How to turn words into statistics?

Sandeep Soni

09/03/2024

Today's slides draw heavily from David Bamman's slides on the same topic

Agenda

- What is a word?
- What is a vocabulary?
- How to segment text into words?
- What are the issues that we should care about?
- How to compare different text collections?


Dog

Words represent concepts

- Words are symbolic representations of a concept

Words represent concepts

- Words are symbolic representations of a concept
  - Concise

Words represent concepts

- Words are symbolic representations of a concept
  - Concise
  - Shared

Words represent concepts

- Words are symbolic representations of a concept
  - Concise
  - Shared
  - Memorable

Words represent concepts

- Words are symbolic representations of a concept
  - Concise
  - Shared
  - Memorable
- Words are a form of dimensionality reduction

What are words?

- May the Force be with you
- I'm going to make him an offer he can't refuse.
- And I says, “What kinda action?”
- I’da rather die-died then do that

Word boundaries are not clear-cut in other languages.

What about these?

What about these?

- #MarchMadness

What about these?

- #MarchMadness
- @KamalaHarris

What about these?

- #MarchMadness
- @KamalaHarris
- 9423164156

What about these?

- #MarchMadness
- @KamalaHarris
- 9423164156
- https://www.google.com

Types and tokens

Types and tokens

- Types: Symbolic representation of real-world concepts
- Tokens: Instances of types
- Tokenization: process to segment text (or speech) into tokens

Types and tokens

- Types: Symbolic representation of real-world concepts
- Tokens: Instances of types
- Tokenization: process to segment text (or speech) into tokens

"to be or not to be"

Types and tokens

- Types: Symbolic representation of real-world concepts
- Tokens: Instances of types
- Tokenization: process to segment text (or speech) into tokens

"to be or not to be"

Types	to, be, or, not	4
Tokens	to, be, or, not, to, be	6

Vocabulary

- Types are the unique tokens in a text
- Types are what we typically use to define a lexicon or a vocabulary or a dictionary.

How can we segment Text into Tokens?

Just search?

Just search?

Start with a long vocabulary and then lookup

Just search?

Just search?

- Fast: Linear time to tokenize any text
- Easy: Just lookup in a dictionary

Just search?

- Fast: Linear time to tokenize any text
- Easy: Just lookup in a dictionary

- Not robust: the vs them
- Brittle:supercalifragilisticexpialidocious
- Unmanageable: move, moves, moving

Whitespace

```
text.split(" ")
```

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

Whitespace

```
text.split(" ")
```

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

- Delimiting on whitespace often misses to separate punctuations from words

- `grep -E -o 'wife.?'`

```
853 wife
439 wife,
229 wife.
 93 wife'
 61 wife
 43 wife;
 36 wife?
 28 wife
 27 wife-
 17 wife!
  7 wifey
  7 wifel
  7 wife’
  4 wife_
  2 wifer
  2 wife—
  1 wifeh
  1 wife
```

Sample of ~100 books in PG corpus

Punctuation

- Naively removing punctuation can lead to loss of information about:
  - sentence boundaries
  - pauses
  - parentheticals or asides

Whitespace

Whitespace

- Fast
- Easy
- Less brittle

Whitespace

- Fast
- Easy
- Less brittle

- Lossy or noisy segmentation

Regular expressions

- Instead of separating tokens by whitespace, define a pattern for a token
- A pattern like `'wife.?'` is called a regular expression

Regular expressions

- More formally, regular expressions are a compact language to specify search patterns in text

/single/

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

Regular expressions

- More formally, regular expressions are a compact language to specify search patterns in text

/?n/

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

Regular expressions

- More formally, regular expressions are a compact language to specify search patterns in text

`/[Ii]t/`

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

Regular expressions

regex	matches	doesn't match
/the/	the, isothermally	The
/[Tt]he/	the, isothermally, The	
/\b[Tt]he\b/	the, The	—The

Slide credit to David Bamman

Regular expressions

- Use brackets to specify a range or specific alternations

Regex	Range	Example matches
`[Ii]t`	It or it	It, it
`[a-z]`	{a,b,c,…,z}	for, she,hhh
`[0-9]`	{0,1,2,3,…,9}	123, 0123
`[A-Z]`	{A,B,C,…,Z}	FOR, SHE, HHH
`[a-zA-Z0-9]`	{a,b,..,z,A,B,..,Z,0,1,..,9}	1aSBv2

Regular expressions

Term	Meaning	Sample regex	Matches
+	one or more	he+y	hey, heeeeeey
?	optional	colou?r	color, colour
*	zero or more	toys*	toy, toys, toysss

Slide credit to David Bamman

Symbols

Symbol	Function
\b	Word boundary (zero width)
\d	Any decimal digit (equivalent to [0-9])
\D	Any non-digit character (equivalent to [^0-9])
\s	Any whitespace character (equivalent to [\t\n\r\f\v])
\S	Any non-whitespace character (equivalent to [^ \t\n\r\f\v])
\w	Any alphanumeric character (equivalent to [a-zA-Z0-9_])
\W	Any non-alphanumeric character (equivalent to [^a-zA-Z0-9_])
\t	The tab character
\n	The newline character

https://www.nltk.org/book/ch03.html

Slide credit to David Bamman

Composition of Regexes

- You can create complex regular expressions from simple ones by using disjunction

`/single/|[Ii]t`

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

Python

- Use the `re` module for regex searches
- re.findall (regex, text) will find non-overlapping matches
- `re.findall ("[Ii]t", "If it is meant to be it is meant to be")`
- The function returns a list of all the matches: `[It, it]`
- Also see `re.match`, `re.compile`
- https://www.dataquest.io/wp-content/uploads/2019/03/python-regular-expressions-cheat-sheet.pdf

Back to tokenization

https://spacy.io/usage/spacy-101#annotations-token

Sentence segmentation

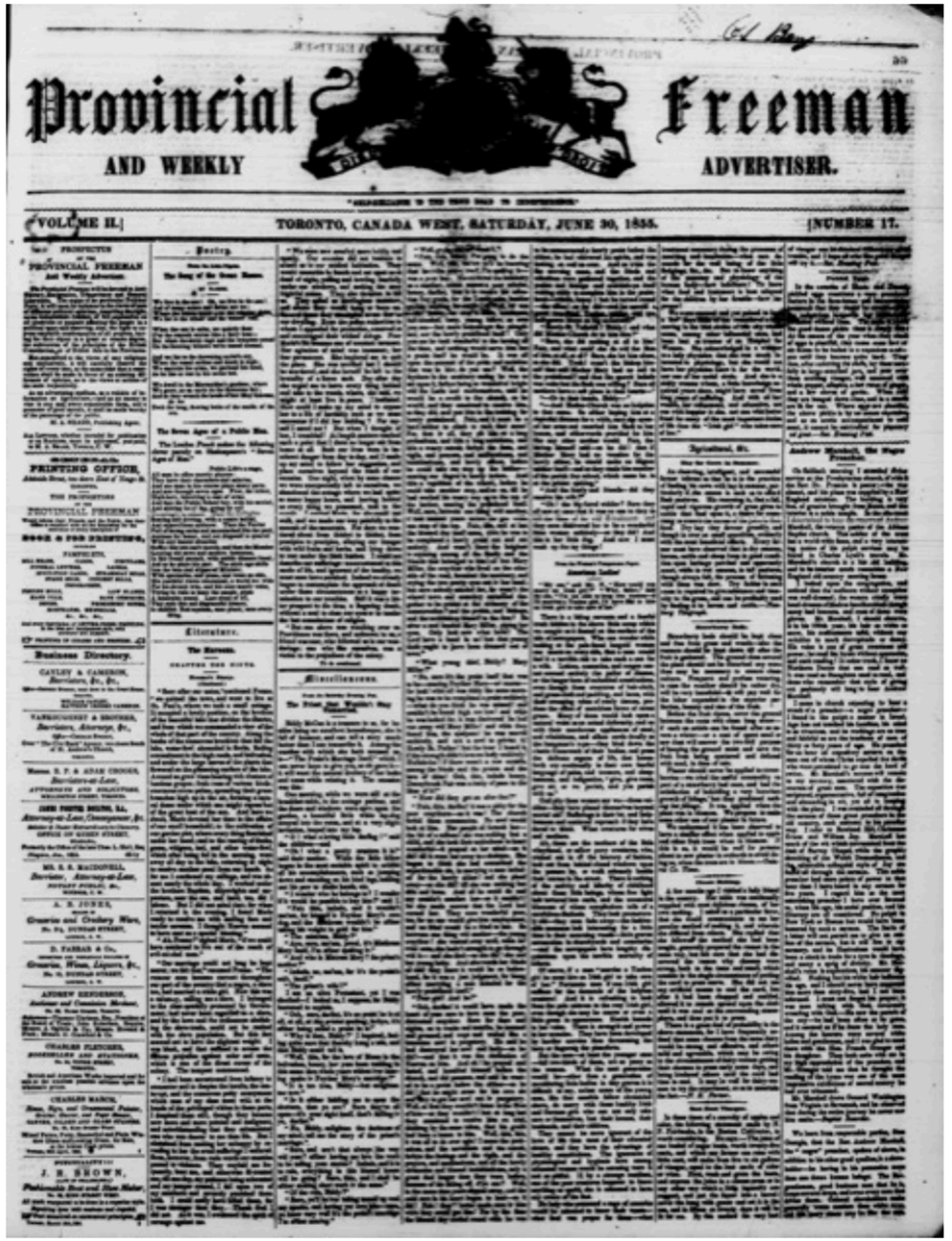
- This is usually not done using regular expressions
- He lives in the U.S. John, however, lives in Canada.
- Typically done using a combination of dependency parsing and unsupervised learning

Stemming and Lemmatization

- Languages have rich inflectional and derivational morphology
- Stemming is a heuristic to clip the suffixes
  - arguing, argues, argue, argued -> argu
- Lemmatization is more principled as it collapses inflectional forms to a typical dictionary entry
  - arguing, argues, argue, argued -> argue

Post processing

- What counts as tokens can have important consequences
- In historical texts, tokenization errors need to be fixed as a post-processing step
- `senatoradmits -> senator admits`

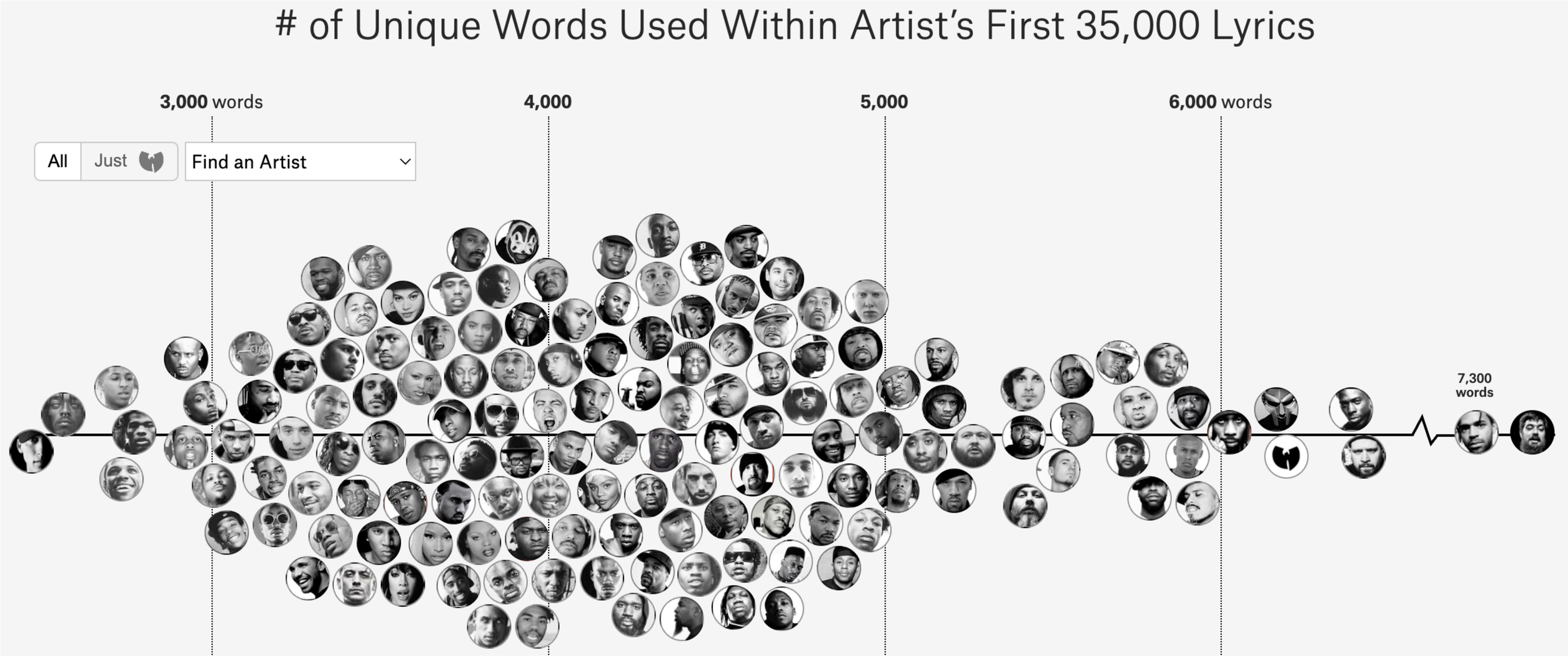

Provincial Freeman
AND WEEKLY ADVERTISER.
VOLUME II.] TORONTO, CANADA WEST, SATURDAY, JUNE 30, 1855. [NUMBER 17.

Soni et. al. 2019

Word analysis

- Vocabulary size and diversity
- Subword analysis
- Word frequency

Vocabulary Size

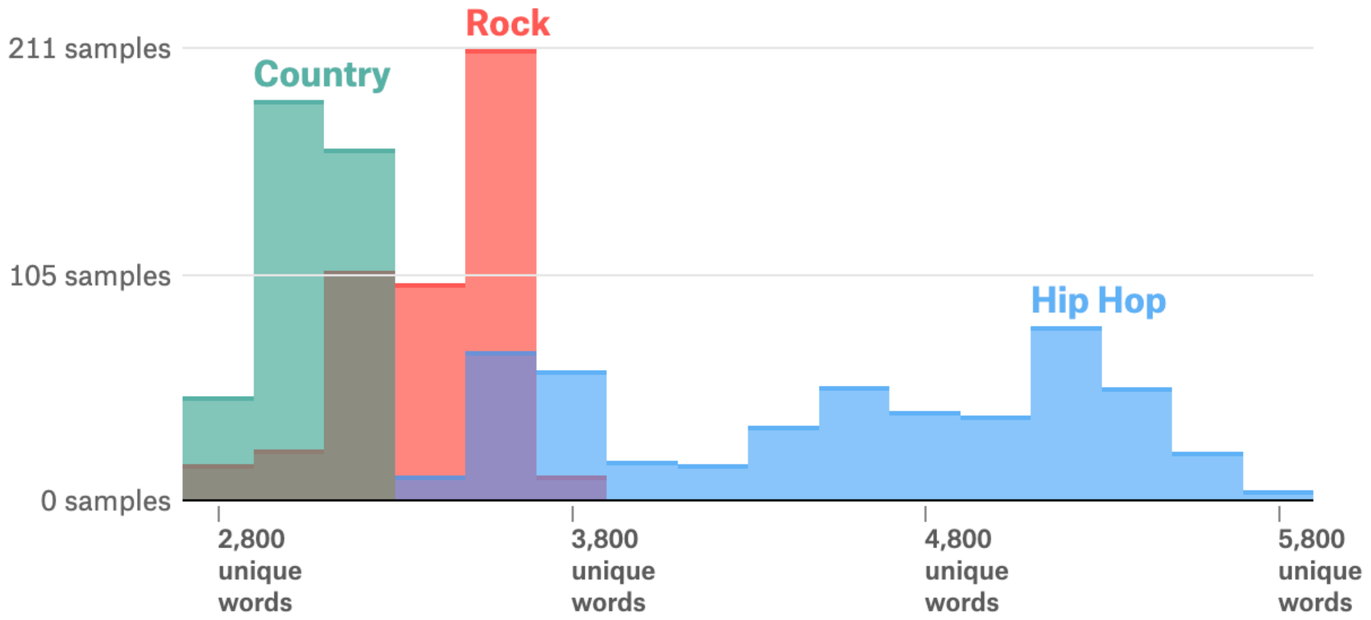


Source: https://pudding.cool/projects/vocabulary/index.html

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

BY ERA[1]

1980s | 1990s | 2000s | 2010s



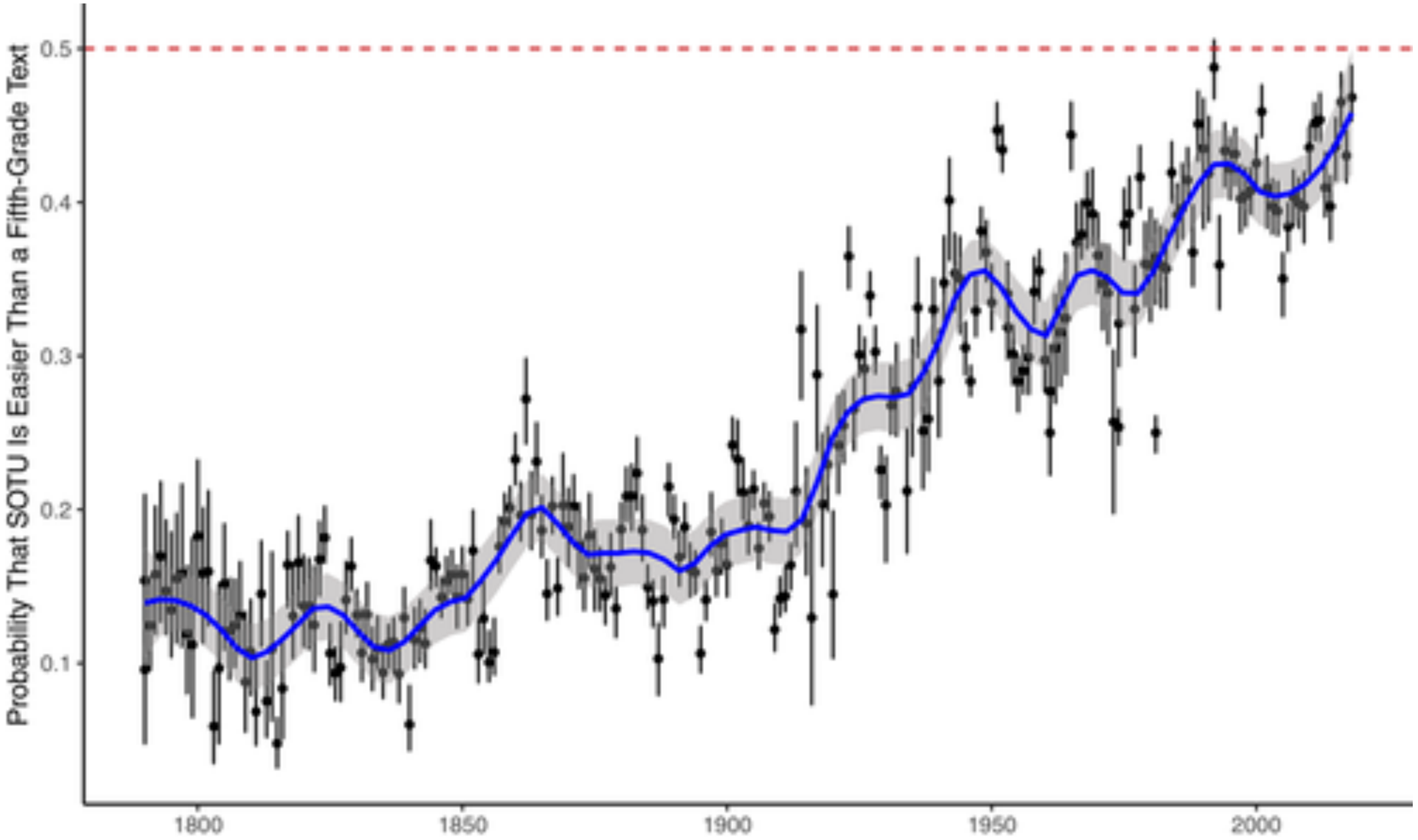
The contraction in vocabulary could be because of the style and structure of some genres

> 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 5 mil - I ain’t been rhyming like Common since

Or it could be that artists make a genuine choice to make songs simpler and shorter

Word Characteristics

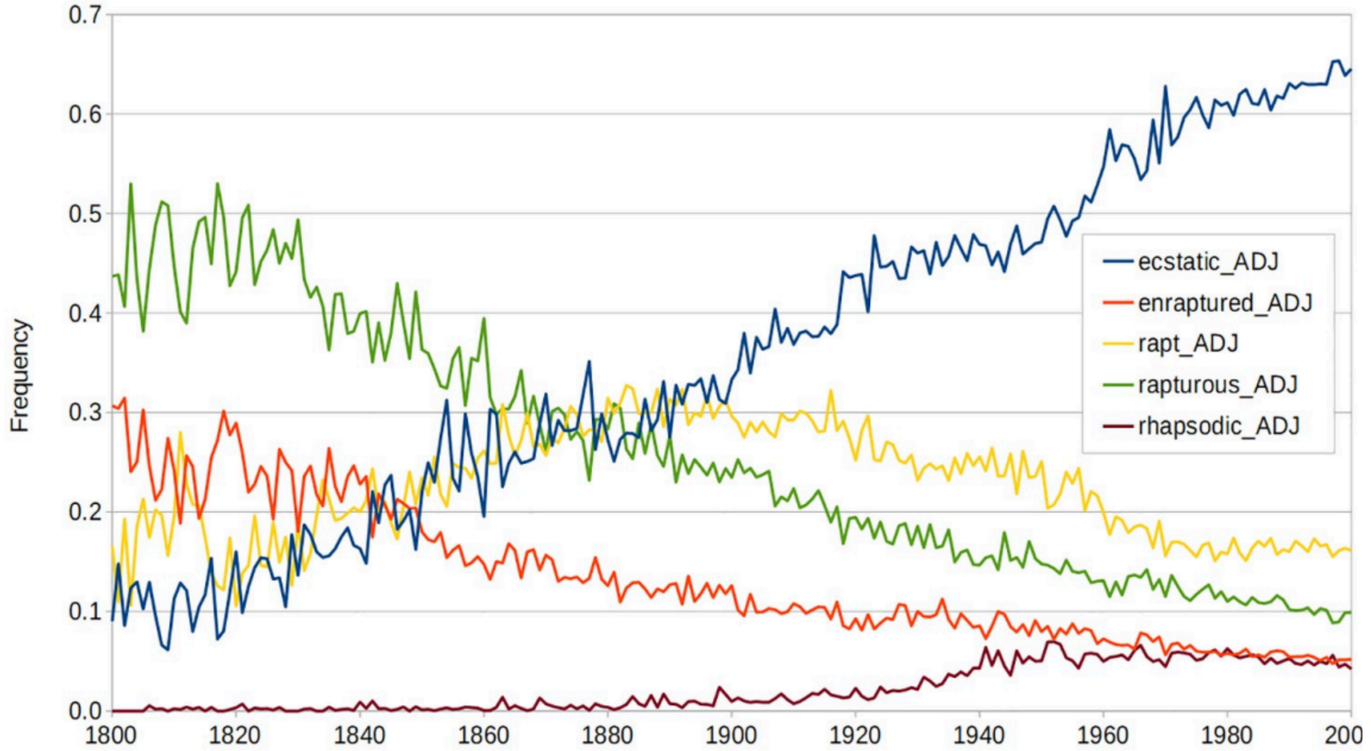
Source of Complexity
Long Words
Mean characters per word
Words with at least 7 characters
Words with at least 6 characters
Mean syllables per word
Words with at least 3 syllables
Words with fewer than 3 syllables
Words with 2 syllables
Words with 1 syllable
Rare Words
Google Books baseline usage
Brown corpus baseline usage



Benoit et. al. (2019) Measuring and Explaining Political Sophistication through Textual Complexity

Word Frequency

- We can create frequency profiles of words and compare them based on their frequencies



How to compare groups using words?

Problem setup

- Assume we have text from two “groups”
  - Twitter in 2014 Vs Twitter in 2024
  - Fiction Vs Non-fiction
  - Donald Trump Speeches Vs Speeches from all other presidents
- Find what differentiates the groups using words and their statistics

Difference in proportions

w	Indexes a word from our lexicon
k_i	Indexes a category
c_{w,k_i}	Count of w in category k_i
$f_{w,k_i} = \frac{c_{w,k_i}}{\sum_{w'} c_{w',k_i}}$	Normalized count or proportion of w in k_i

$f_{w,k_1} - f_{w,k_2}$ is the difference in proportions across two groups

Difference in proportions

Difference in proportions

w="students"

law of large numbers is a rule in math that students learn in school It is not a law though

k_1="maths"

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

Difference in proportions

w="students"

law of large numbers is a rule in math that students learn in school It is not a law though

k_1="maths"

$c_{\texttt{students,maths}} = 1$

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

$c_{\texttt{students,legal}} = 1$

Difference in proportions

w="students"

law of large numbers is a rule in math that students learn in school It is not a law though

k_1="maths"

$c_{\texttt{students,maths}} = 1$

$f_{\texttt{students, maths}} = \frac{1}{20}$

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

$c_{\texttt{students,legal}} = 1$

$f_{\texttt{students, legal}} = \frac{1}{20}$

Difference in proportions

w="students"

law of large numbers is a rule in math that students learn in school It is not a law though

k_1="maths"

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

$$c_{\texttt{students,maths}} = 1$$

$$f_{\texttt{students, maths}} = \frac{1}{20}$$

$$c_{\texttt{students,legal}} = 1$$

$$f_{\texttt{students, legal}} = \frac{1}{20}$$

$$f_{\texttt{students,maths}} - f_{\texttt{students,legal}} = \frac{1}{20} - \frac{1}{20} = 0$$

Difference in proportions

w="law"

law of large numbers is a rule in math that students learn in school It is not a law though

k_1="maths"

$c_{\texttt{law,maths}} = ?$

$f_{\texttt{law, maths}} = ?$

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

$c_{\texttt{law,legal}} = ?$

$f_{\texttt{law, legal}} = ?$

$$f_{\texttt{law,maths}} - f_{\texttt{law,legal}} = ?$$

Difference in proportions

w="law"

law of large numbers is a rule in math that students learn in school
It is not a law though

k_1="maths"

$$c_{\texttt{law,maths}} = 2$$

$$f_{\texttt{law,maths}} = \frac{2}{20}$$

The law school teaches law to students who want to learn the law and practice law in their future careers

k_2="legal"

$$c_{\texttt{law,legal}} = 4$$

$$f_{\texttt{law,legal}} = \frac{4}{20}$$

$$f_{\texttt{law, maths}} - f_{\texttt{law,legal}} = \frac{2}{20} - \frac{4}{20} = -0.1$$

Difference in proportions

- Simple and easy to measure and interpret
- Overemphasizes common words; for common words, there differences are also large
- No correction for chance or determination of statistical significance

$$\chi^2$$

χ^2

Does the word “robot” occur significantly more frequently in science fiction?

χ^2

Does the word “robot” occur significantly more frequently in science fiction?

	robot	¬robot	
sci-fi	104	1004	= 10.3%
¬sci-fi	2	13402	= 0.015%

Slide credit: David Bamman’s Info 256 class

χ^2

We can calculate the following statistic, which is the sum of squared difference between the observed value in each cell and the expected value assuming independence

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Slide adapted from David Bamman's class

χ^2

	robot	¬robot	sum	frequency
sci-fi	104	1004	1108	0.076
¬sci-fi	2	13402	13404	0.924
sum	106	14406		
frequency	0.007	0.993		

Assuming independence:

$$P(\text{robot}, \text{scifi}) = P(\text{robot}) \times P(\text{scifi})$$
$$= 0.007 \times 0.076 = 0.00053$$

Among 14512 words, we would expect to see 7.69 occurrences of *robot* in sci-fi texts.

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

$P(\text{scifi})$	0.076
$P(\neg\text{scifi})$	0.924

$P(\text{robot})$	$P(\neg\text{robot})$
0.007	0.993

Slide adapted from David Bamman's class

χ^2

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

Left is observed counts; right is expected counts assuming complete independence

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Slide adapted from David Bamman's class

χ^2

How different are these two tables?

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

Left is observed counts; right is expected counts assuming complete independence

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

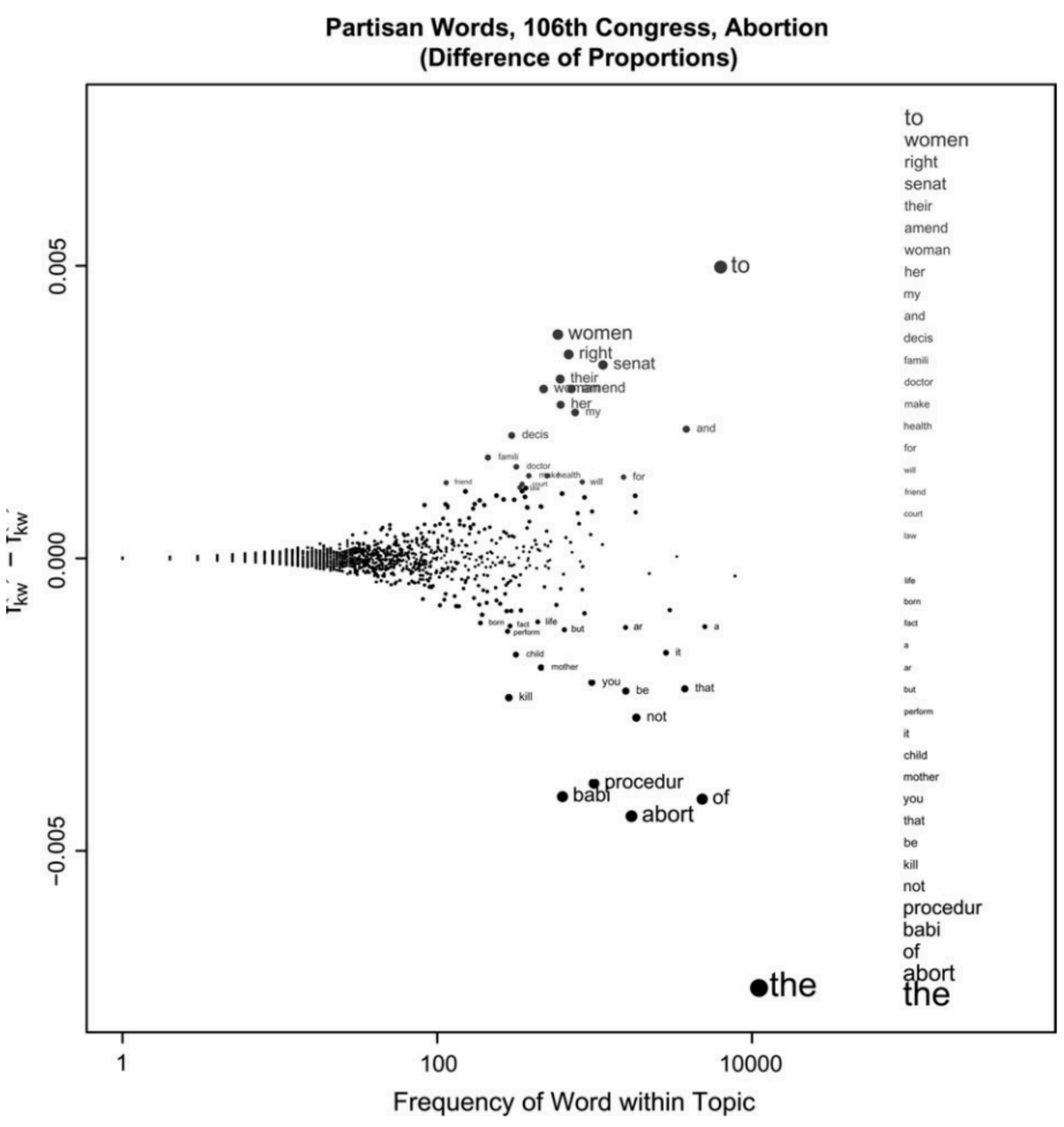
Slide adapted from David Bamman's class

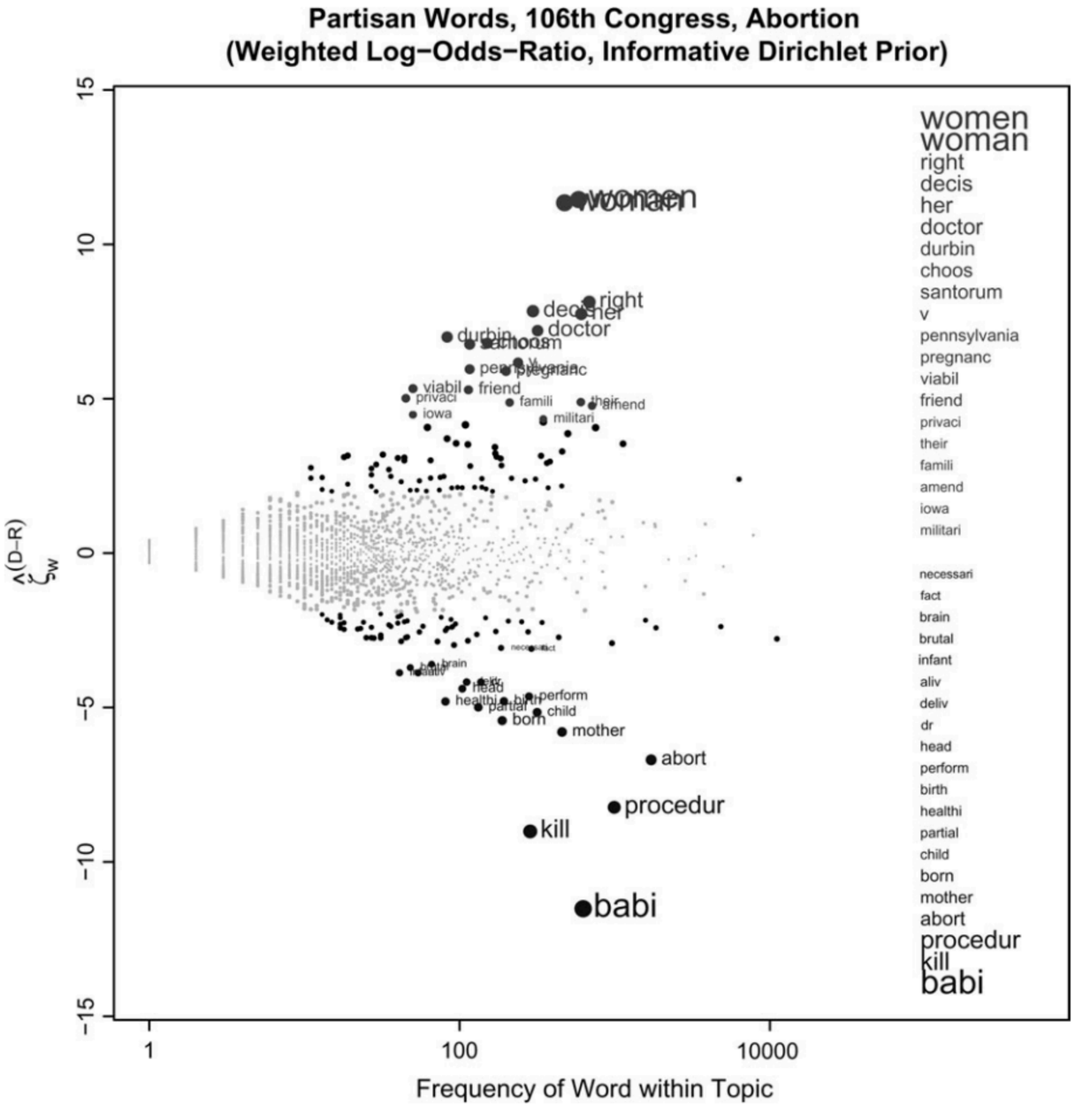
χ^2

- Useful statistic to find differentiating markers
- We have a way to test for statistical significance
- Assumes each word is independent from the others

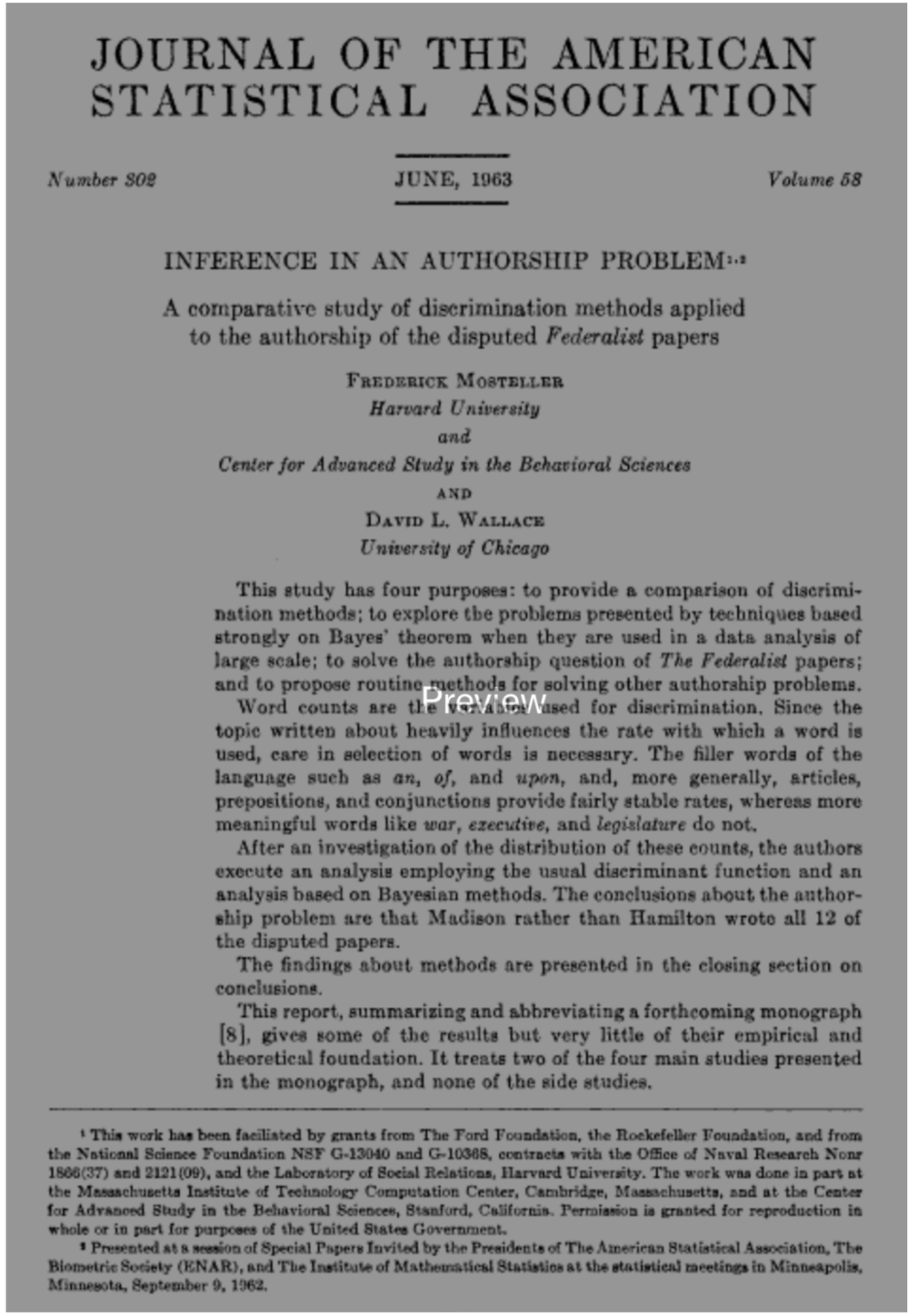
Other methods

- Many other methods to characterize differences
  - Model based methods that assume parametric distributions and Bayesian priors (e.g., Monroe et. al. 2009)
  - Unsupervised and Bayesian (e.g., SAGE; Eisenstein et. al. 2011)
  - Supervised learning to learn precise features that are informative in separating categories (e.g., Underwood et. al. 2018)


Partisan Words, 106th Congress, Abortion
(Difference of Proportions)
0.005
0.000
−0.005
1
100
10000
Frequency of Word within Topic
to
women
right
senat
their
amend
woman
her
my
and
decis
famili
doctor
make
health
for
will
friend
court
law
life
born
fact
a
ar
but
perform
it
child
mother
you
that
be
kill
not
procedur
babi
of
abort
the


Partisan Words, 106th Congress, Abortion
(Weighted Log−Odds−Ratio, Informative Dirichlet Prior)
15
10
5
0
−5
−10
−15
1
100
10000
Frequency of Word within Topic
women
woman
right
decis
her
doctor
durbin
choos
santorum
v
pennsylvania
pregnanc
viabil
friend
privaci
their
famili
amend
iowa
militari
necessari
fact
brain
brutal
infant
aliv
deliv
dr
head
perform
birth
healthi
partial
child
born
mother
abort
procedur
kill
babi

- Who wrote the disputed 12 federalist papers?
- Compare the frequency of some basic stylistic words (e.g., upon)
- Answer: Madison



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of *The Federalist* papers; and to propose routine methods for solving other authorship problems.

Word counts are the basic data used for discrimination. Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like *war*, *executive*, and *legislature* do not.

After an investigation of the distribution of these counts, the authors execute an analysis employing the usual discriminant function and an analysis based on Bayesian methods. The conclusions about the authorship problem are that Madison rather than Hamilton wrote all 12 of the disputed papers.

The findings about methods are presented in the closing section on conclusions.

This report, summarizing and abbreviating a forthcoming monograph [8], gives some of the results but very little of their empirical and theoretical foundation. It treats two of the four main studies presented in the monograph, and none of the side studies.

[1] This work has been facilitated by grants from The Ford Foundation, the Rockefeller Foundation, and from the National Science Foundation NSF G-13040 and G-10368, contracts with the Office of Naval Research Nonr 1866(37) and 2121(09), and the Laboratory of Social Relations, Harvard University. The work was done in part at the Massachusetts Institute of Technology Computation Center, Cambridge, Massachusetts, and at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. Permission is granted for reproduction in whole or in part for purposes of the United States Government.

[2] Presented at a session of Special Papers Invited by the Presidents of The American Statistical Association, The Biometric Society (ENAR), and The Institute of Mathematical Statistics at the statistical meetings in Minneapolis, Minnesota, September 9, 1962.

In class

- Tokenization demo
- Distinctive terms demo