CSCI 491/591: Data Visualization

28- Text Visualization-1

"The contents of this presentation have been prepared based on Jason Chuang's lecture on text visualization at the University of Washington."

What is considered as ”text”?

- Documents (Articles, books and novels, E-mails, web pages, blogs)
- Text Snippets (Tweets, SMS messages, Tags, comments, profiles)
- And More... (Computer programs, logs, even this slide!)

Why visualize text?

- Understanding – read a document
- Summaries – get the "gist" of a document
- Clustering – group together similar contents
- Quantify – convert to numerical measures
- Correlate – compare patterns in text to those in other data, e.g., correlate with social network

Example: Health Care Reform

- Initiatives by President Clinton (Clinton healthcare plan)
- Overhaul by President Obama (the Affordable Care Act (ACA))
- Text data
  - News articles
  - Speech transcriptions
  - Legal documents
- **What questions might you want to answer?**
- **What visualizations might help?**

Example: Health Care Reform

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

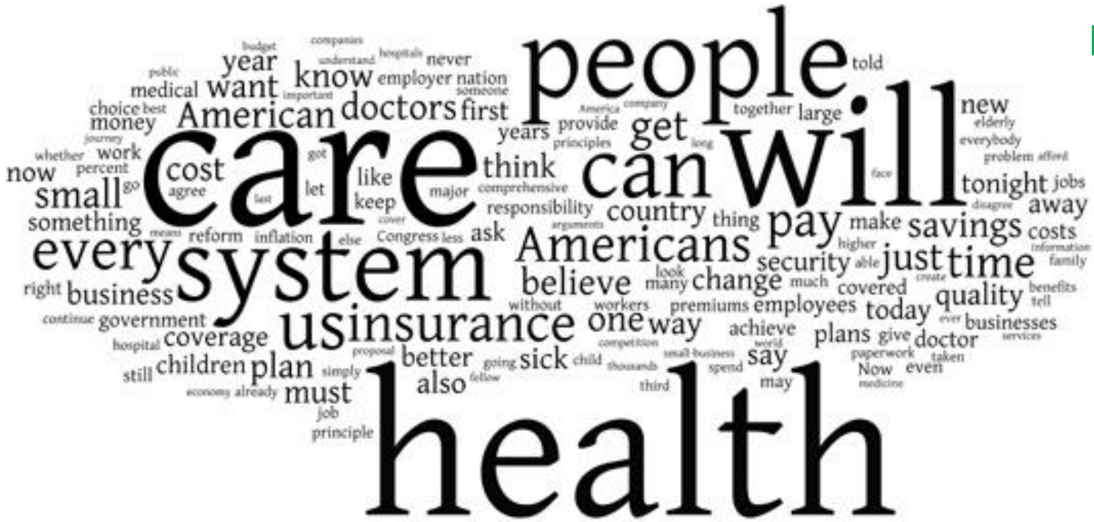
I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo

Tag Clouds: Word Count

- President Obama's Health Care Speech to Congress [New York Times]

https://archive.nytimes.com/economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93/

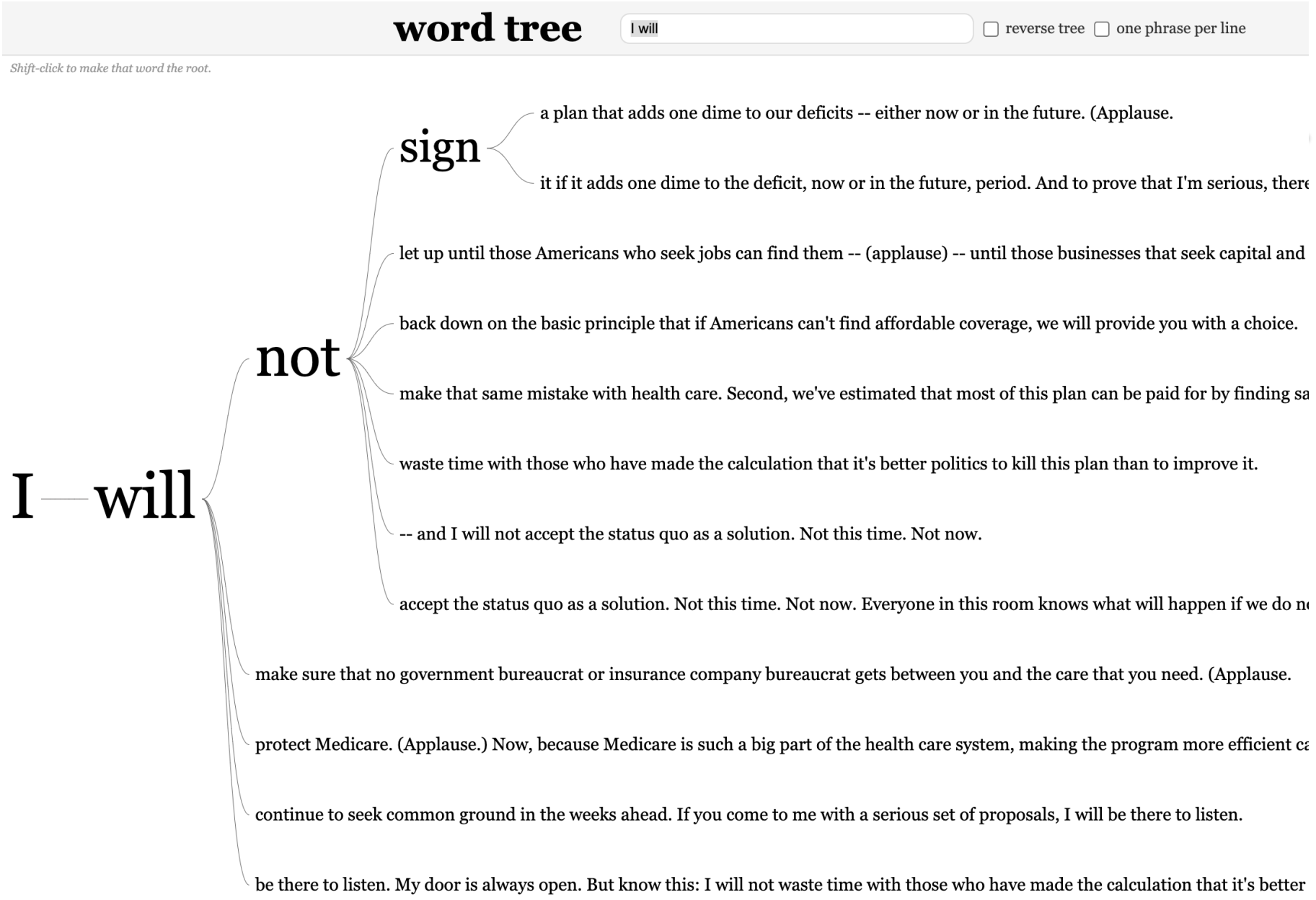


President Clinton, 1993

President Obama, 2009

WordTree: Word Sequences

https://www.jasondavies.com/wordtree



https://www.jasondavies.com/wordtree

Text mining models

- Many (most?) text visualizations do not represent the text directly. They represent the output of a language model (word counts, word sequences, etc.).
- Can you interpret the visualization? How well does it convey the properties of the model?

Topics

- Summarizing with Words
- Visualizing Themes in a Document Collection
- Quantifying Textual Content
- Performing Text Analysis

Summarize with Words

- What kind of data are words? Are they nominal?
- High dimensional (10,000+)
- Words have meanings and relations
  - Correlations: Hong Kong, San Francisco, Bay Area
  - Order: April, February, January, June, March, May
  - Membership: Tennis, Running, Swimming, Hiking, Piano
  - Hierarchy, antonyms & synonyms, entities, …

Text Processing Pipeline

1. Tokenization
   - Segment text into terms. "Text mining is fascinating → ["Text", "mining", "is", "fascinating"]
   - Remove stop words? a, an, the, of, to, be
   - Numbers and symbols? #GocatsGo, @MSUBobcats
   - Entities? San Francisco, O'Connor, U.S.A.
2. Stemming
   - Group together different forms of a word.
   - Porter stemmer? visualization(s), visualize(s), visually → visual
   - Lemmatization? goes, went, gone → go
3. Ordered list of terms Alphabetical, Frequency, Semantic similarity, Part-of-speech

Bag of Words Model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
  - Each dimension corresponds to a term (10,000+)
  - Each value represents the relevance, i.e, term counts
- Aggregate into a document-term matrix
  - Document vector space model

Bag of Words Model Example

Here's a simplified example of how the Bag of Words model works:

Suppose you have two documents:

- Document 1: "Text mining is fascinating."
- Document 2: "NLP and text mining are related."

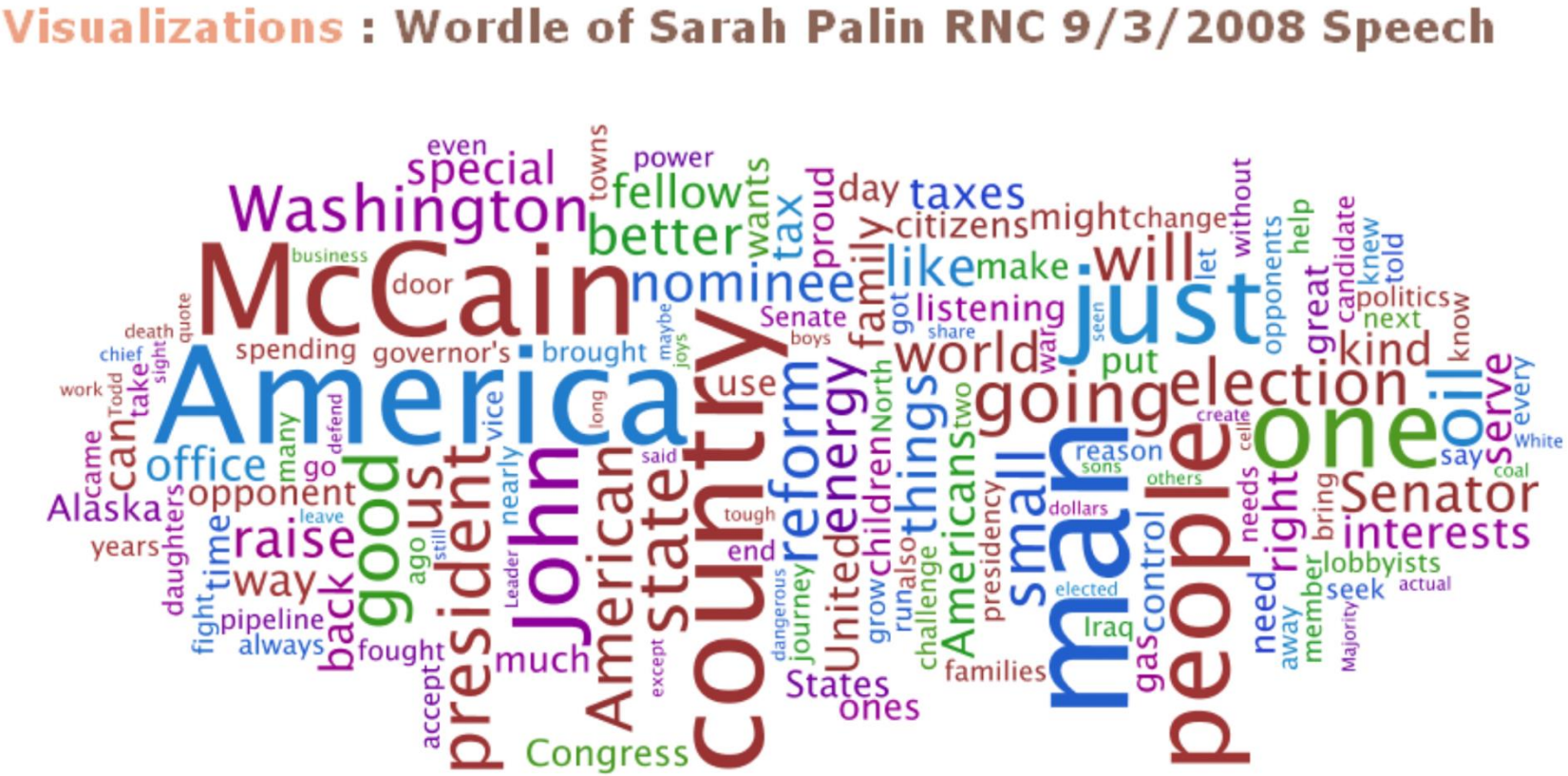
After tokenization and removing punctuation, the vocabulary (unique words) for the entire corpus is:

- ["Text", "mining", "is", "fascinating", "NLP", "and", "are", "related"]

The BoW representation of the documents is then:

- Document 1: [1, 1, 1, 1, 0, 0, 0, 0]
- Document 2: [0, 1, 0, 0, 1, 1, 1, 1]

WordTag Clouds



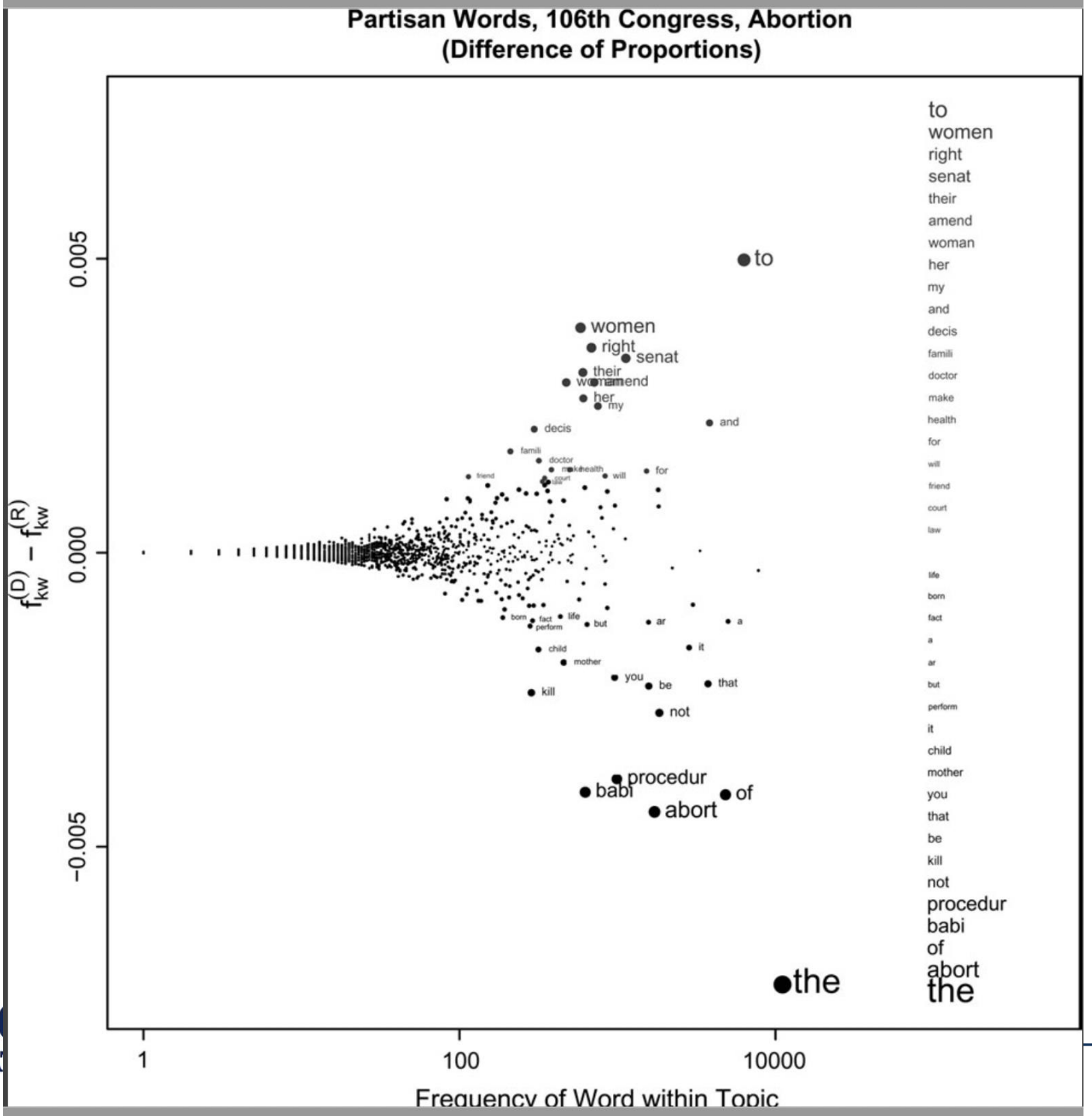
Tag Clouds

- Strengths
  - Can help with gisting and initial query formation.
- Weaknesses
  - Sub-optimal visual encoding (size vs. position)
  - Inaccurate size encoding (long words are bigger)
  - May not facilitate comparison (unstable layout)
  - Term frequency may not be meaningful
  - Does not show the structure of the text

Keyword Weighting

Term Frequency

- tf_{td} = count(t) in d
- Can take log frequency: $\log(1 + tf_{td})$
- Can normalize to show proportion: $tf_{td} / \Sigma_t\ tf_{td}$


Partisan Words, 106th Congress, Abortion
(Difference of Proportions)
$f_{kw}^{(D)} - f_{kw}^{(R)}$
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

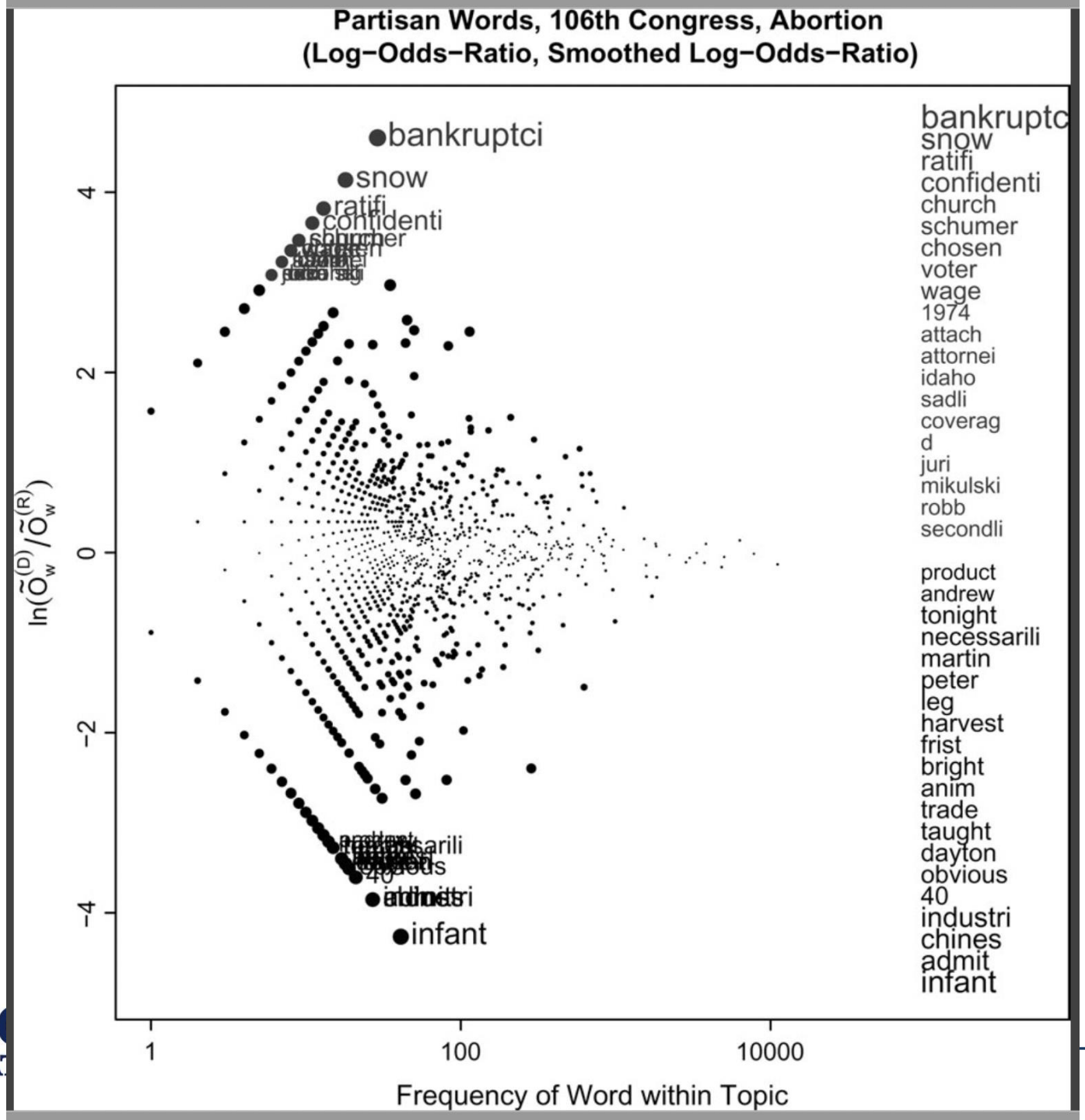
Keyword Weighting

Term Frequency

- tf_{td} = count(t) in d
- Can take log frequency: $\log(1 + tf_{td})$
- Can normalize to show proportion: $tf_{td} / \Sigma_t\ tf_{td}$
- It measures the importance of a term within a single document. A higher term frequency indicates that the word is more significant in the context of the document.

TF.IDF: Term Freq by Inverse Document Freq

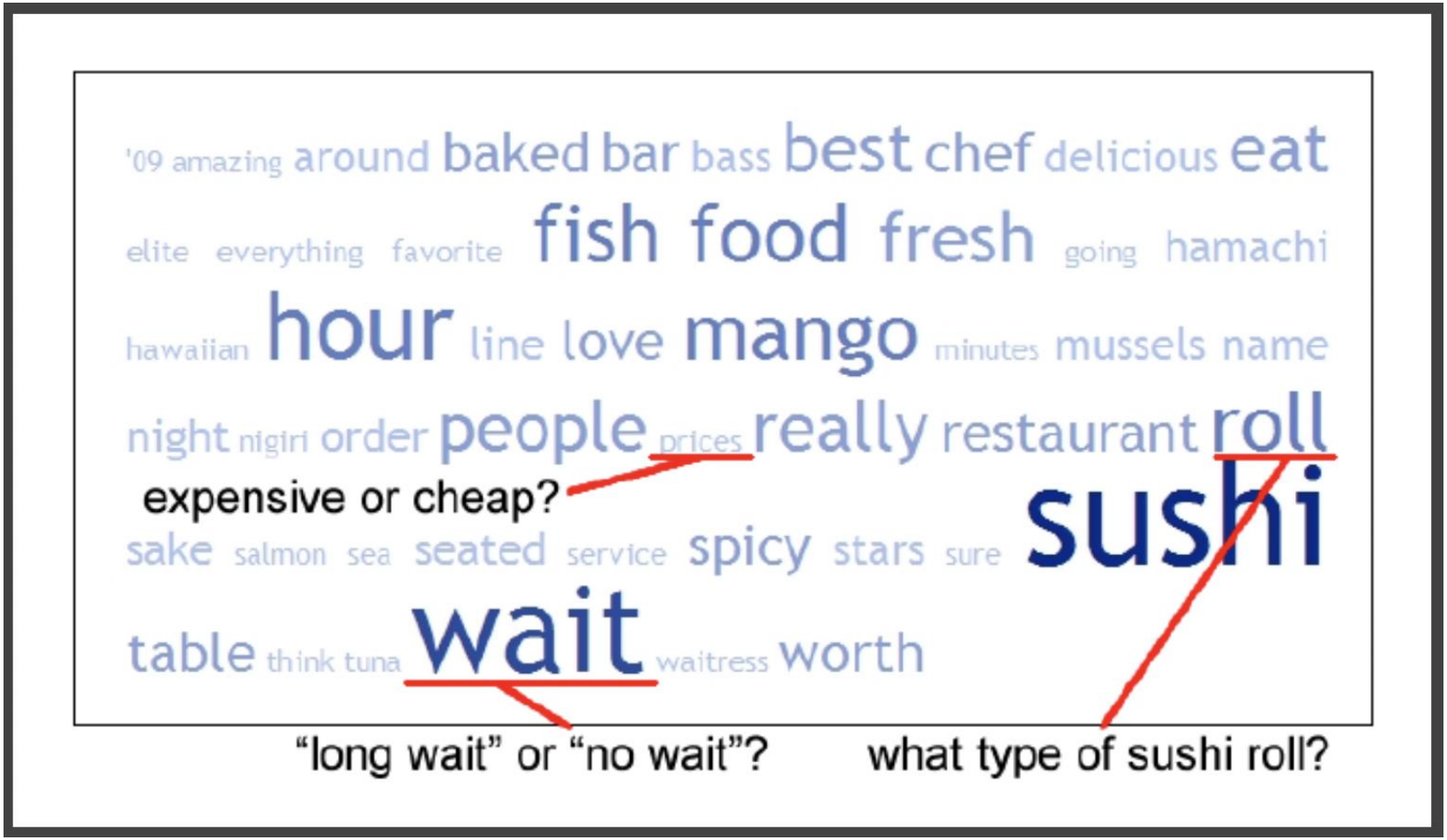
- $tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$
- df_t = # docs containing t; N = # of docs
- IDF measures the importance of a term across the entire corpus. A higher IDF value indicates that the word is less common across the documents and, therefore, more distinctive or informative.


Partisan Words, 106th Congress, Abortion
(Log−Odds−Ratio, Smoothed Log−Odds−Ratio)
bankruptci
snow
ratifi
confidenti
church
schumer
chosen
voter
wage
1974
attach
attornei
idaho
sadli
coverag
d
juri
mikulski
robb
secondli
product
andrew
tonight
necessarili
martin
peter
leg
harvest
frist
bright
anim
trade
taught
dayton
obvious
40
industri
chines
admit
infant
4
2
0
−2
−4
1
100
10000
Frequency of Word within Topic

Limitations of Frequency Statistics?

- Typically focus on unigrams (single terms)
- Often favors frequent (TF) or rare (IDF) terms
  - Not clear that these provide best description
- A "bag of words" ignores additional information
  - Grammar / part-of-speech
  - Position within document
  - Recognizable entities

Yelp: Review Spotlight



Yelp: Review Spotlight