CMPU 366 · Computational Linguistics

Tokens and Normalization

10 September 2025

Assignment 1

Out later today

Due at the start of class next Wednesday

Where are we?

In the first class, we talked about language as linguists think about it, through the lens of levels of linguistic abstraction.

But we saw that when we handle raw text, all we have is a sequence of characters in some encoding, and those layers aren't always easy to pull apart.

We've seen that seemingly hard tasks – like psychoanalysis or winning at Jeopardy! – can be approximated with very little language understanding, using regular expressions to capture text patterns.

We'll continue considering how we can describe and process text data in deeper ways.

Most NLP tasks involve *segmenting* language into workable units…

narratives,

paragraphs,

sentences,

words,

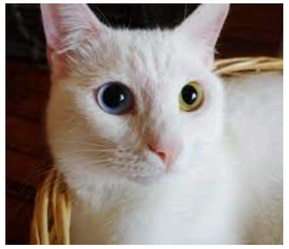
morphemes

…which undergo varying degrees of *normalization*.

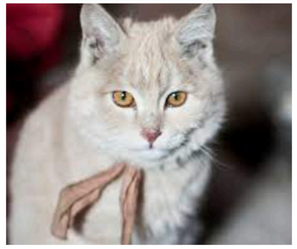
Words

What are words?

How the Cat Gets Its Stripes: It's...
nytimes.com

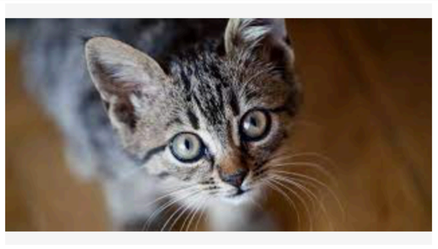


Van cat - Wikipedia
en.wikipedia.org

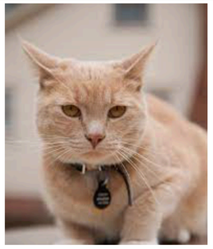


Cat | National Geographic
nationalgeographic.com

The inner lives of cats: what our ...
theguardian.com



Thinking of getting a cat ...
icatcare.org



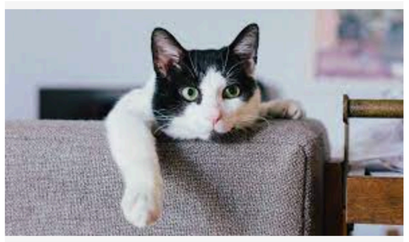
Why You're Probably Tra...
nationalgeographic.com

Cat - Wikipedia
en.wikipedia.org

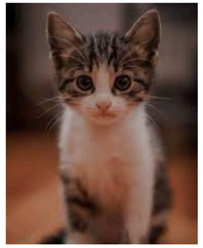
Is my cat happy? This app predicts ...
wired.co.uk

Reading the Eyes of Your Cat
thesprucepets.com

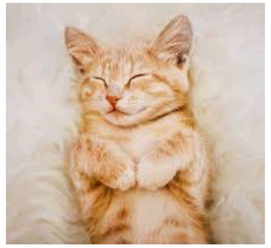
pet guru Yuki Hattori explain |...
theguardian.com



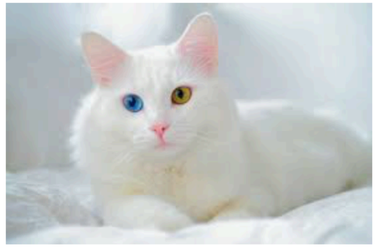
Are Cats Really Aloof? : Short Wave : NPR
npr.org



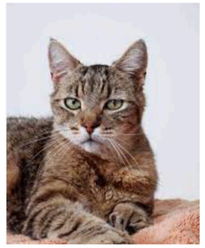
500+ Domestic Cat Pic...
unsplash.com



100 Best Girl Cat Names - Uniqu...
goodhousekeeping.com



World's most popular cat breeds | Daily ...
dailysabah.com



The 10 Best Types of C...
britannica.com

Cat infected with COVID-19 from owner ...
livescience.com

Illustrated by a Cat | History ...
smithsonianmag.com

“cat”

One morning I shot an elephant in my pajamas.

One morning I shot an elephant in my pajamas.

I didn't shoot an elephant.

One morning I shot an elephant in my pajamas.

I didn't shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.

One morning I shot an elephant in my pajamas.

I didn't shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.

I do uh main- mainly business data processing.

One morning I shot an elephant in my pajamas.

I didn't shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.

I do uh main- mainly business data processing.

Have a great day! :-)

When we talk about a “word”, we might mean

an abstract *vocabulary item*, or — *word type*

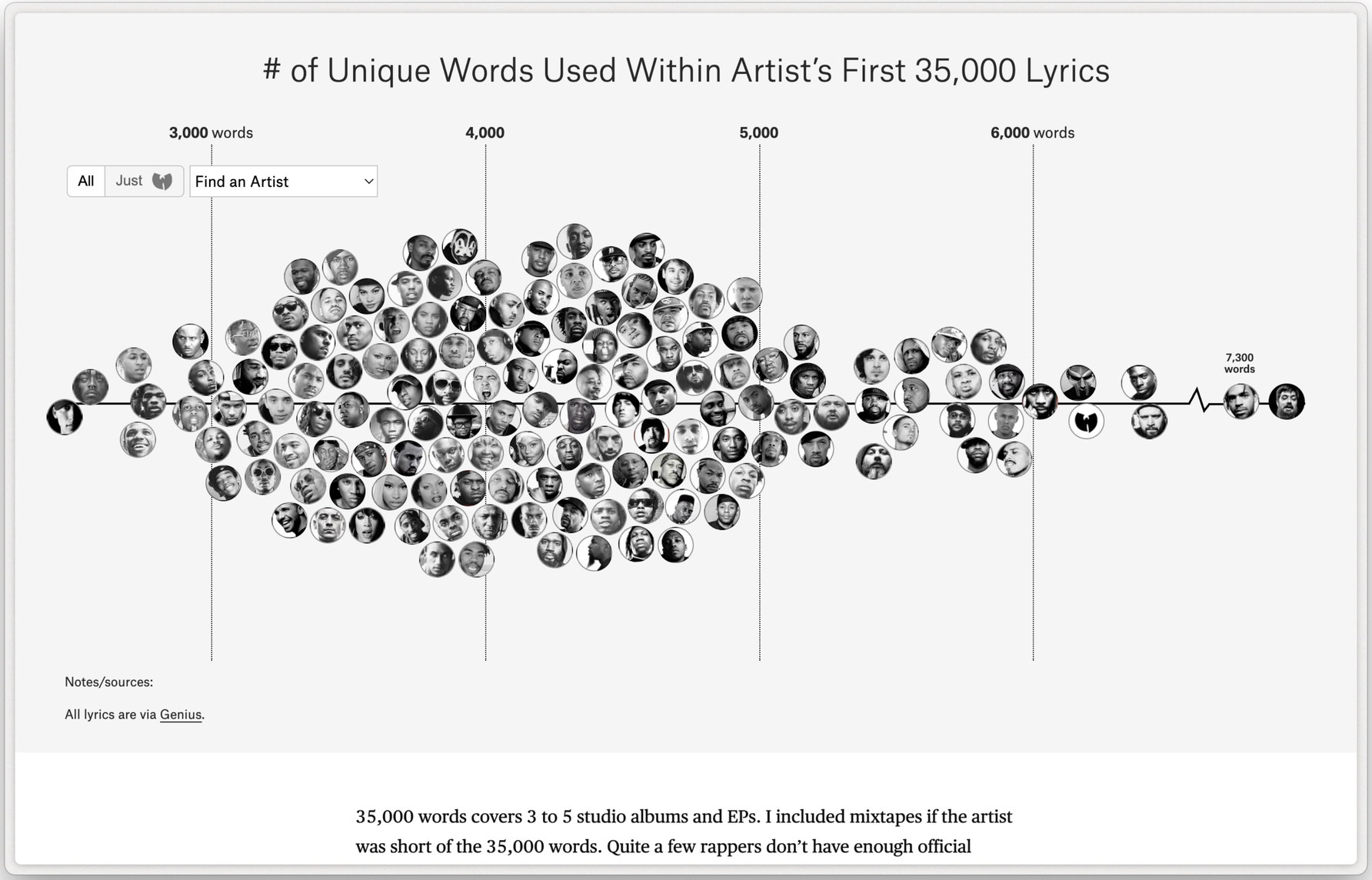
an individual *occurrence* of a word. — *word token*

To be or not to be

6 tokens: [*to*, *be*, *or*, *not*, *to*, *be*]

4 types: {*be*, *not*, *or*, *to*}

When we look at a corpus, it can be informative just to count the word types and word tokens, giving us a measure of vocabulary richness/variation.



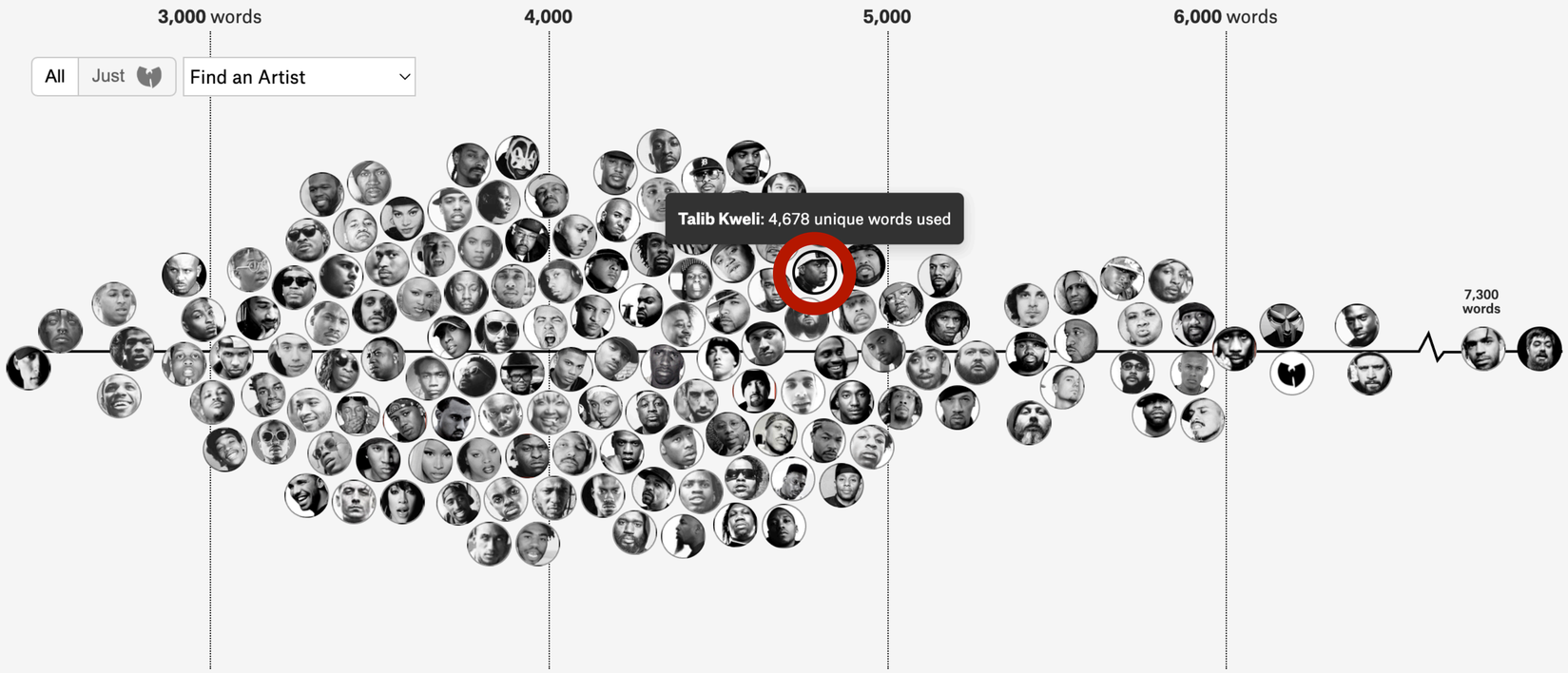
35,000 words covers 3 to 5 studio albums and EPs. I included mixtapes if the artist was short of the 35,000 words. Quite a few rappers don't have enough official

pudding.cool/projects/vocabulary

I dumbed down for my audience to double my dollars.
They criticized me for it, yet they all yell "holla".
If skills sold, truth be told, I'd probably be
Lyrically Talib Kweli.
Truthfully, I wanna rhyme like Common Sense,
But I did five mil.; I ain't been rhyming like Common since.

Jay-Z, "Moment of Clarity", 2003

of Unique Words Used Within Artist's First 35,000 Lyrics

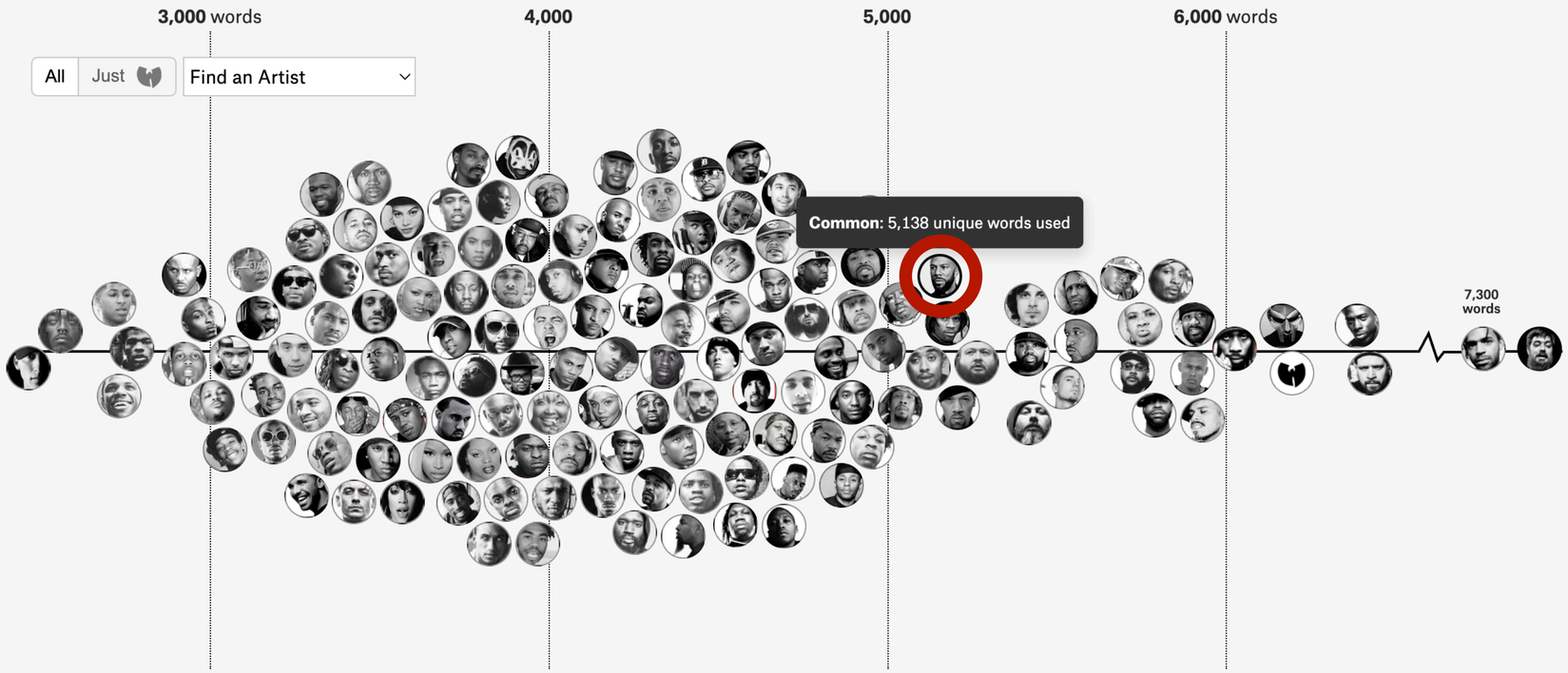


Notes/sources:

All lyrics are via Genius.

35,000 words covers 3 to 5 studio albums and EPs. I included mixtapes if the artist was short of the 35,000 words. Quite a few rappers don't have enough official

of Unique Words Used Within Artist's First 35,000 Lyrics

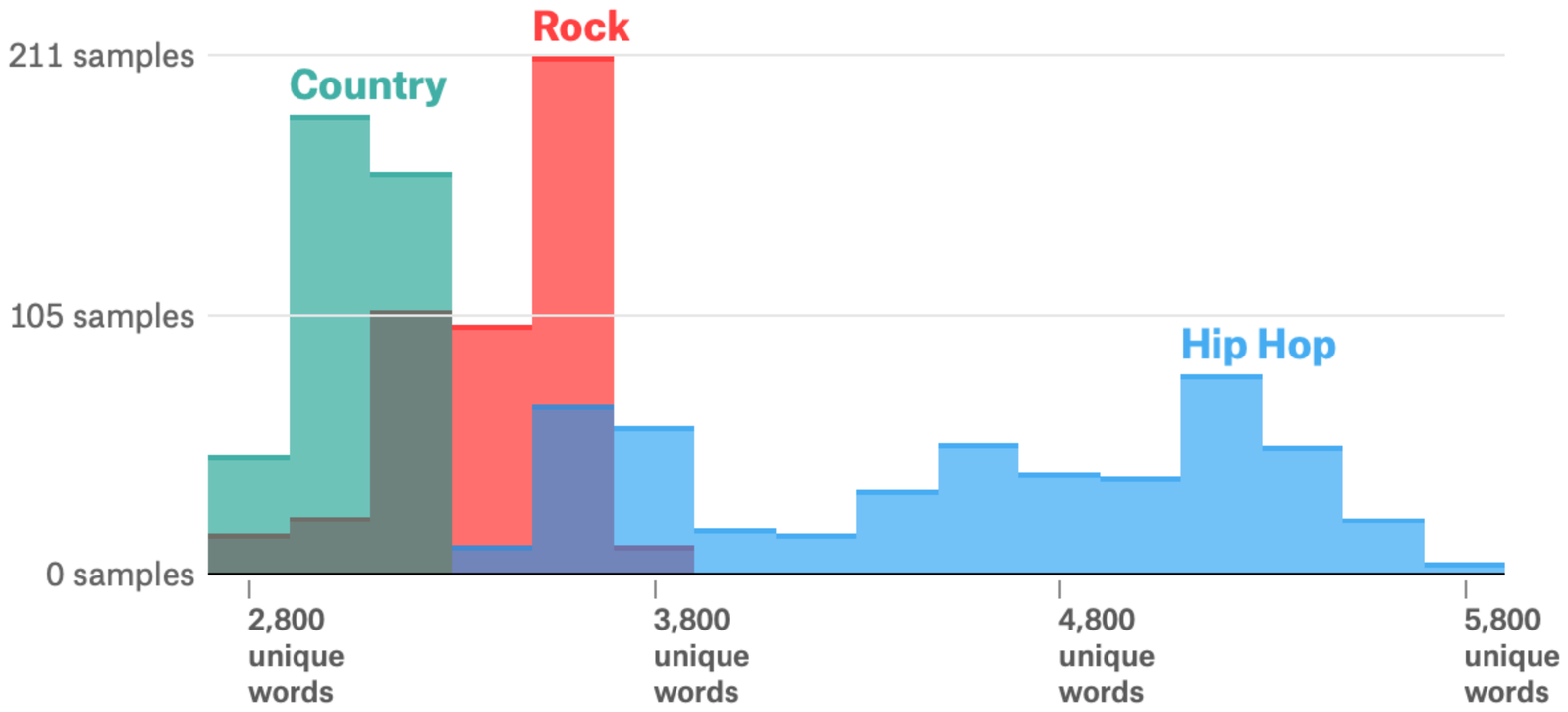


Notes/sources:

All lyrics are via Genius.

35,000 words covers 3 to 5 studio albums and EPs. I included mixtapes if the artist was short of the 35,000 words. Quite a few rappers don't have enough official

of Unique Words Used in 500 Random Samples of 35,000 Lyrics from Country, Rock, Hip Hop



Raw Lyrics Data via John W. Miller

Separating words

Tokenization is splitting a text into the word tokens we want to process.

Why can’t we just use white space to tokenize a text?

That old bell, presage of a train, had just sounded through Oxford station; and the undergraduates who were waiting there, gay figures in tweed or flannel, moved to the margin of the platform and gazed idly up the line.

Max Beerbohm, *Zuleika Dobson*, 1911

That old bell, presage of a train, had just sounded through Oxford station; and the undergraduates who were waiting there, gay figures in tweed or flannel, moved to the margin of the platform and gazed idly up the line.

Max Beerbohm, *Zuleika Dobson*, 1911

In a sample of 100 books from Project Gutenberg,
we find these tokens when we split on white space:

889	*earth*	6	*earth!"*	2	*earth--"*
345	*earth,*	5	*earth?"*	2	*earth),*
213	*earth.*	4	*earth.'*	1	*earth?'*
76	*earth's*	3	*earth._"*	1	*earth;_*
49	*earth;*	3	*earth--he*	1	*earth;"*
19	*earth."*	3	*earth--*	1	*earth:--*
14	*earth?*	3	*earth)*	1	*earth._*
9	*earth!*	3	*earth!'*	1	*earth.[5]*
8	*earth:*	2	*earth.'"*	1	*earth....*
8	*earth,"*	2	*earth-goddess*	1	*earth.--Thou*
6	*earth"*	2	*earth--as*	1	*earth."--The*
					…

Can we just strip out all punctuation?

Punctuation can be helpful:

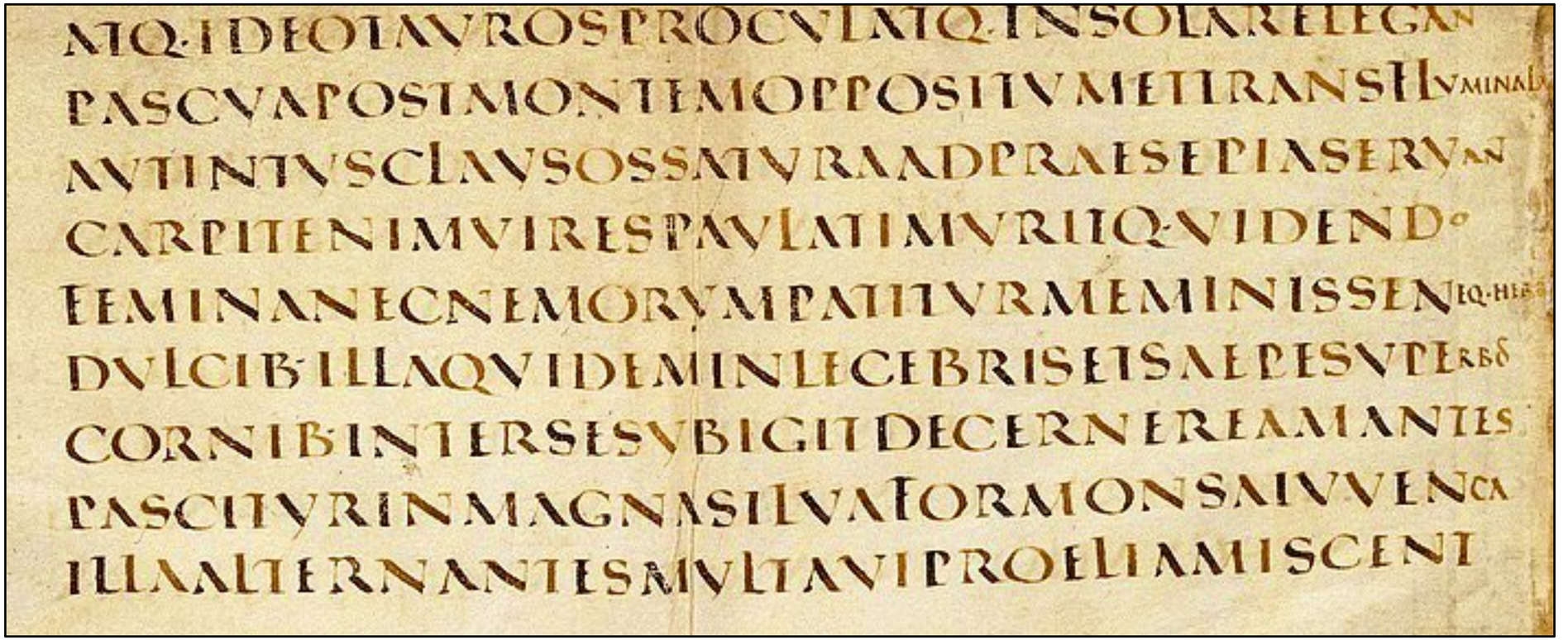
- It can mark boundaries for sentences, clauses, parentheticals, asides.
- Some punctuation has illocutionary force, like exclamation points and question marks.
- Emoticons are strong signals of sentiment

Exercise: Write a regular expression to tokenize English and then count each token.

tinyurl.com/2025-09-10-starter

But, English is an easy one!

What if you don’t have spaces and punctuation to help you?



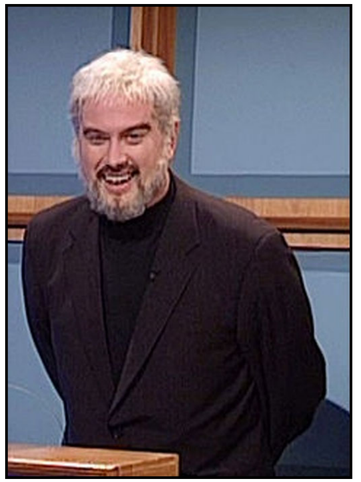
Vergilius Augusteus

Susan Boyle
@SusanBoyleHQ

Susan will be answering your questions at her exclusive album listening party on Saturday. Send in your questions #susanalbumparty Susan HQ

Susan Boyle
@SusanBoyleHQ

Susan will be answering your questions at her exclusive album listening party on Saturday. Send in your questions #susanalbumparty Susan HQ



姚明进入总决赛

姚明	进入	总决赛
Yao Ming	*reaches*	*finals*

姚	明	进入	总	决赛
Yao	*Ming*	*reaches*	*overall*	*finals*

姚	明	进	入	总	决	赛
Yao	*Ming*	*enter*	*enter*	*overall*	*decision*	*game*

Tokenization's hard, and the correct behavior depends on what we want to treat as a token.

Consider how to handle clitics.

We can leave them alone:

Can't → *Can't*

We can separate them:

Can't → *Can 't*

Can't → *Ca n't*

We can expand them:

Can't → *Can not*

What are the dangers of simple tokenization?

Why not leave words like *can't* alone?

This doesn't mean we always split distinct units.

German compound nouns like

Lebensversicherungsgesellschaftsangestellter

"life insurance company employee"

are traditionally treated as individual tokens.

Subword tokenization

Subword tokenization approaches like *byte pair encoding* are typically bottom-up algorithms – they use the data to learn what the tokens should be!

Corpus

Corpus

Token learner

Corpus

Token learner

Vocabulary

```
un
friendly
person
's
...
```

Corpus

Token learner

Vocabulary

```
un
friendly
person
's
...
```

Token segmenter

Corpus

String to tokenize

```
my cat's
friendly
```

Token learner

Vocabulary

```
un
friendly
person
's
...
```

Token segmenter

Corpus

String to tokenize

```
my cat's
friendly
```

Token learner

Vocabulary

```
un
friendly
person
's
...
```

Token segmenter

```
my cat 's
friend ly
```

Tokens!

Corpus

Token learner

Vocabulary

```
un
friendly
person
's
...
```

Let the vocabulary be the set of all individual characters: {A, B, C, D, …, a, b, c, d, …}

Repeat *k* times:

Choose the two symbols that are most frequently adjacent in the training corpus (say, A, B)

Add a new merged symbol AB to the vocabulary

Replace every adjacent A B in the corpus with AB.

Return the vocabulary.

Since most subword algorithms are run *inside* the initial space-separated tokens, we commonly add a special end-of-word symbol _ before spaces in the training corpus.

Consider this – somewhat repetitious – corpus:

> *low low low low low lowest lowest newer newer newer newer newer newer wider wider wider new new*

Adding end-of-word tokens results in this initial vocabulary:

> `_`, `d`, `e`, `i`, `l`, `n`, `o`, `r`, `s`, `t`, `w`

Corpus:

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w

Corpus:

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w

Merge e r *to* er

Corpus:

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w

Merge e r *to* er

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er _
3 w i d er _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er _
3 w i d er _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er _
3 w i d er _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er

Merge er _ *to* er_

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er _
3 w i d er _
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er

Merge `er _` *to* `er_`

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er, er_

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er, er_

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er, er_

Merge n e *to* ne

Corpus:

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er, er_

Merge n e *to* ne

Corpus:

5 l o w _
2 l o w e s t _
6 ne w er_
3 w i d er_
2 ne w _

Vocabulary: _, d, e, i, l, n, o, r, s, t, w, er, er_, ne

Merge	*Vocabulary*
(ne, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new
(l, o)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo
(lo, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low
(new, er_)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_
(low, _)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_

String to tokenize

```
my cat's
friendly
```

Vocabulary

```
un
friendly
person
's
...
```

Token segmenter

```
my cat 's
friend ly
```

Tokens!

After training BPE on a corpus of data, we can now run it on new data, applying each merge learned from the training data, greedily in the order we learned them.

So, merge every `e r` to `er`, then merge `er _` to `er_`, etc.

Result:

If we see `n e w e r _`, it will be tokenized as a full word.

If we see `l o w e r _`, it will be tokenized as two tokens: `low er_`.

The tokens found by BPE usually include frequent words and frequent subwords.

Many of these subwords are *morphemes* – the smallest meaning-bearing units of a language, e.g., *un-*, *-er*, *-est*.

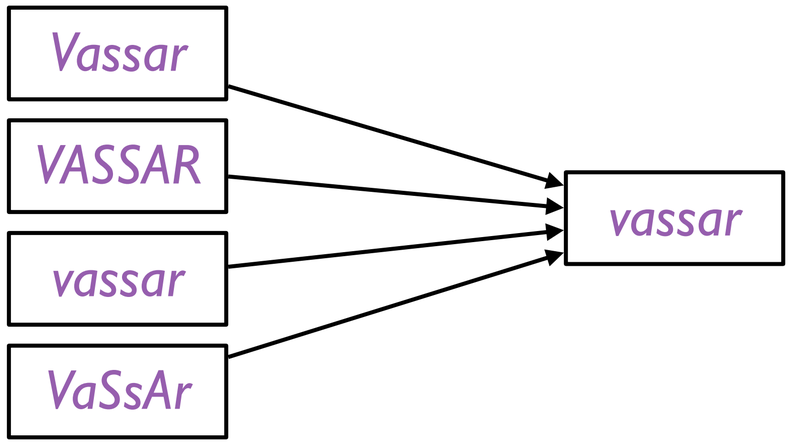
> Byte Pair Encoding (BPE) (Sennrich et al., 2015) is a practical middle ground between character and word level language modeling which effectively interpolates between word level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences.

Radford et al., 2019

This is the basic approach that's used by systems like ChatGPT.

Try it out: `platform.openai.com/tokenizer`

Normalizing words



What does case folding get us?

What do we lose?

organizes

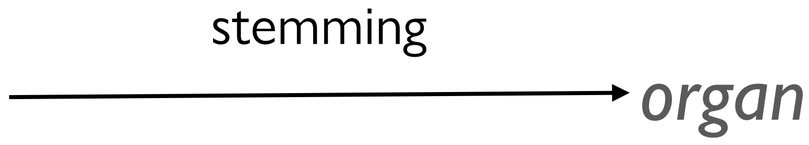
organized

organizing

organizes

organized

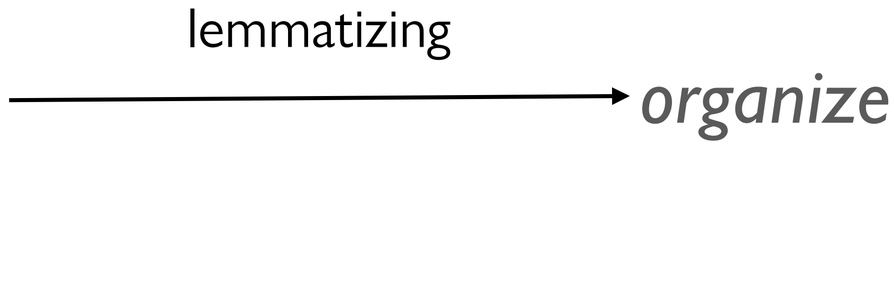
organizing



organizes

organized

organizing



the boy's cars are different colors

→ *the boy car be different color*

He is reading detective stories

→ *he be read detective story*

Why might we prefer lemmatization to stemming?

What does lemmatizing get us?

What do we lose?

Beyond tokens

In addition to splitting a text into tokens, we can think about finding other boundaries, like separating sentences. You'll get a chance to try this on Assignment 1!

Whenever we're making decisions about text, like identifying words or identifying sentence boundaries, we need to think about evaluation.

Error types

	Actually a sentence boundary	*Not actually a sentence boundary*
Program thinks it's a sentence boundary	*True positive* 	*False positive* 
Program thinks it isn't a sentence boundary	*False negative* 	*True negative* 

In NLP, we're always dealing with these kinds of errors.

Reducing the error rate for an application often involves two antagonistic efforts:

- Increasing accuracy or precision (minimizing false positives)
- Increasing coverage or recall (minimizing false negatives)

Acknowledgments

This class incorporates material from:

Carolyn Anderson, Wellesley College

David Bamman, UC Berkeley

Matt Daniels, “The Largest Vocabulary in Hip Hop”

Katie Keith, Williams College

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