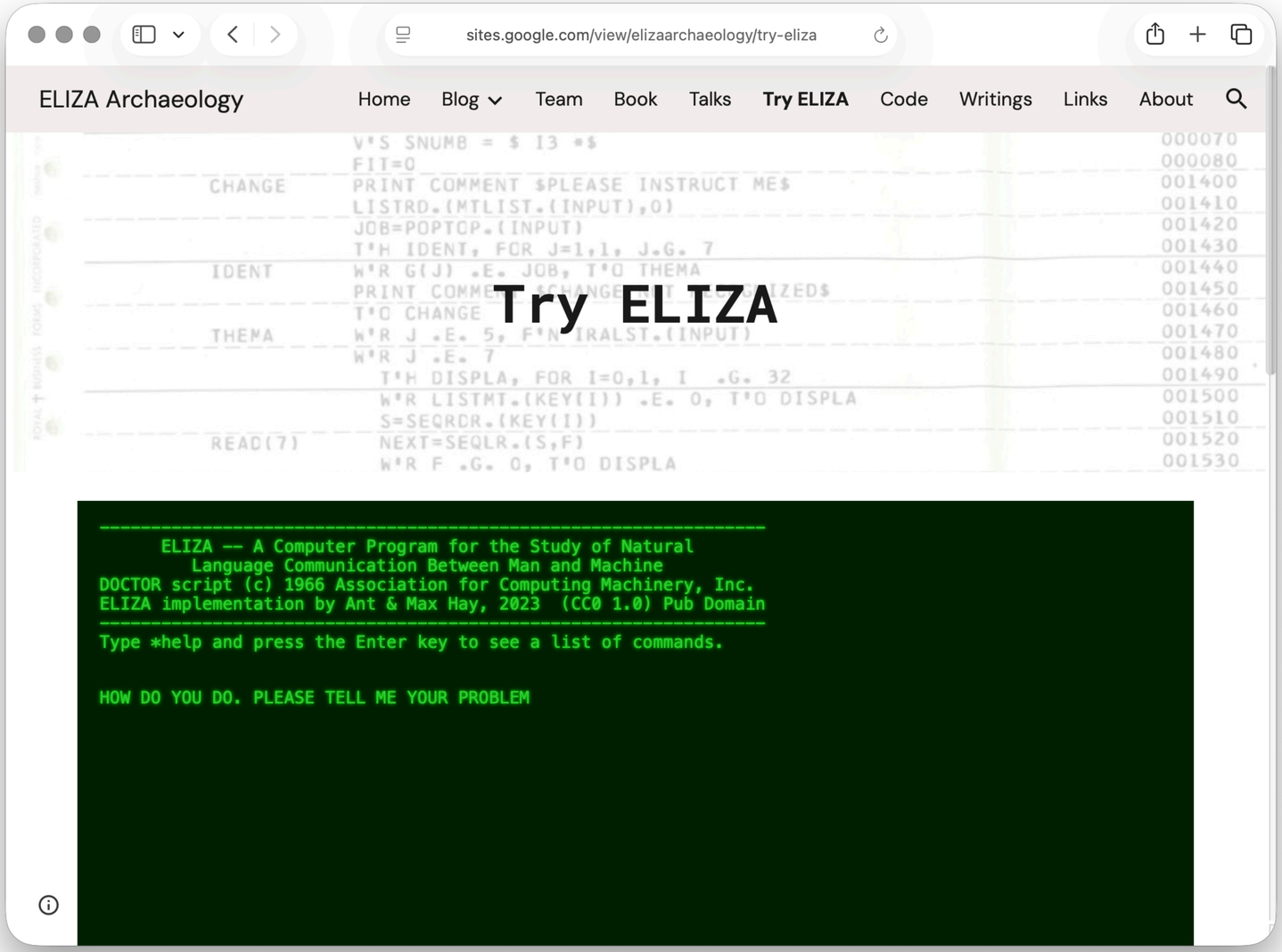


Discussion board?

Where are we?

Our textbook is called *Speech and Language Processing*, but this class will be focused on text.

Why might we do that?



sites.google.com/view/elizaarchaeology

```
        001910
POPMOR          W'R LISTMT.(KEY(I)) .E. 0, T'O WRITE
        001920
                LPRINT.(POPTOP.(KEY(I)),SCRIPT)
        001930
                T'O POPMOR
        001940
WRITE           CONTINUE
        001950
                LPRINT.(MTLIST.(INPUT),SCRIPT)
        001960
                EXIT.
        001970
               R* * * * * * * * * * SCRIPT ERROR EXIT
        001980
NOMATCH(1)      PRINT COMMENT $PLEASE CONTINUE $
        002200
                T'O START
        002210
NOMATCH(2)      PRINT COMMENT $HMMM $
        002220
                T'O START
        002230
NOMATCH(3)      PRINT COMMENT $GO ON , PLEASE $
        002240
                T'O START
        002250
NOMATCH(4)      PRINT COMMENT $I SEE $
        002260
                T'O START
        002270
                VECTOR VALUES SNUMB= $I3 * $
        002280
                E'M
        002290
```

NLP tasks: inputs and outputs

NLP *tasks* have helped break down the complex goals of NLP into well-defined inputs and outputs.

They are often accompanied by well-defined *benchmark datasets*.

Task	*Input*	*Output*

Task	***Input***	***Output***
Sentiment analysis	Sentence(s)	Classification (categories): Positive, Neutral, Negative

Task	***Input***	***Output***
Sentiment analysis	Sentence(s)	Classification (categories): Positive, Neutral, Negative
Machine translation	Text in Language 1	Text in Language 2

Task	***Input***	***Output***
Sentiment analysis	Sentence(s)	Classification (categories): Positive, Neutral, Negative
Machine translation	Text in Language 1	Text in Language 2
Coreference resolution	Sentence(s), target pronoun	A span of text in the sentence, which the pronoun refers to

Activity instructions

1. (8–10 min.) Work through *examples*, then brainstorm how you would program a computer to perform this task for *any* example.

   This is just for you; it won't be collected.

   Feel free to skip around.

Activity instructions

1. (8–10 min.) Work through *examples*, then brainstorm how you would program a computer to perform this task for *any* example.

   This is just for you; it won't be collected.

   Feel free to skip around.

2. (5 min.) Turn to a partner to discuss.

Activity instructions

1. (8–10 min.) Work through *examples*, then brainstorm how you would program a computer to perform this task for *any* example.

   This is just for you; it won't be collected.

   Feel free to skip around.

2. (5 min.) Turn to a partner to discuss.

3. (7 min.) We'll return together as a whole class.

Debrief

Corpora

Dictionary
10 found

corpus

All Apple Dictionary Thesaurus Wikipedia

- corpus
- corpus callosum
- corpus cavernosum
- Corpus Christi
- corpus delicti
- corpus luteum
- corpus spongiosum
- corpus striatum
- corpuscle
- corpuscular

cor·pus | ˈkôrpəs |

noun (*plural* **corpora** | ˈkôrp(ə)rə | or *plural* **corpuses** | ˈkôrpəsəz |)

1 a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus.*
- a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc..

2 *Anatomy* the main body or mass of a structure.
- the central part of the stomach, between the fundus and the antrum.

ORIGIN

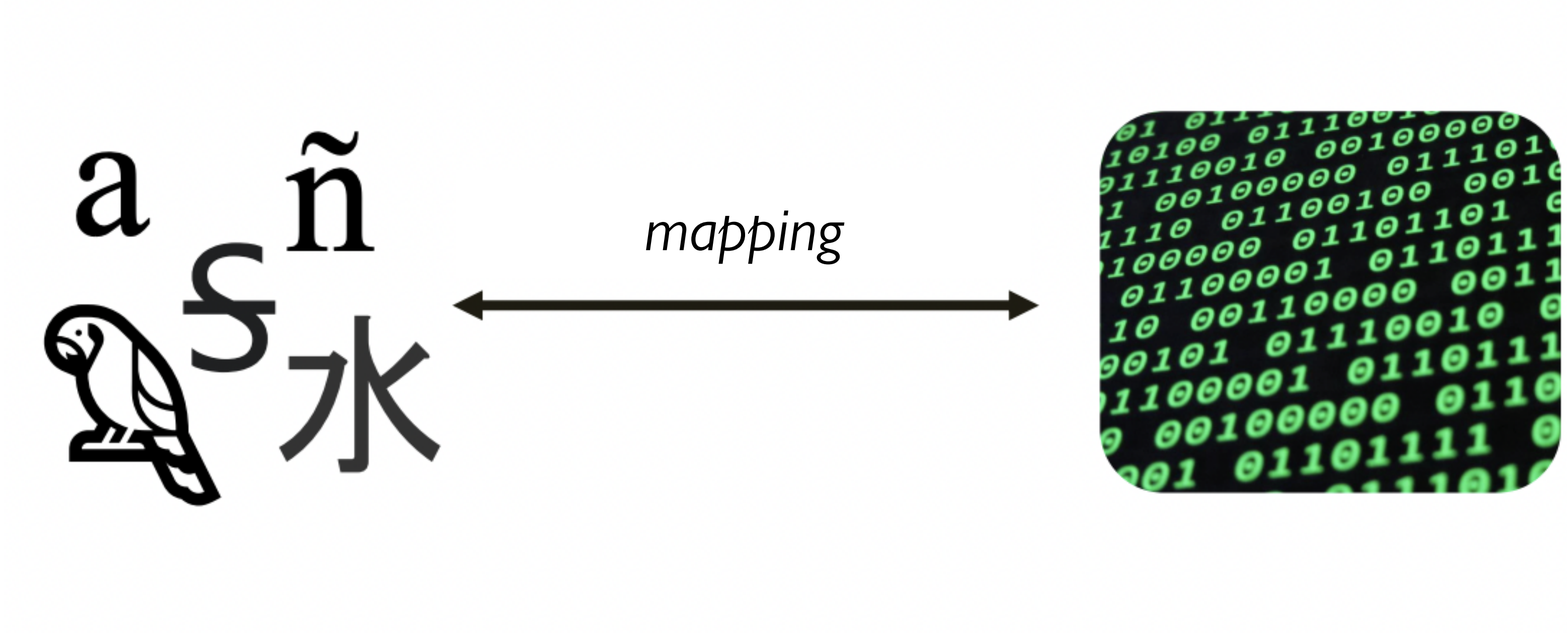
late Middle English (denoting a human or animal body): from Latin, literally **'body'**. corpus (sense 1 of the noun) dates from the early 18th century.

What *isn't* a text corpus?

For the NLP tasks we considered, the benchmark datasets behind those tasks are corpora with labels attached (and someone had to decide what the labels were).

Text on computers

Encodings



ASCII

ASCII

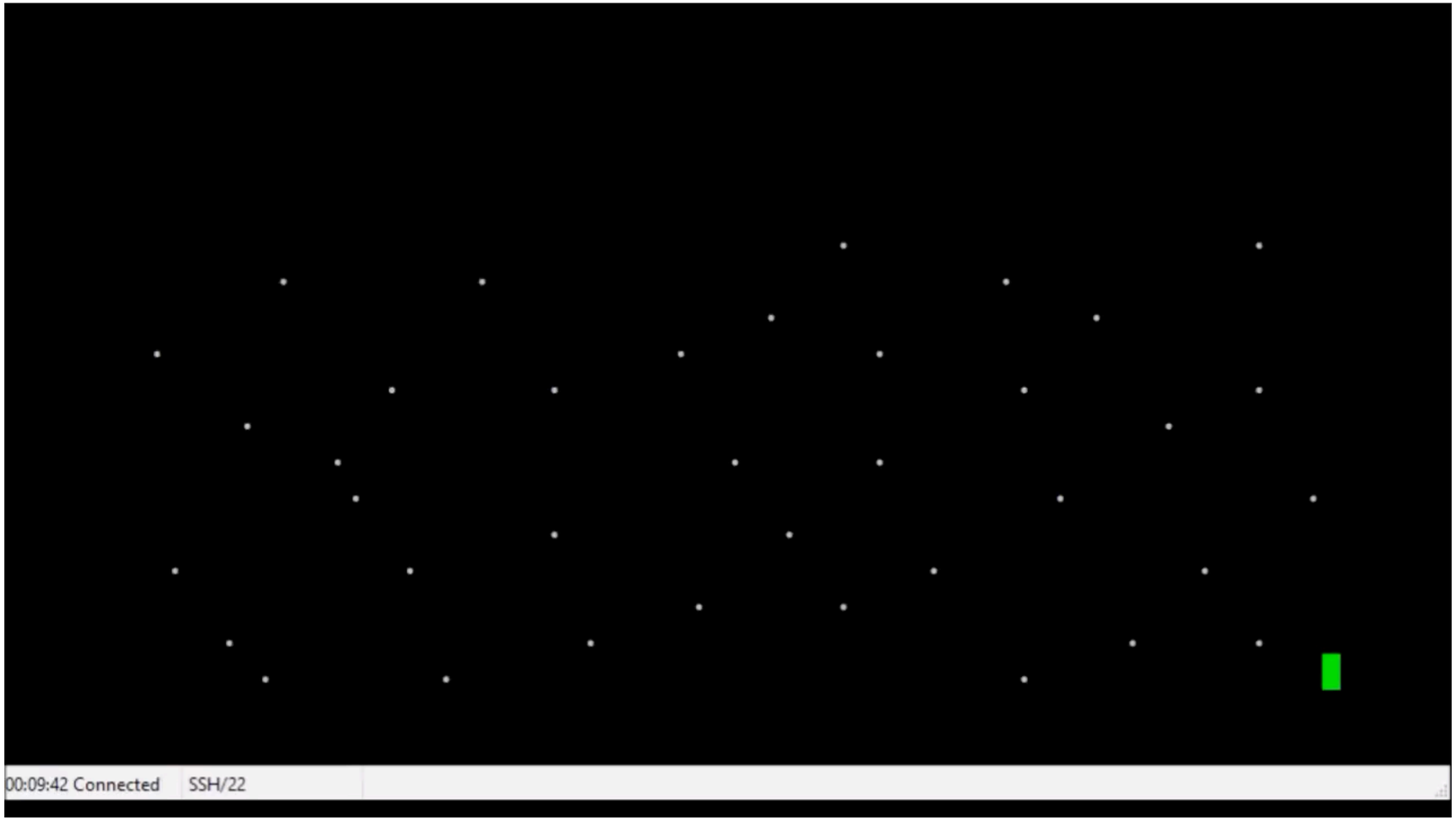
```
snd
```

```
      _  .   .   .
    .' '; '_' '_'|-.
   (     '------.'   )
    ;             \ /
     :      '    ' |/
     '              \     .;
    .:-._           '__' /
   .-'    ;--.
  /       \-'---.___.'
 |      / 7 \(>o<) /\
 |      | \ |   . \    \
 |=====|    |   .   \ |-)
  |-'-'    ./_.-._.\|"
  '-.----'          |
     |        |      |
     |      | |    | |
     |_____|_|___|_|
     (-------',----'.
      '-'-----'-----'
```

Babar

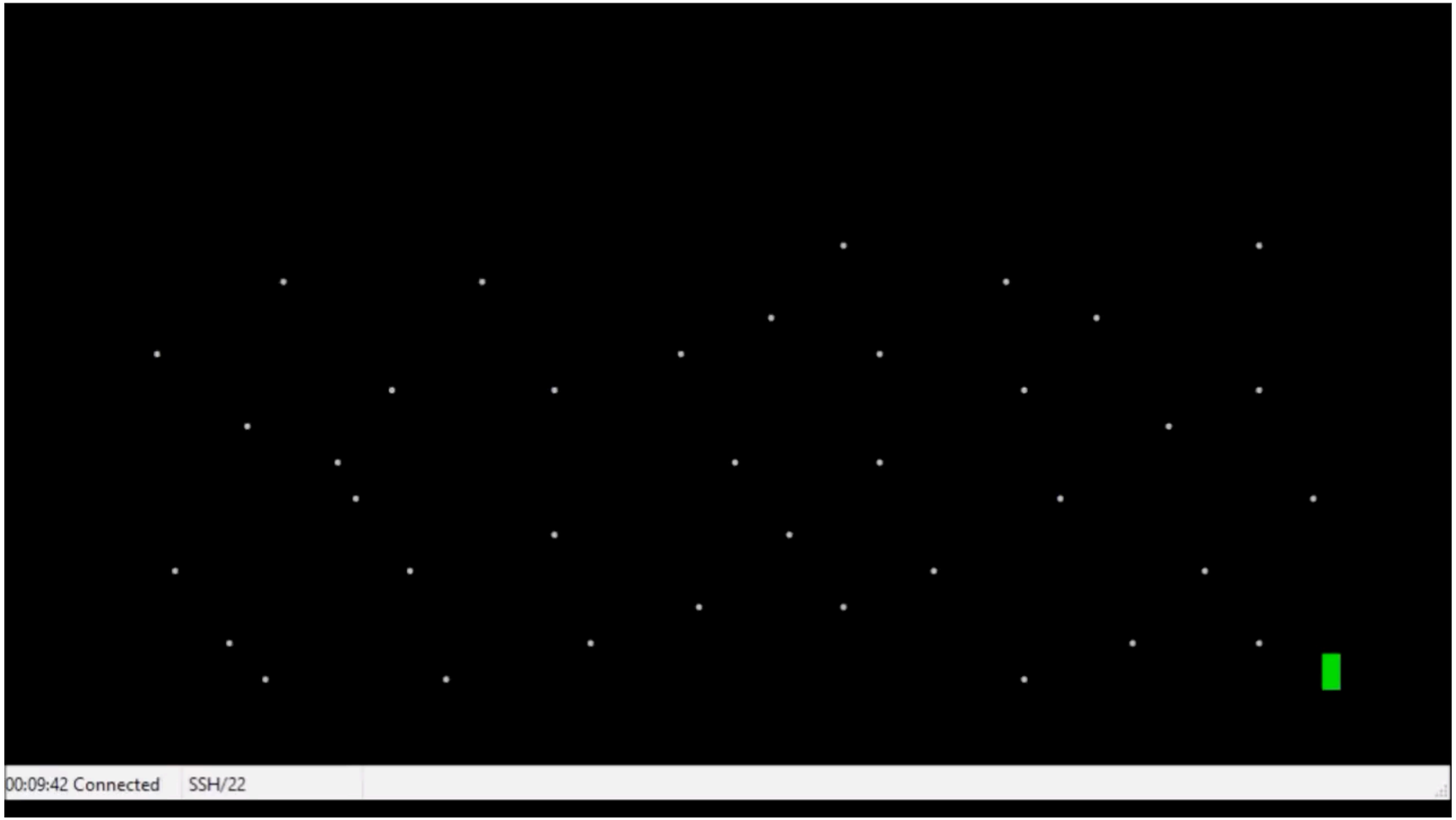
By Shanaka Dias

ASCII




```
telnet towel.blinkenlights.nl
```

ASCII




```
telnet towel.blinkenlights.nl
```

ASCII *American Standard Code for Information Interchange*

01000001	01010011	01000011	01001001	01001001	*Bits*
65	83	67	73	73	*Code point*
"A"	"S"	"C"	"I"	"I"	*Character*

Dec	Char	Dec	Char	Dec	Char	Dec	Char
0	NUL (null)	32	SPACE	64	@	96	`
1	SOH (start of heading)	33	!	65	A	97	a
2	STX (start of text)	34	"	66	B	98	b
3	ETX (end of text)	35	#	67	C	99	c
4	EOT (end of transmission)	36	$	68	D	100	d
5	ENQ (enquiry)	37	%	69	E	101	e
6	ACK (acknowledge)	38	&	70	F	102	f
7	BEL (bell)	39	'	71	G	103	g
8	BS (backspace)	40	(	72	H	104	h
9	TAB (horizontal tab)	41	)	73	I	105	i
10	LF (NL line feed, new line)	42	*	74	J	106	j
11	VT (vertical tab)	43	+	75	K	107	k
12	FF (NP form feed, new page)	44	,	76	L	108	l
13	CR (carriage return)	45	-	77	M	109	m
14	SO (shift out)	46	.	78	N	110	n
15	SI (shift in)	47	/	79	O	111	o
16	DLE (data link escape)	48	0	80	P	112	p
17	DC1 (device control 1)	49	1	81	Q	113	q
18	DC2 (device control 2)	50	2	82	R	114	r
19	DC3 (device control 3)	51	3	83	S	115	s
20	DC4 (device control 4)	52	4	84	T	116	t
21	NAK (negative acknowledge)	53	5	85	U	117	u
22	SYN (synchronous idle)	54	6	86	V	118	v
23	ETB (end of trans. block)	55	7	87	W	119	w
24	CAN (cancel)	56	8	88	X	120	x
25	EM (end of medium)	57	9	89	Y	121	y
26	SUB (substitute)	58	:	90	Z	122	z
27	ESC (escape)	59	;	91	[	123	{
28	FS (file separator)	60	<	92	\	124	|
29	GS (group separator)	61	=	93	]	125	}
30	RS (record separator)	62	>	94	^	126	~
31	US (unit separator)	63	?	95	_	127	DEL

*These are **all** of the characters in ASCII.*

1963 ASCII, a 7-bit standard (with an unused 8th bit)

1987–2000 ISO-8859, a series of 256-character 8-bit standards

Examples: Latin-1, Latin/Cyrillic, Latin/Greek, Latin/Hebrew,

But CJKV languages need bigger character sets, e.g., Big5 (two bytes), GB 18030 and ISO 2022 (more complex).

How do you tell which one a file is using?

Today we mostly use *Unicode*. Unicode isn't an encoding; it's a listing of code points.

It can be encoded in different ways:

- UTF-32 is a 32-bit fixed-length encoding
- UTF-8 and UTF-16 are variable-length encodings

Character	*Code point*	*UTF-8 bits*
A	U+0041	`01000001`
ਐ	U+0A10	`11100000 10101000 10010000`

Unicode® 17.0.0

2025 September 9 (Announcement)

This page summarizes the important changes for the Unicode Standard, Version 17.0.0. This version supersedes all previous versions of the Unicode Standard.

A. Summary
B. Technical Overview
- Core Specification
- Code Charts
- Han Radical-Stroke Indices
- Unicode Standard Annexes
- Unicode Character Database
- Version References
- Errata

C. Stability Policy Update
D. Textual Changes and Character Additions
E. Conformance Changes
F. Changes in the Unicode Character Database
G. Changes in the Unicode Standard Annexes
H. Changes in Synchronized Unicode Technical Standards
I. List of Components
M. Implications for Migration

A. Summary

Unicode 17.0 adds 4803 characters, for a total of 159,801 characters. The new additions include 4 new scripts:

- Sidetic
- Tolong Siki
- Beria Erfe

blog.emojipedia.org/whats-new-in-unicode-17-0/

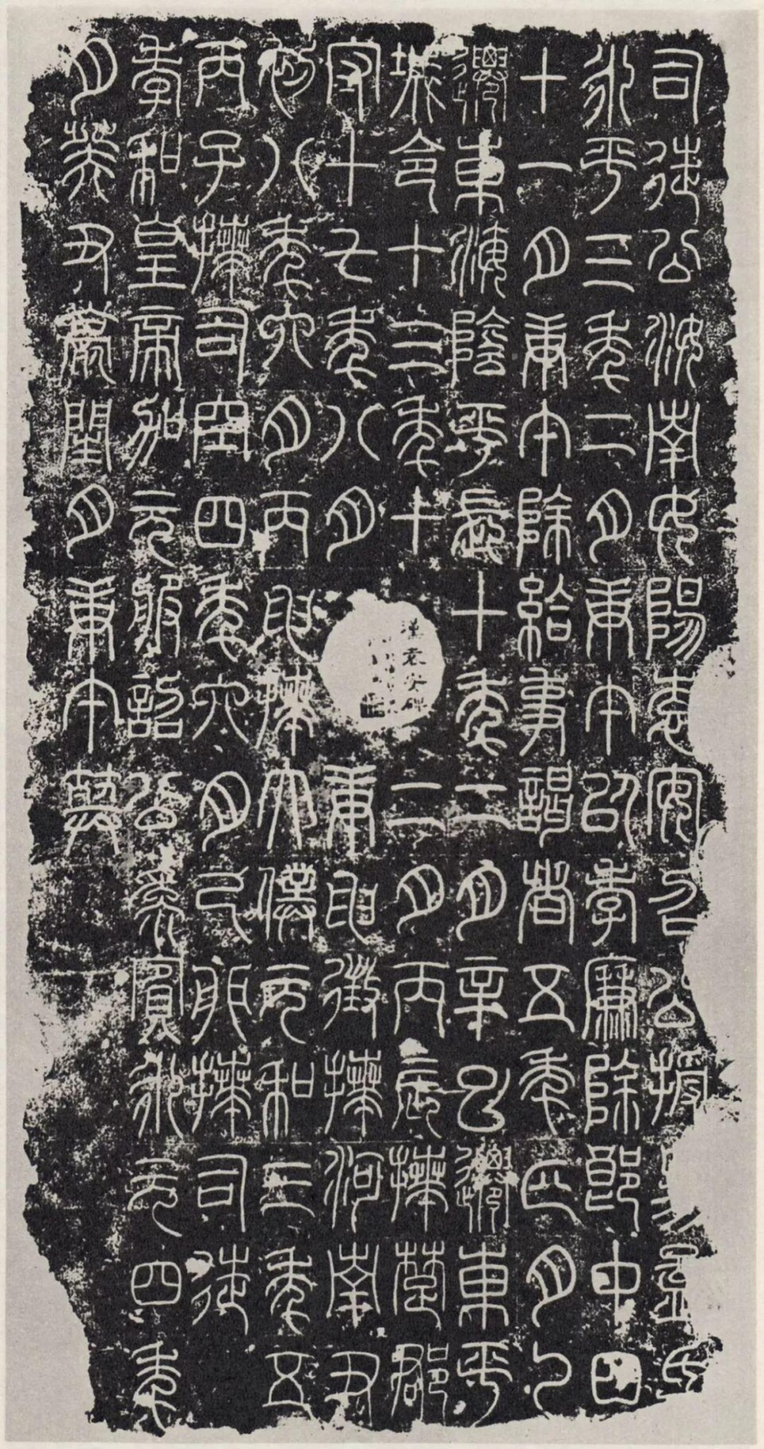
Emoji Search Botmoji World Emoji Day

UNICODE EMOJI UPDATE

What's New In Unicode 17.0

Today Unicode has officially confirmed the new emojis that will be finding their way to our emoji keyboards over the next year or so.

Keith Broni
Sep 9, 2025 • 5 min read



Epitaph of Yuan An written in seal script.

Wikipedia

㕓

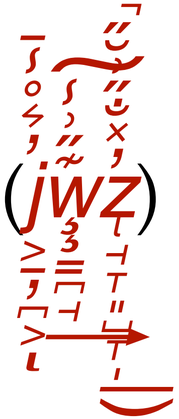
Introduced by Unicode 12.1 in 2019

Computers have gotten better at typography, but they've made things more difficult for NLP.

Computers have gotten better at typography, but they've made things more difficult for NLP.

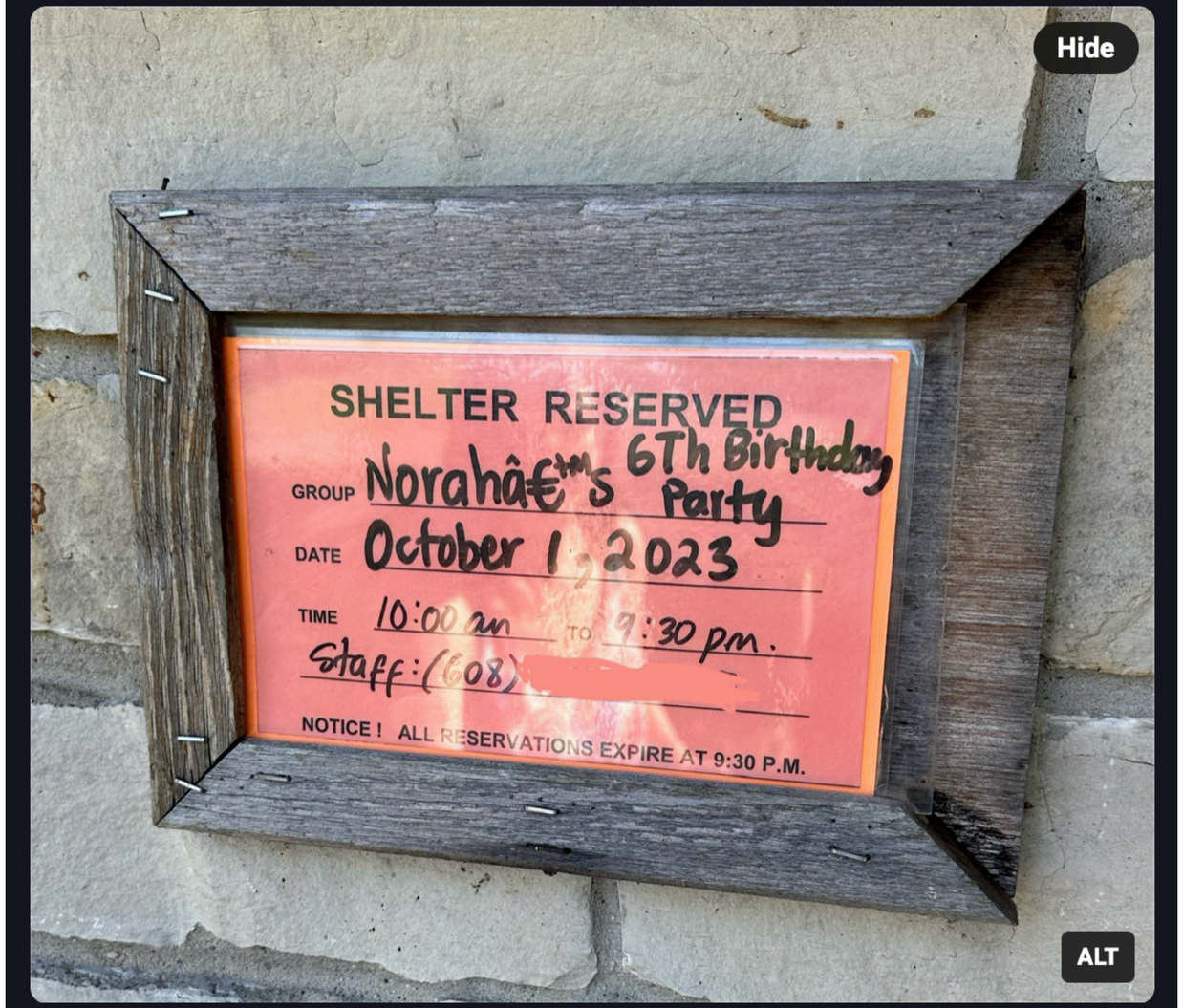
Things that can cause trouble:

- Contextual characters (“quotation marks” rather than "quotation marks")
- Newlines (LF, CR, or CRLF)
- Ligatures (sco*ff*)
- Combining characters (*jWZ*)



Brad Grzesiak 🪿
@listrophy@ruby.social

Computers were a mistake



Oct 01, 2023, 10:51 PM · · Ivory for iOS

Next time

(Almost?) everything you proposed today was a rule, and a traditional NLP system like ELIZA is a pile of them.

Next class, we'll be looking at how far we can get at solving NLP problems with rule-based systems – and where those fail.

Acknowledgments

This class incorporates material from:

Carolyn Anderson, Wellesley College

Katrin Erk, University of Texas at Austin

Katie Keith, Williams College

Xanda Schofield, Harvey Mudd College

Jurafsky & Martin, *Speech and Language Processing*, 3rd ed. draft