Finding Distinctive terms

Sandeep Soni

01/25/2024

Story so far

- Words are linguistic units that reference concepts
- Type Vs Token
- 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

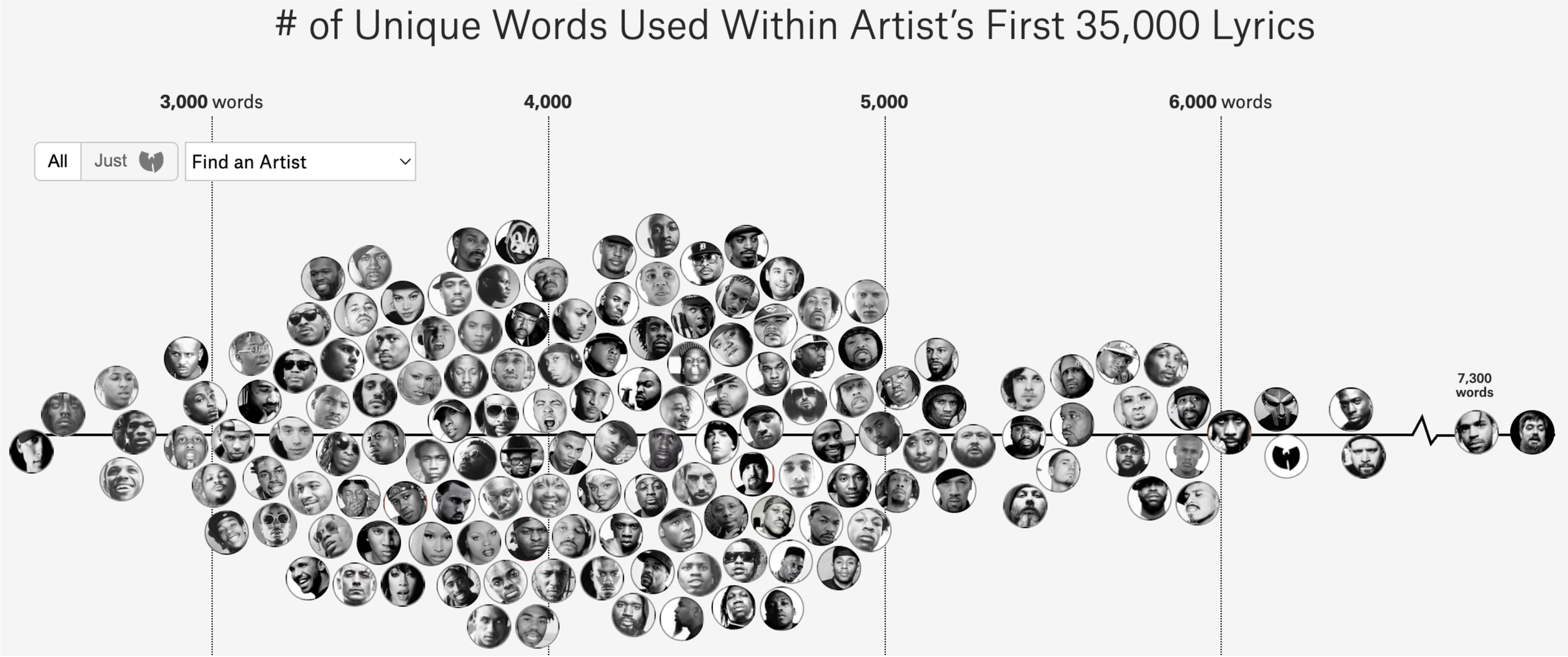
Story so far

- Tokenization is a typical pre-processing step
- Tokenization is done by matching regular expressions
- Lexicon or vocabulary is a collection of word types
- Stemming/lemmatization can reduce the vocabulary size

Word analysis

- Vocabulary size and diversity
- Subword analysis
- Word frequency

Vocabulary Size

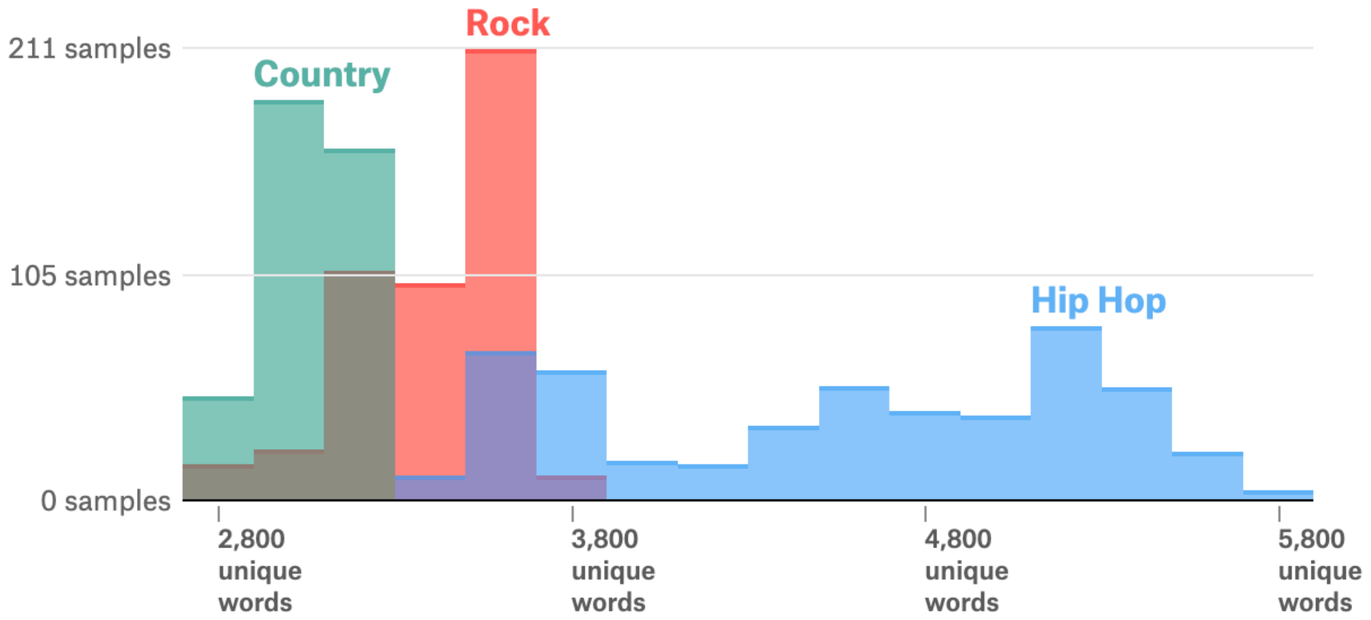


Source: https://pudding.cool/projects/vocabulary/index.html

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

BY ERA[1]

1980s | 1990s | 2000s | 2010s



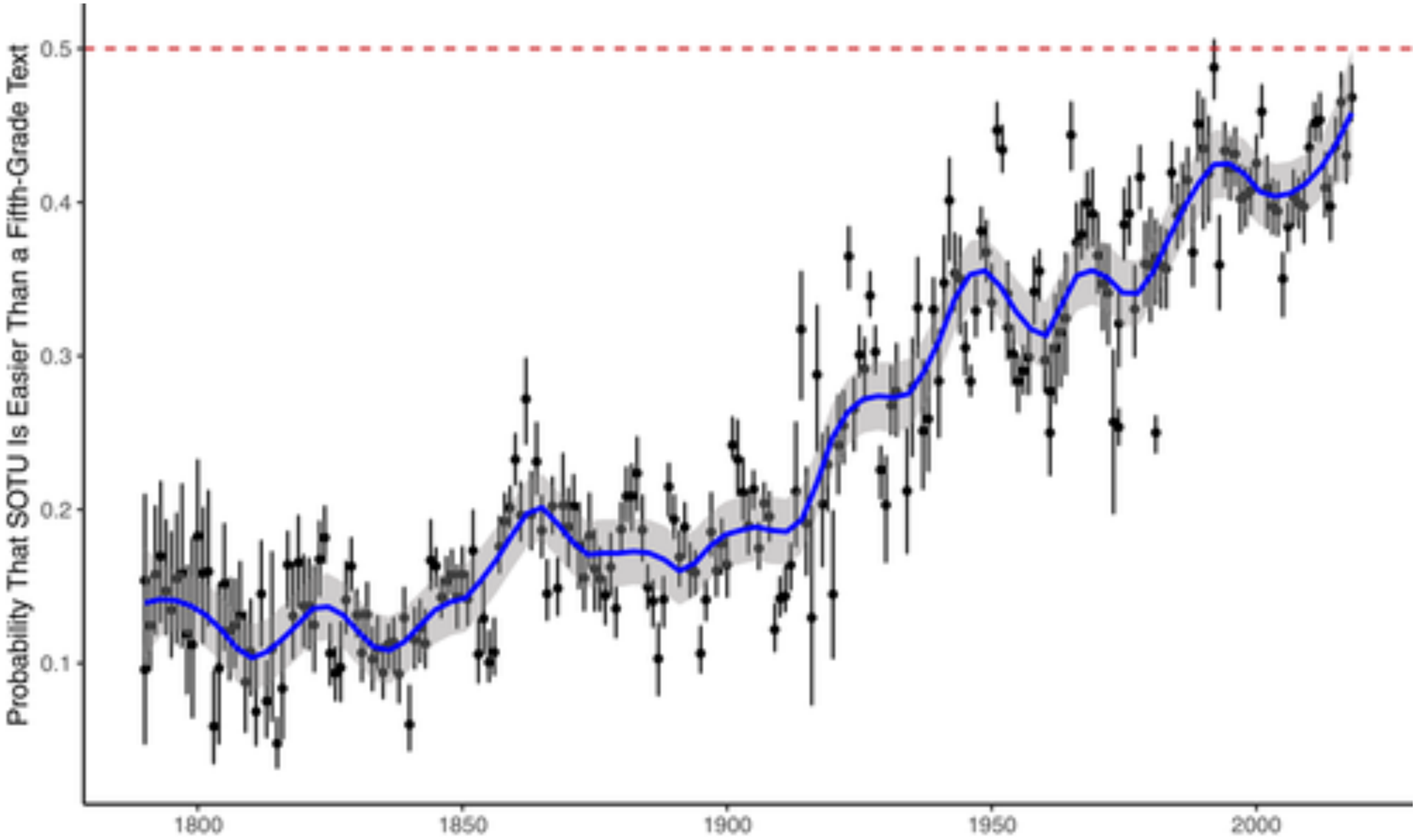
The contraction in vocabulary could be because of the style and structure of some genres

> 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

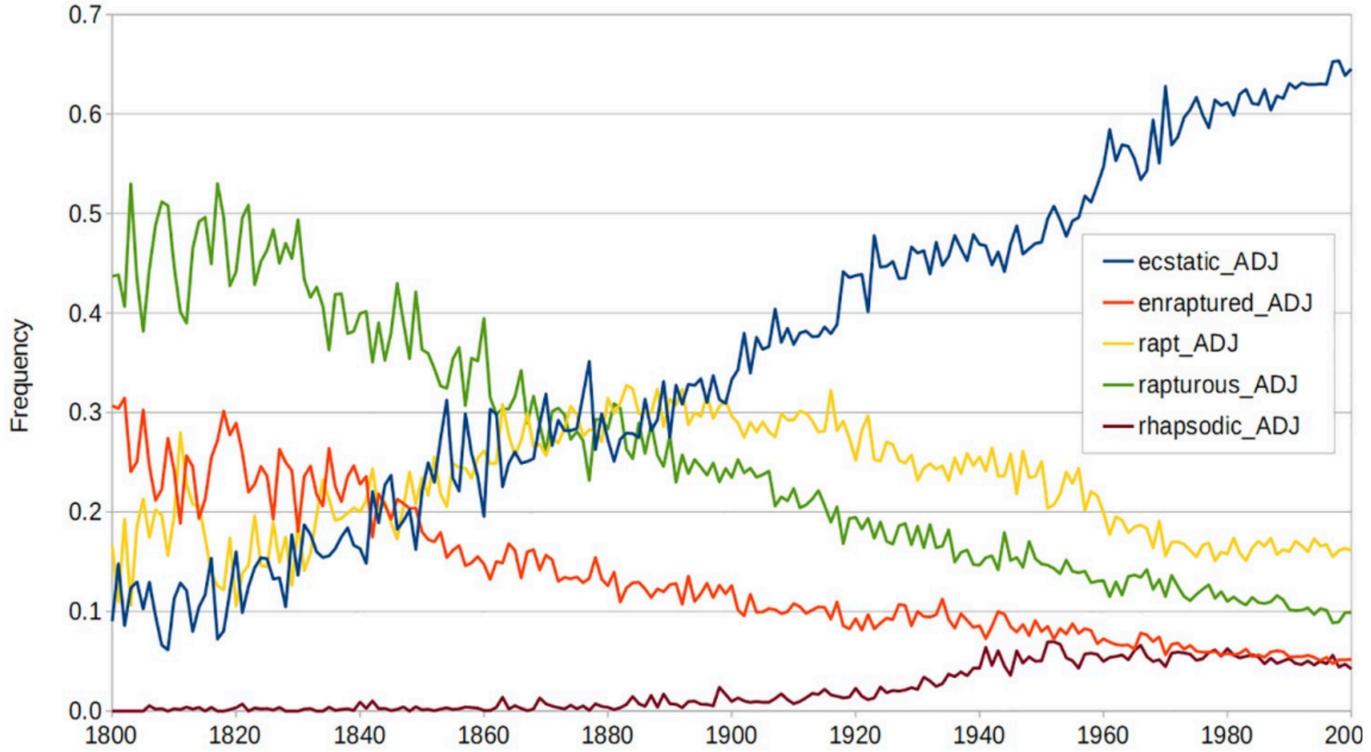
Source of Complexity
Long Words
Mean characters per word
Words with at least 7 characters
Words with at least 6 characters
Mean syllables per word
Words with at least 3 syllables
Words with fewer than 3 syllables
Words with 2 syllables
Words with 1 syllable
Rare Words
Google Books baseline usage
Brown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



Question for the day

"How to compare groups using words?"

Problem setup

- Given two text collections corresponding to two “groups”
  - SotU speeches from 1800s Vs 1900s
  - Fiction Vs Non-fiction
  - r/politics Vs everything else
- Find if the two groups are different and what words differentiate the two groups

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"

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

k_2="legal"

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

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

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

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

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

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

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

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

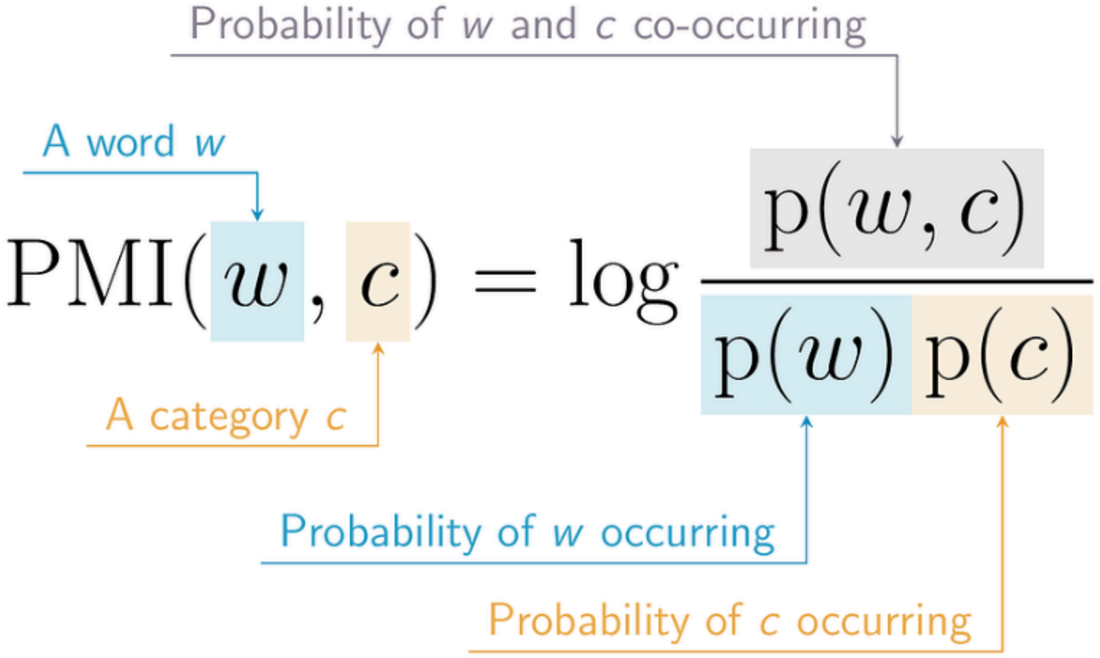
Slide adapted from David Bamman's class

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

Pointwise Mutual Information

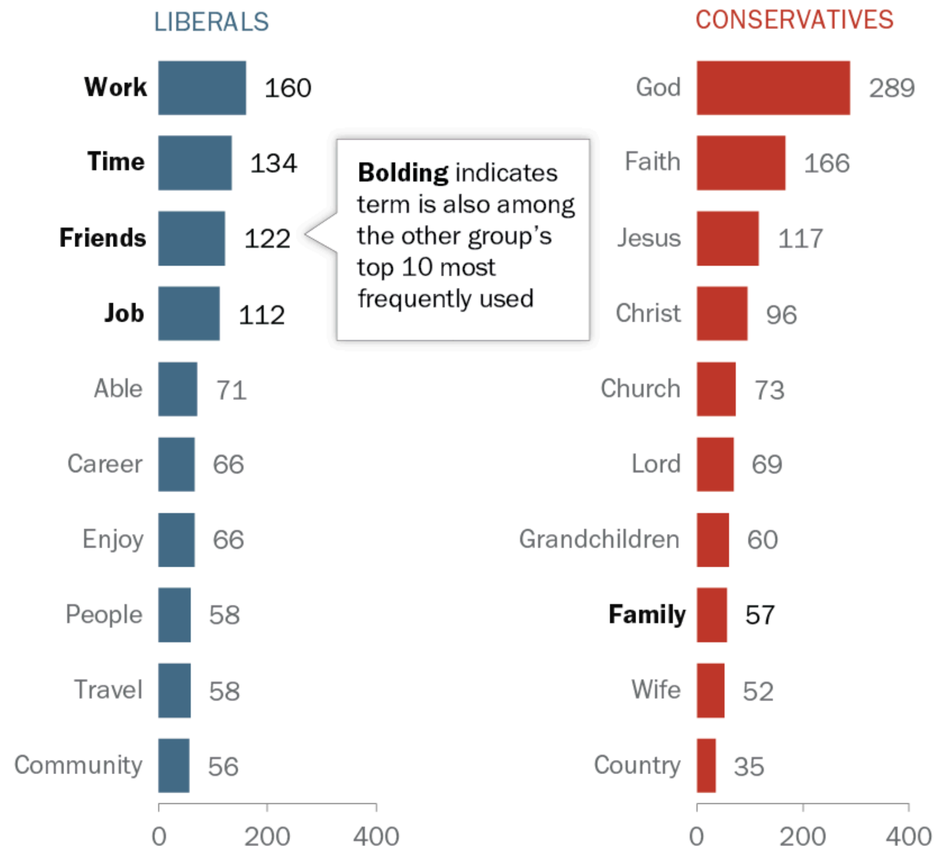
- An information theoretic approach
- Generalizes to more than 2 categories
- Statistical significance can be calculated using non-parametric methods



Source: Bestvater and Shah; Pew research article

Identifying terms used more often by one group than another doesn't always indicate distinctivenes

*Top 10 terms used **more** frequently by ___, by difference in word count*

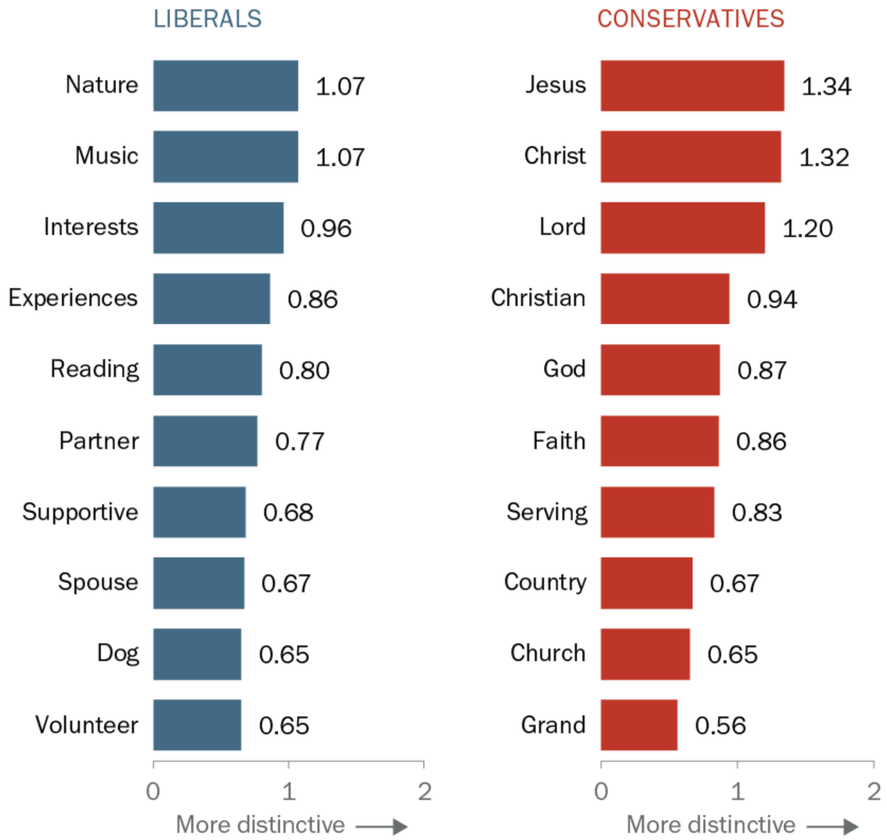


Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information helps identify words that are used distinctively by one group, not just used more

Most dinstinctive terms used by each group, by PMI score



Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information emphasizes different words than simple term frequencies or frequency differences

Top 10 terms used by conservatives, as measured by ...

Rank	Frequency	Difference in frequency	PMI
1	Family	God	Jesus
2	God	Faith	Christ
3	Friends	Jesus	Lord
4	Children	Christ	Christian
5	Health	Church	God
6	Love	Lord	Faith
7	Time	Grandchildren	Serving
8	Work	Family	Country
9	Faith	Wife	Church
10	Job	Country	Grand
.....			
60			Family

Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information emphasizes different words than simple term frequencies or frequency differences

Top 10 terms used by conservatives, as measured by ...

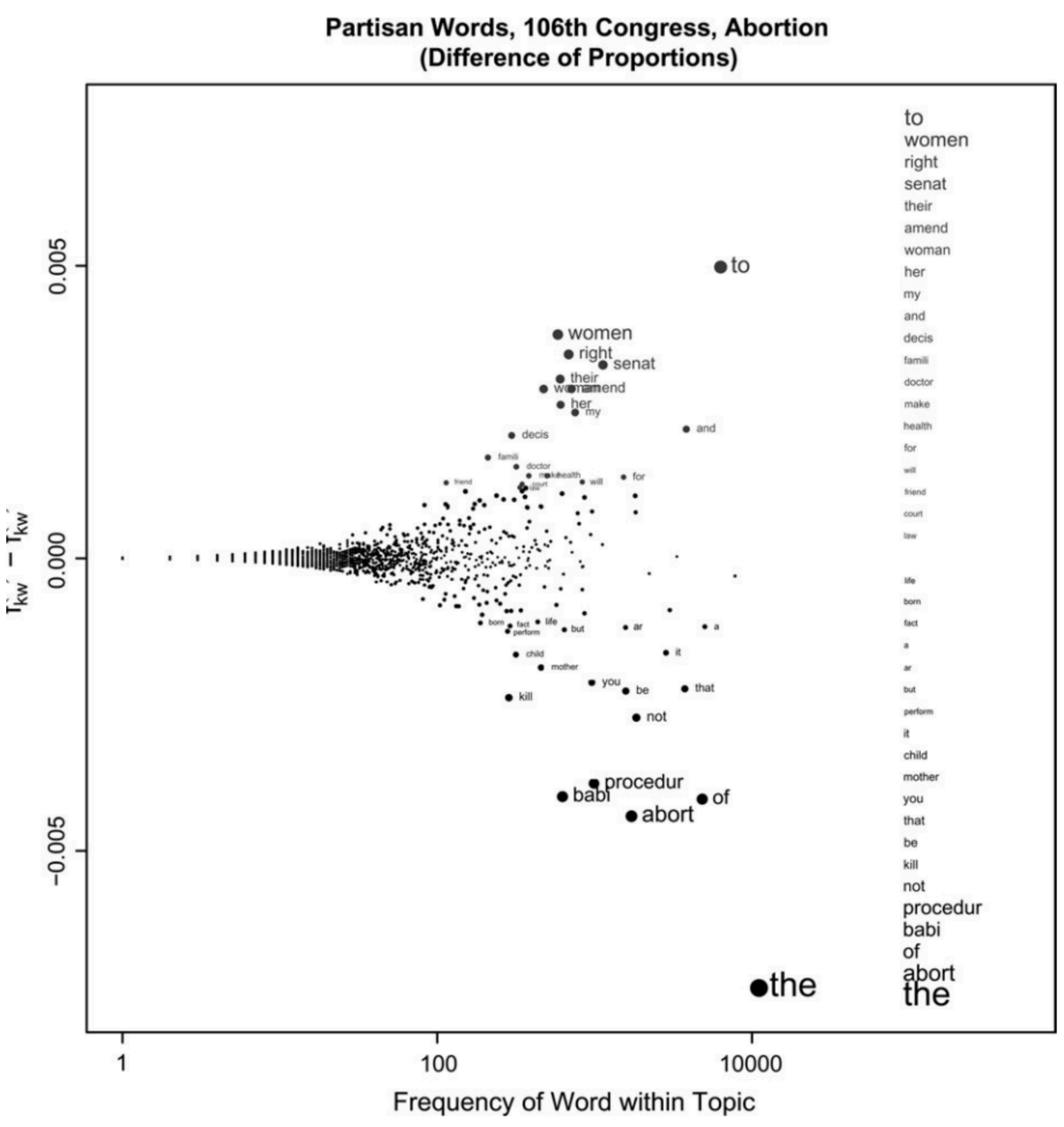
Rank	Frequency	Difference in frequency	PMI
1	Family	God	Jesus
2	God	Faith	Christ
3	Friends	Jesus	Lord
4	Children	Christ	Christian
5	Health	Church	God
6	Love	Lord	Faith
7	Time	Grandchildren	Serving
8	Work	Family	Country
9	Faith	Wife	Church
10	Job	Country	Grand
.....			
60			Family

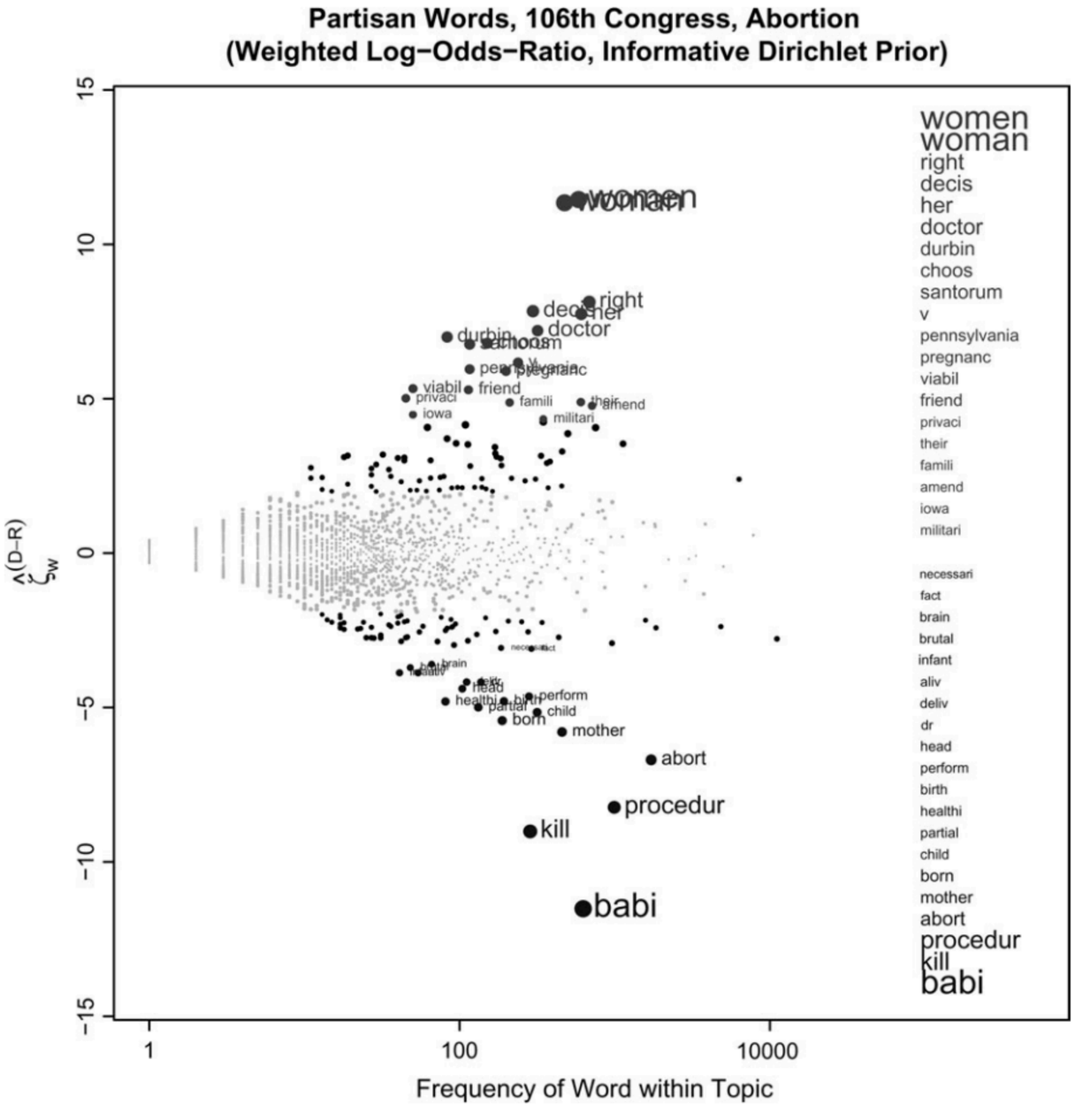
Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

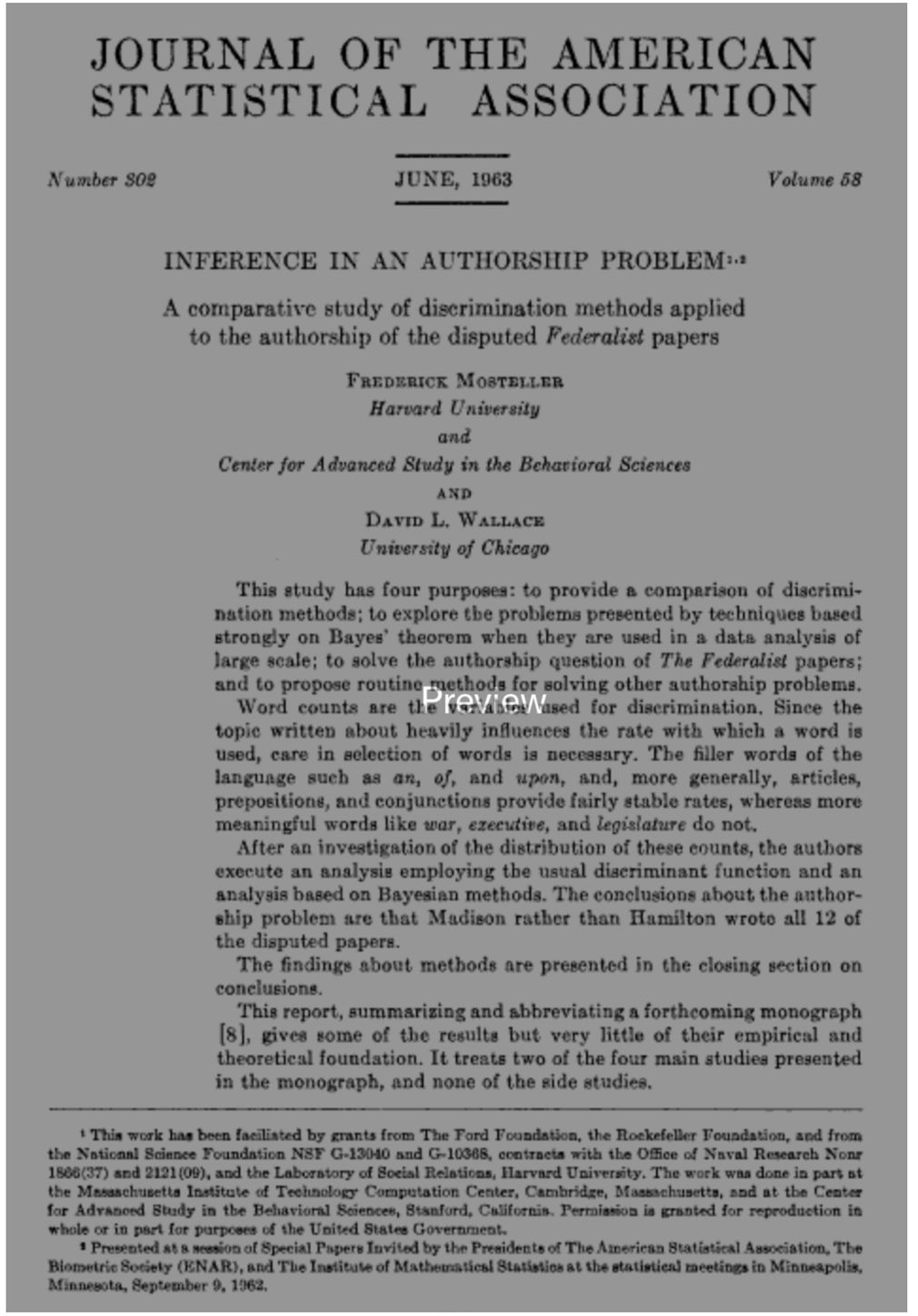
Other methods

- 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 variables 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

- Distinctive terms demo