TOKENIZATION

SYRACUSE UNIVERSITY
School of Information Studies

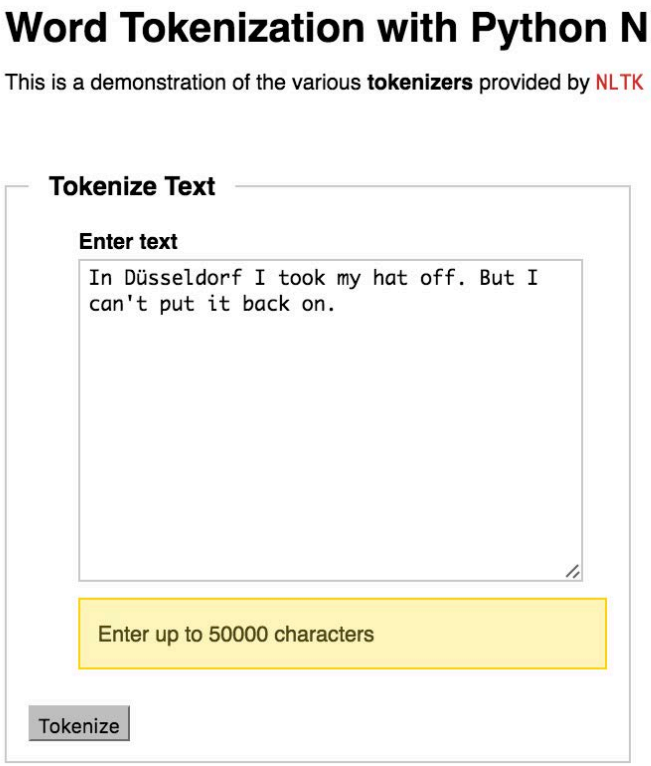
TEXT REPRESENTATION AND VECTORIZATION

Convert text to numbers

Computers can do only ONE thing, that is, COUNTING!

TOKENIZATION

A tokenizer has a set of rules about grouping characters into tokens.



TOKENIZATION RULES

2.

2.

2.

2.

TOKENIZATION IS NOT EASY

Tokenizing URLs

Choosespain.com

TOKENIZATION IS NOT EASY

Tokenize text strings with no white space

Chinese (New Year couplets):
养猪大如山老鼠头头死

Raise|pigs|big|as|mountain|rats|all|die

养|猪|大|如|山|老鼠|头头|死

Raise|pigs|big|as|mountain rats, all|die

养|猪|大|如|山老鼠| 头头|死

TOKENIZATION IS NOT EASY

Lowercase vs. uppercase

Words with inflected forms

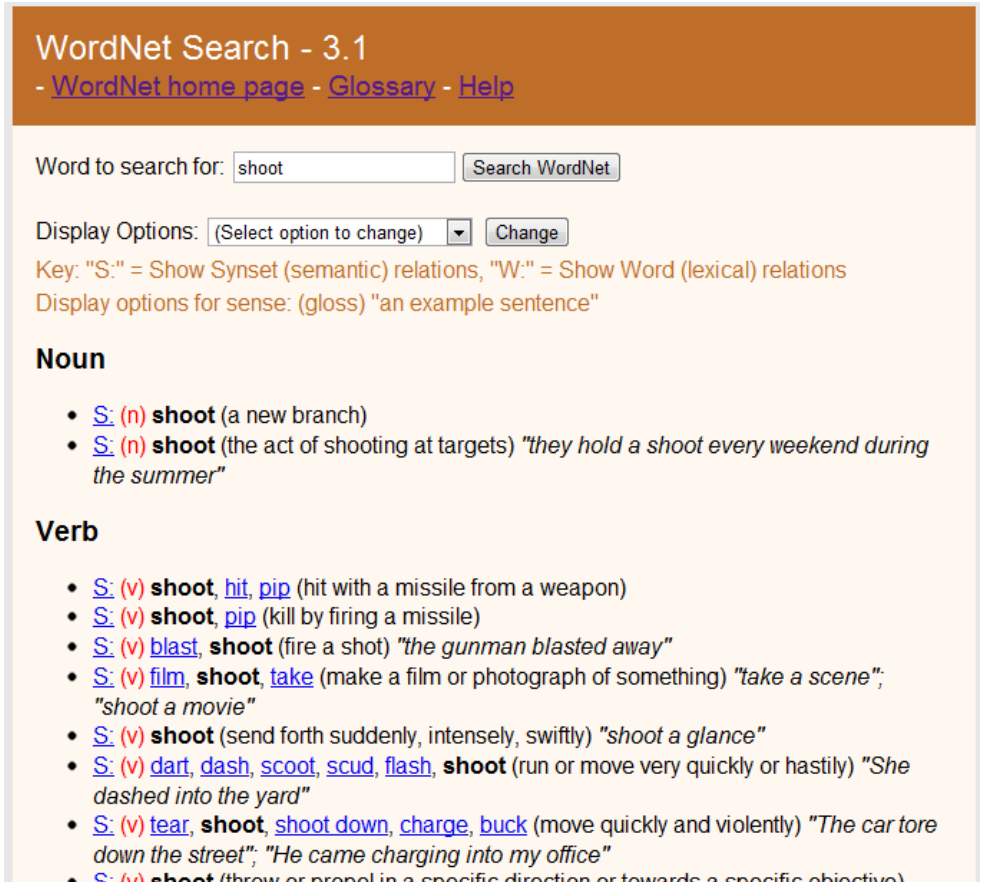
"dishwasher" vs. "dishwashers"

Words with multiple senses

"There is a money **bank** near the river **bank.**"

WORDNET

http://wordnetweb.princeton.edu/perl/webwn



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: shoot Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **shoot** (a new branch)
- S: (n) **shoot** (the act of shooting at targets) *"they hold a shoot every weekend during the summer"*

Verb

- S: (v) **shoot**, hit, pip (hit with a missile from a weapon)
- S: (v) **shoot**, pip (kill by firing a missile)
- S: (v) blast, **shoot** (fire a shot) *"the gunman blasted away"*
- S: (v) film, **shoot**, take (make a film or photograph of something) *"take a scene"; "shoot a movie"*
- S: (v) **shoot** (send forth suddenly, intensely, swiftly) *"shoot a glance"*
- S: (v) dart, dash, scoot, scud, flash, **shoot** (run or move very quickly or hastily) *"She dashed into the yard"*
- S: (v) tear, **shoot**, shoot down, charge, buck (move quickly and violently) *"The car tore down the street"; "He came charging into my office"*
- S: (v) **shoot** (throw or propel in a specific direction or towards a specific objective)

WORD SENSE DISAMBIGUATION (WSD)

WSD techniques use word context to decide the word sense

Could introduce more errors to next steps

So far does not help search engines significantly

Not widely used in text mining

- Text mining tends to use shallow features to process large amount of text data.

VECTORIZATION

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO COUNT TOKENS

Convert documents into word vectors

Bag of words (BOW)

- Boolean
- Term frequency
- Normalized term frequency
- Tf–idf

VECTORIZATION

Step 1: Create a dictionary of unique words.

1 “vector”
2 “number”
3 “text”
…

Step 2: Represent every document as a word vector; each word is an attribute or feature.

	“vector”	“number”	“text”	…
Doc1	1	0	0	
Doc2	1	1	1	
Doc3	1	0	1	

VALUES OF WORD FEATURES

Boolean value: Word presence or absence

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	1	1	1	
Doc3	1	0	1	

VALUES OF WORD FEATURES

Word frequency: The number of word occurrences

	"vector"	"number"	"text"	…
Doc1	5	0	0	
Doc2	1	3	6	
Doc3	2	0	8	

VALUES OF WORD FEATURES

Normalized word frequency: Word frequency normalized by the document length

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	0.1	0.3	0.6	
Doc3	0.2	0	0.8	

VALUES OF WORD FEATURES

Tf–idf weighting

Tf: Term (word) frequency

Df: Document frequency, i.e, how many documents contain this term (e.g., 8 out of 100 documents -> 8/100)

Idf: Inverse document frequency, 100/8

Tf–idf = tf*log(idf)

	"vector"	"number"	"text"
Doc1	1	0	0
Doc2	0.1	0.3	0.6
Doc3	0.2	0	0.8

	"vector"	"number"	"text"
Doc1	0	0	0
Doc2	0	0.3*log3	0.6*log 1.5
Doc3	0	0	0.8*log 1.5

TF–IDF

Concept borrowed from information retrieval

A “blind” weighting strategy for text classification

REDUCING VOCABULARY SIZE

SYRACUSE UNIVERSITY
School of Information Studies

APPROACHES TO REDUCE THE VOCABULARY SIZE

- Stemming
- Case merging
- Removing stop words
- Word clustering

STEMMING

Characteristic of inflected language like English

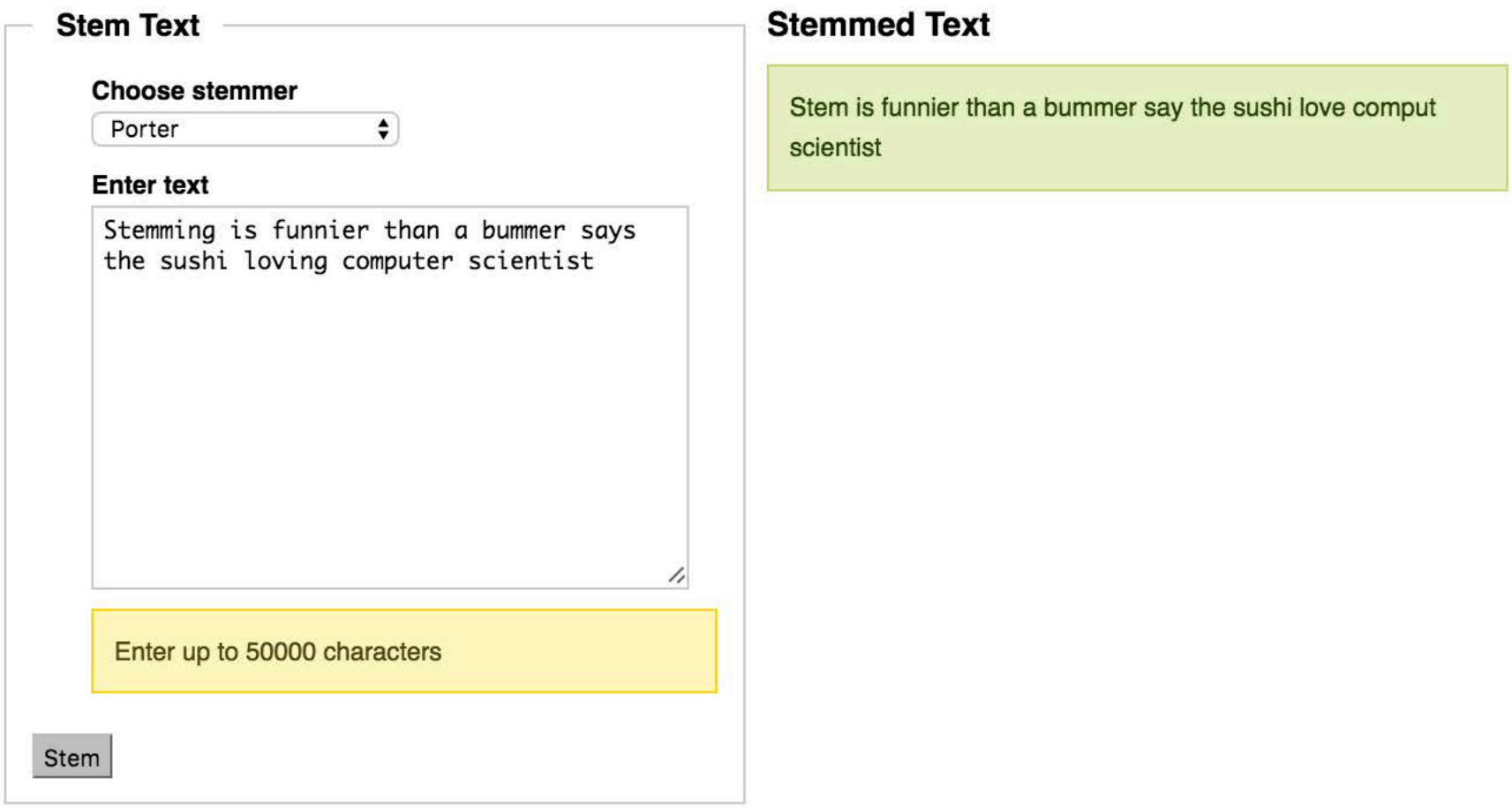
Stemmer: Remove postfixes to find the root form

"applied" and "application" -> "appli"

Lemmatizer: Transform the root to a real word

"applied" and "application" -> apply

NLTK STEMMING DEMO



http://text-processing.com/demo/stem/

STEMMING ISSUES

How far should it go?

"denormalization" -> denormalize -> denormal -> normal -> norm?

How accurate can it be?

"bore"/ he wanted to bore a hole / he bore the students on his heart

HOW USEFUL IS STEMMING?

No consistent conclusion

Information retrieval

- Search “dishwasher” to know how it works
- Search “dishwashers” to shop around

Text categorization

- Future tense vs. past tense in company performance report
  - “Will do” vs. “have done”

CONVERT UPPERCASE TO LOWERCASE?

Emily Dickinson's poem

"Joy" vs. "joy"

"Love" vs. "love"

UPPERCASE

But pompous Joy Betrays us, as his first Betrothal Betrays a Boy.	The Treason of an Accent Might vilify the Joy - To breathe - corrode the rapture Of Sanctity to be	Boundlessness - Expanse cannot be lost - Not Joy, but a Decree Is Deity - His Scene, Infinity -

LOWERCASE

Could she have guessed that it would be - Could but a Crier of the joy Have climbed the distant hill! -	I want to send you joy, I have half a mind to put up one of these dear little Robin's, and . . .	I can't believe you are coming - but when I think of it, and tell myself it's so, a wondrous joy comes over me, and my old fashioned life . . .

REMOVE STOP WORDS

Observation: Words occur in most documents that are not useful for distinguishing documents.

Stop words are usually function words that bear no specific meaning, compared to content words.

EXAMPLE OF THE START OF A STOP WORD LIST

a
about
across
after
afterwards
again
against
all
alone
along
already
also
always
among
an
and
another
any
anyhow
anyone
anything
are
around
as
be
because
becomes
becoming
been
before
beforehand
behind
being
below
besides
between
beyond
but
can

LITTLE WORDS CAN MAKE A BIG DIFFERENCE

SYRACUSE UNIVERSITY
School of Information Studies

LITTLE WORDS CAN MAKE A BIG DIFFERENCE

Function words are useful for certain text mining tasks:

Genre classification

Authorship attribution

Gender detection

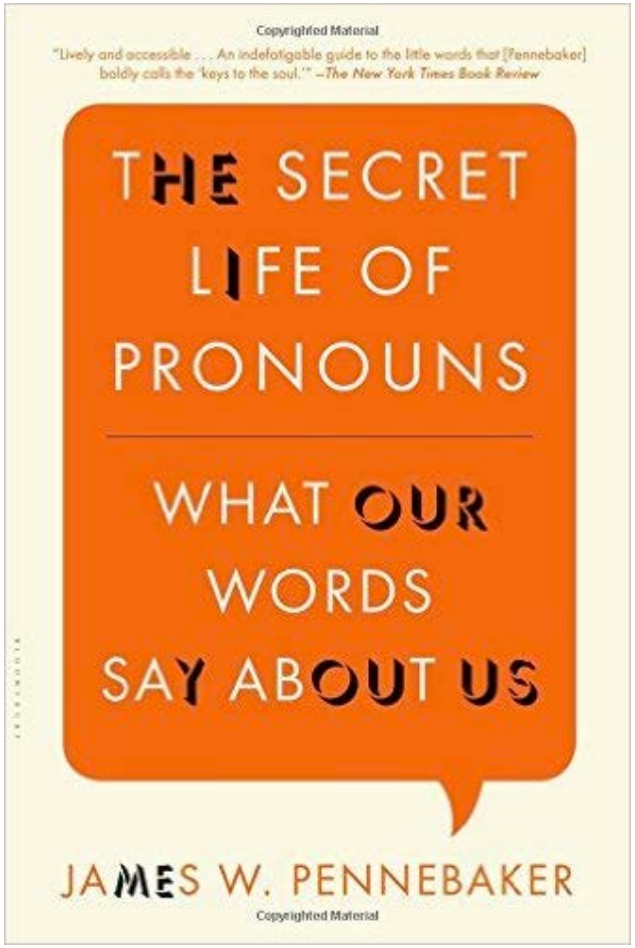
GENRE CLASSIFICATION

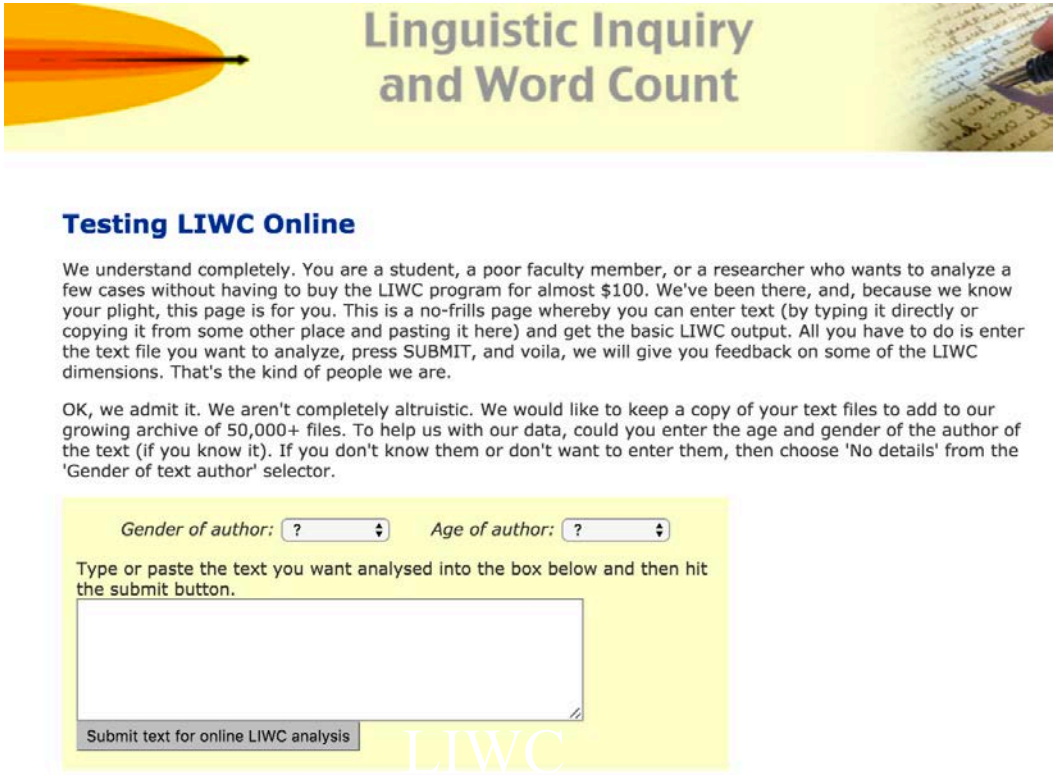
Personal home page identification (Riloff, 1995)

Top features “I” and “my”

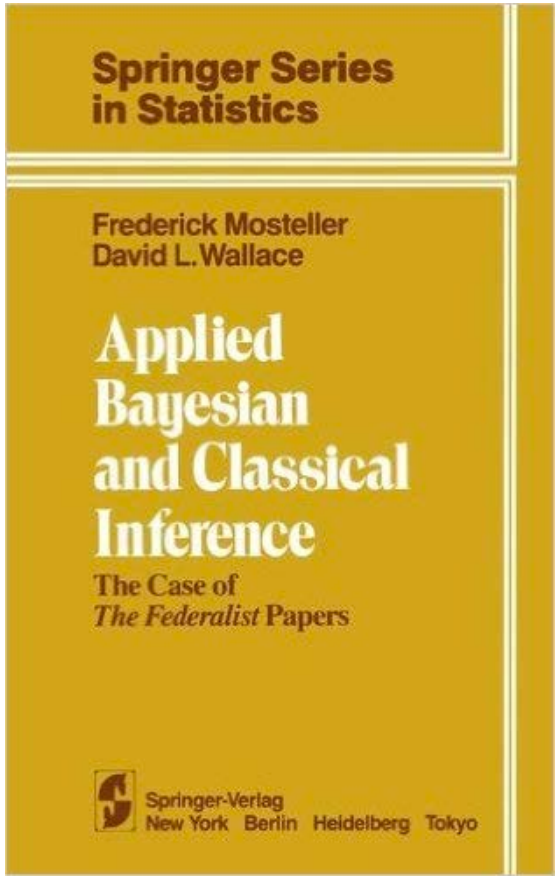
Riloff, E. (1995, July). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 130–36). Association for Computing Machinery.

PERSONAL PRONOUNS





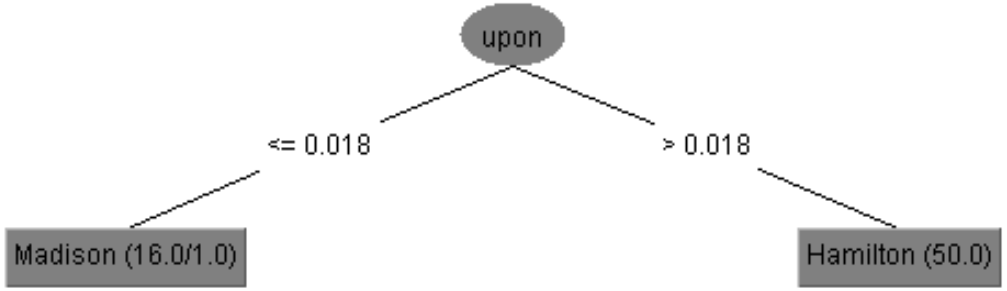
FUNCTION WORDS FOR AUTHORSHIP ATTRIBUTION



The Federalist Papers

This web-friendly presentation of the original text of the Federalist Papers (also known as The Federalist) was obtained from the e-text archives of Project Gutenberg. For more information, see About the Federalist Papers.

No.	Title	Author	Publication	Date
1	General Introduction	Hamilton	For the Independent Journal	- -
2	Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
3	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
4	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
5	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
6	Concerning Dangers from Dissensions Between the States	Hamilton	For the Independent Journal	- -



https://www.congress.gov/resources/display/content/The+Federalist+Papers

GENDER CLASSIFICATION IN GENERAL TEXTS

TABLE 1 (*Continued*)

LIWC Dimension	*Examples*	*Female*		*Male*		*Effect Size (d)*
		M	*SD*	*M*	*SD*	
Pronouns		14.24	4.06	12.69	4.63	0.36
First-person singular	I, me, my	7.15	4.66	6.37	4.66	0.17
First-person plural	we, us, our	1.17	2.15	1.07	2.12	*ns*
Second person	you, you're	0.59	1.05	0.65	1.15	–0.06
Third person	she, their, them	3.41	3.45	2.74	3.01	0.20

Newman, M.L., Groom, C.J., Handelman, L.D., & Pennebaker, J.W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*(3), 211–36.

GENDER CLASSIFICATION IN CONGRESS

Table 4 Gender differences in selected LIWC categories

LIWC dimension	Corpus	Female		Male		Effect size (*d*)	Result
		Mean	SD	Mean	SD		
Pronoun	NGHP	14.24	4.06	12.69	4.63	0.36	Disagree
	HS	7.55	0.01	7.69	0.01	−0.1	

Table 6 Gender differences in pronoun case use

Pronoun cases		Female		Male		Effect size (*d*)
		Mean	SD	Mean	SD	
Subjective	We	1.18	0.40	1.37	0.51	−0.39
	I	1.48	0.32	1.57	0.43	−0.21
Possessive	Our	0.76	0.30	0.58	0.28	0.64
	My	0.46	0.15	0.40	0.17	0.36
Objective	Us	0.22	0.10	0.22	0.10	0.00
	Me	0.15	0.07	0.15	0.08	−0.09

Congresswomen	Congressmen
"Our community"	"Our enemy"
"Our workforce"	"Our side"
"We honor"	"We ought"
"We share"	"We gave"