15. Text Data Visualization

Prof. Tulasi Prasad Sariki
SCSE, VIT, Chennai

www.learnersdesk.weebly.com

Why Visualize Text?

Understanding – get the “gist” of a document

Grouping – cluster for overview or classification

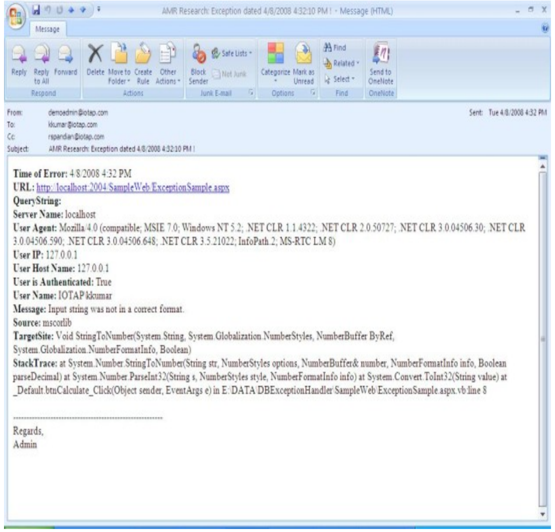
Compare – compare document collections, or inspect evolution of collection over time

Correlate – compare patterns in text to those in other data, e.g., correlate with social network

What is Text Data

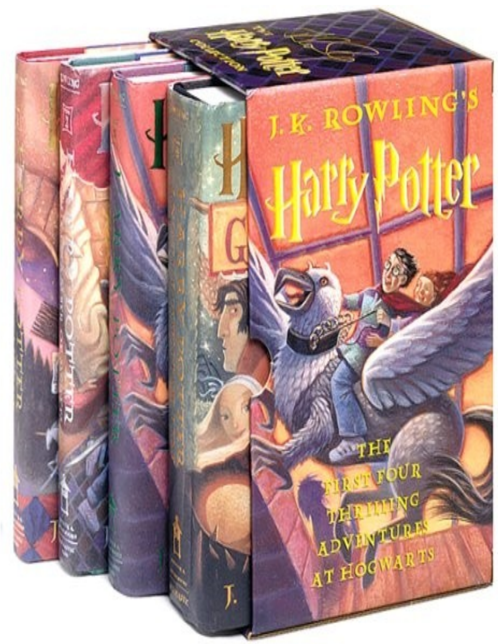
- Documents
  - Articles, books and novels
  - E-mails, web pages, blogs
  - Tags, comments
  - Computer programs, logs
- Collection of documents
  - Messages (e-mail, blogs, tags, comments)
  - Social networks (personal profiles)
  - Academic collaborations (publications)

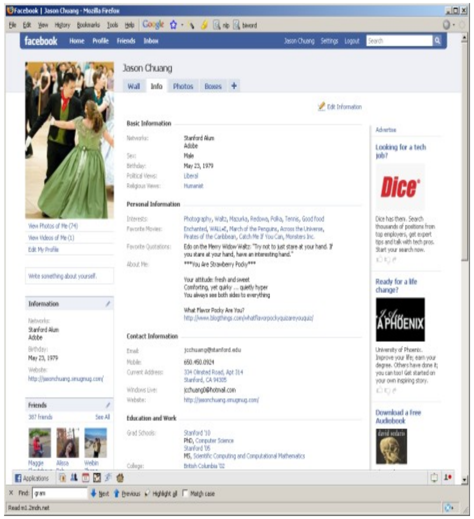
Where Text Data?



Time of Error: 4/8/2008 4:32 PM
URL: http://localhost:2004/SampleWeb/ExceptionSample.aspx
QueryString:
Server Name: localhost
User Agent: Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.2; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.04506.30; .NET CLR 3.0.04506.590; .NET CLR 3.0.04506.648; .NET CLR 3.5.21022; InfoPath.2; MS-RTC LM 8)
User IP: 127.0.0.1
User Host Name: 127.0.0.1
User is Authenticated: True
User Name: IOTAP\kkumar
Message: Input string was not in a correct format.
Source: mscorlib
TargetSite: Void StringToNumber(System.String, System.Globalization.NumberStyles, NumberBuffer ByRef, System.Globalization.NumberFormatInfo, Boolean)
StackTrace: at System.Number.StringToNumber(String str, NumberStyles options, NumberBuffer& number, NumberFormatInfo info, Boolean parseDecimal) at System.Number.ParseInt32(String s, NumberStyles style, NumberFormatInfo info) at System.Convert.ToInt32(String value) at _Default.btnCalculate_Click(Object sender, EventArgs e) in E:\DATA\DBExceptionHandler\SampleWeb\ExceptionSample.aspx.vb:line 8

Regards,
Admin





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 you about an issue that is central to that future – and that is the issue of health care.

I am not the first President to take up this cause, but I am determined to be the last. It has now been nearly a

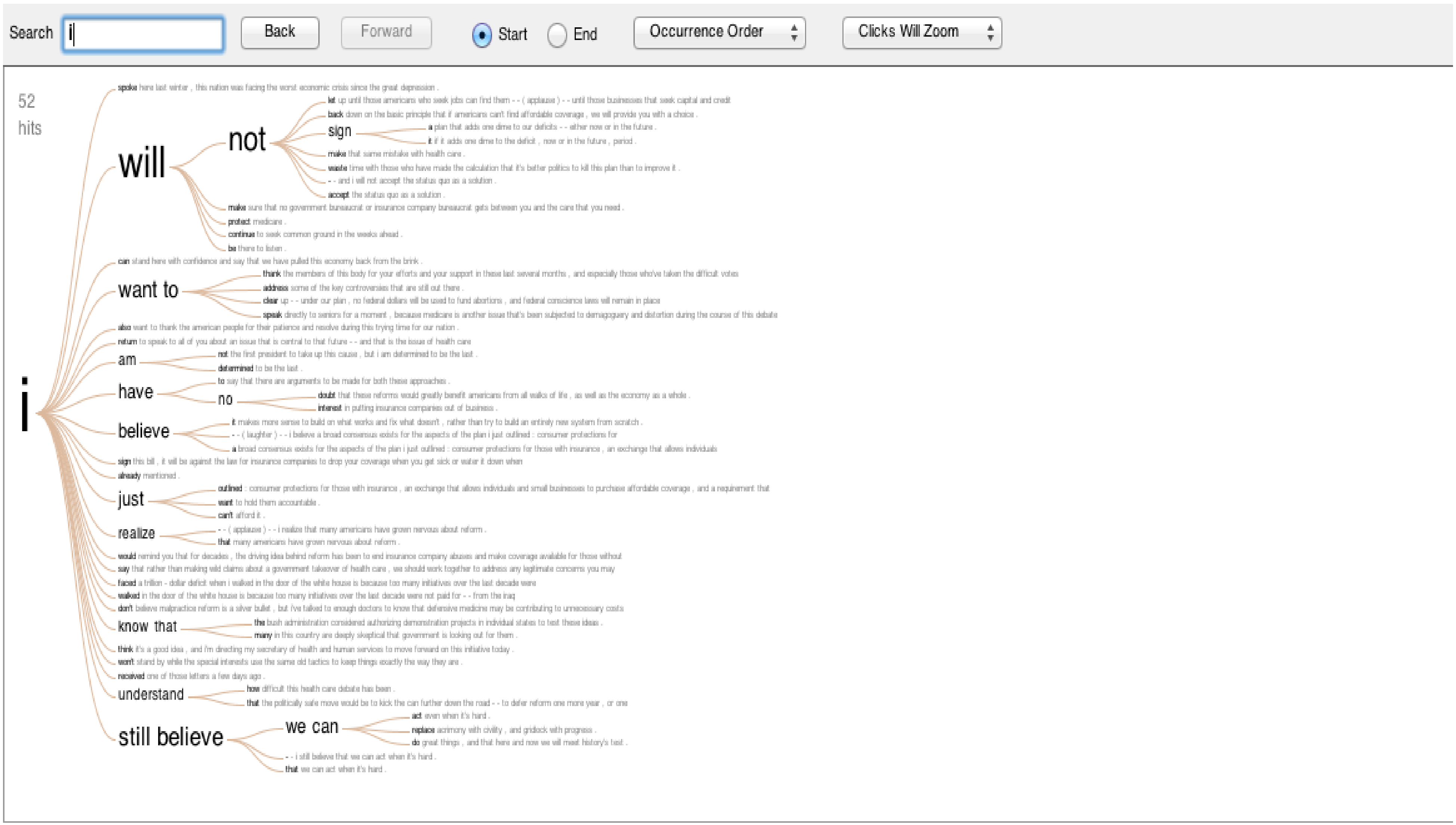
Tag Clouds: Word Count

Tag Clouds: Word Count

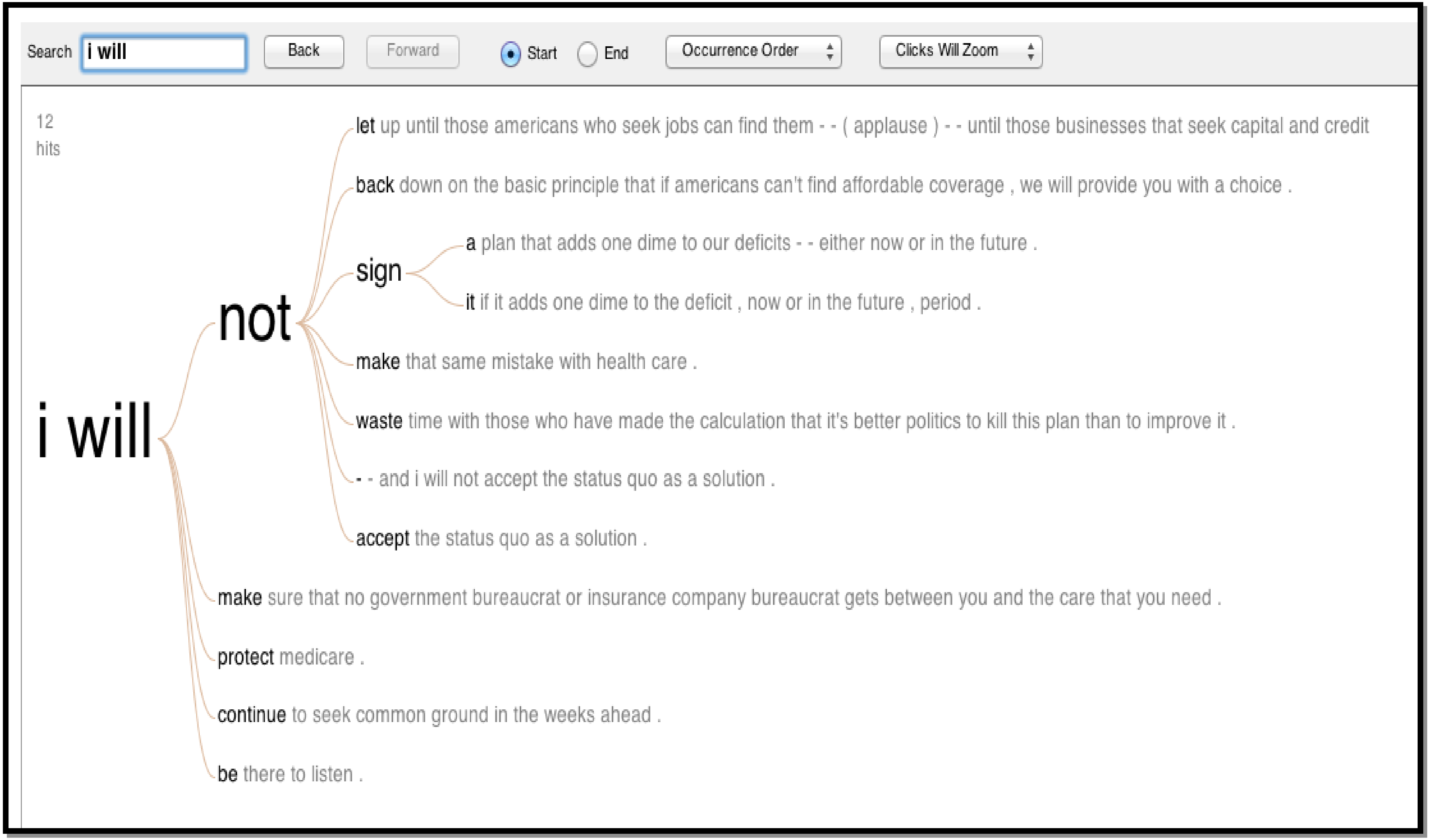
Bill Clinton 1993

Barack Obama 2009

WordTree: Word Sequences



WordTree: Word Sequences



Challenges of Text Visualization

High Dimensionality

Where possible use **text to represent text**…
… which terms are the most descriptive?

Context & Semantics

Provide **relevant context** to aid understanding.

Show (or provide access to) the **source text**.

Modeling Abstraction

Determine your **analysis task**.

Understand abstraction of your **language models**.

Match analysis task with appropriate tools and models.

Topics

- Text as Data
- Visualizing Document Content
- Evolving Documents
- Visualizing Conversation
- Document Collections

Text as Data

Words are (not) nominal?

High dimensional (10,000+)

More than equality tests

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.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#gocard, @stanfordfball, Beat Cal!!!!!!!!*

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

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
  - For example, simple term counts

Aggregate into a document-term matrix

- Document vector space model

Document-Term Matrix

- Each document is a vector of term weights
- Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

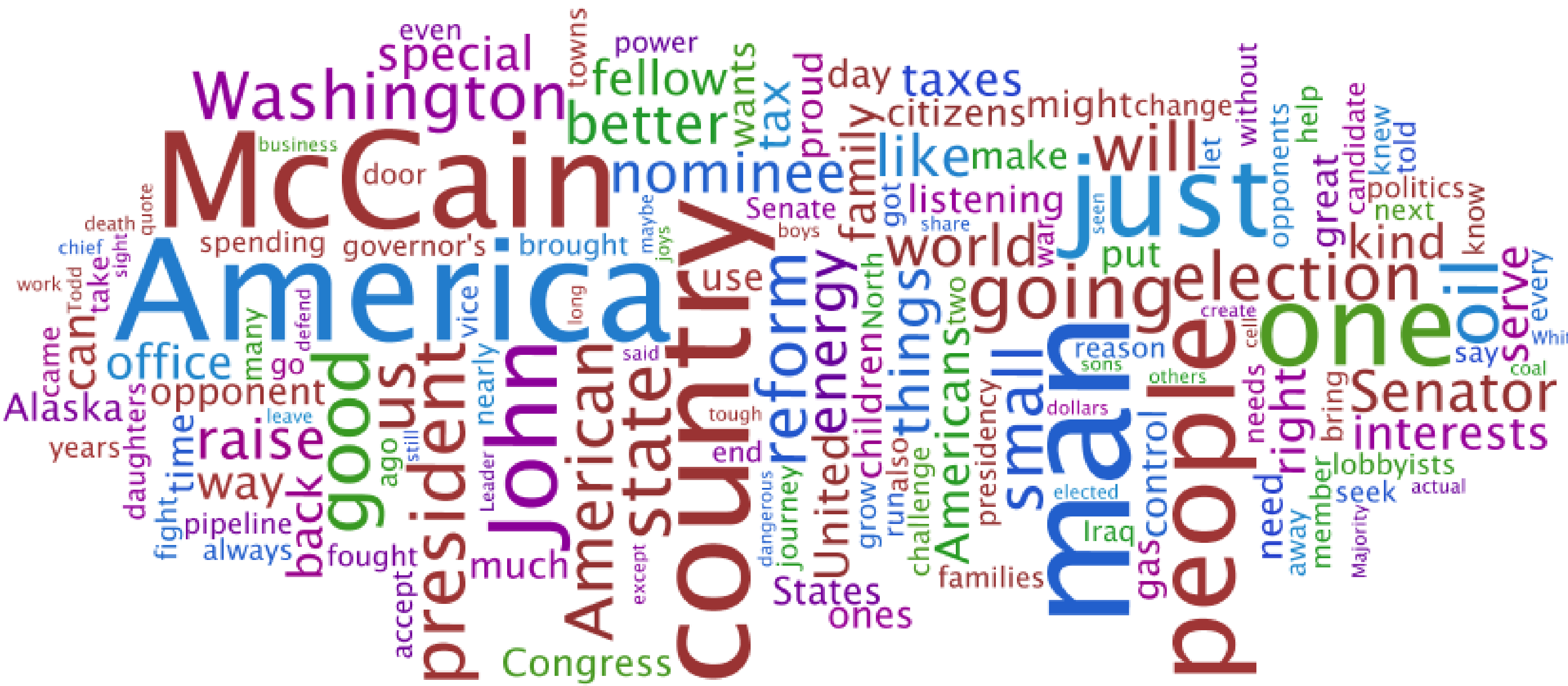
Tag Clouds

- Strength
  - Can help with 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

Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous
Tags:

Edit Language Font Layout Color



Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated

Keyword Weighting

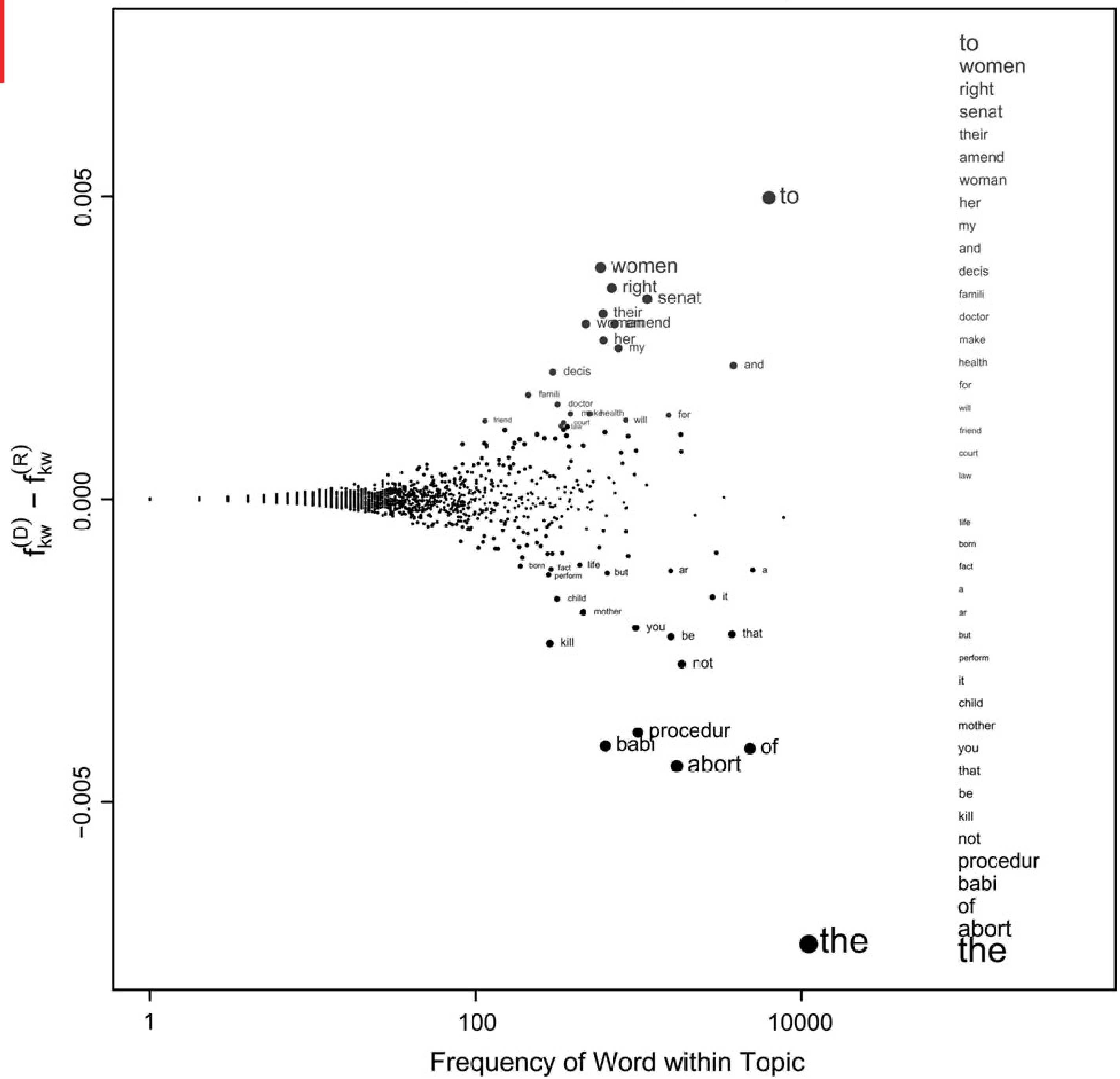
Term Frequency

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)



Keyword Weighting

Term Frequency

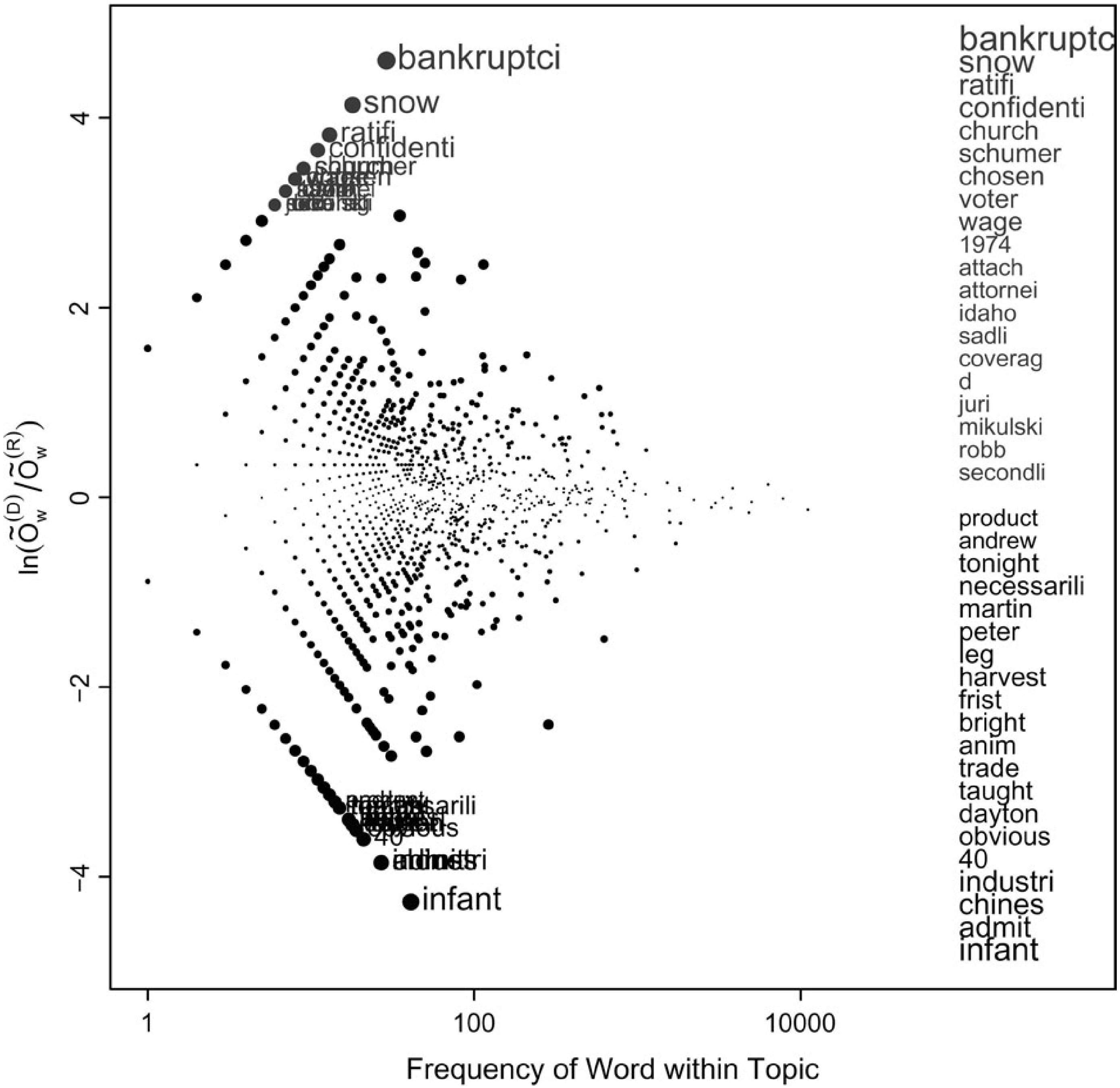
$tf_{td} = \text{count}(t) \text{ in } d$

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

Partisan Words, 106th Congress, Abortion
(Log−Odds−Ratio, Smoothed Log−Odds−Ratio)



Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

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

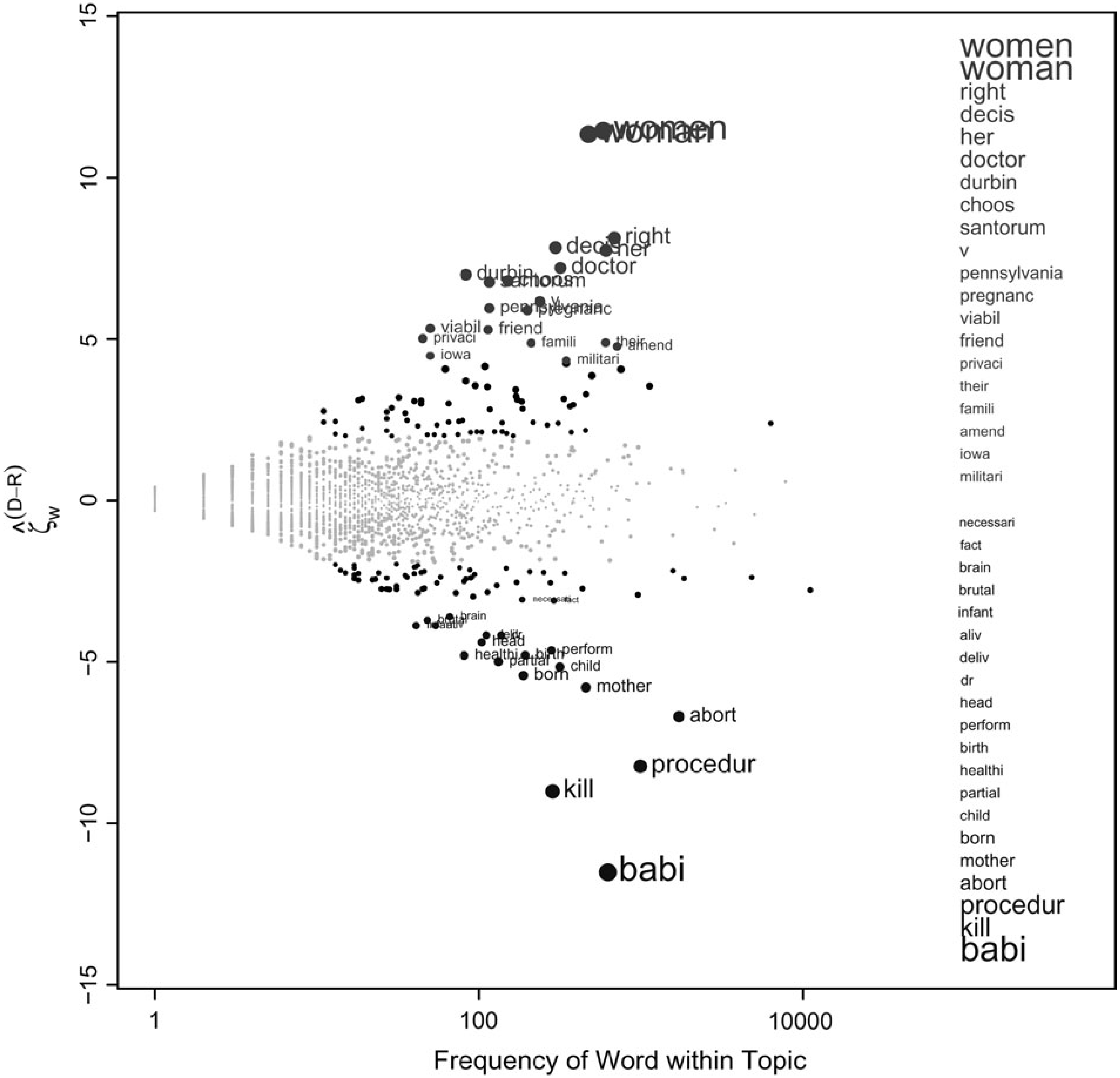
G^2: Probability of different word frequency

$E_1 = |d| \times (tf_{td} + tf_{t(C\text{-}d)}) / |C|$

$E_2 = |C\text{-}d| \times (tf_{td} + tf_{t(C\text{-}d)}) / |C|$

$G^2 = 2 \times (tf_{td} \log(tf_{td}/E_1) + tf_{t(C\text{-}d)} \log(tf_{t(C\text{-}d)}/E_2))$

Partisan Words, 106th Congress, Abortion
(Weighted Log−Odds−Ratio, Informative Dirichlet Prior)



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

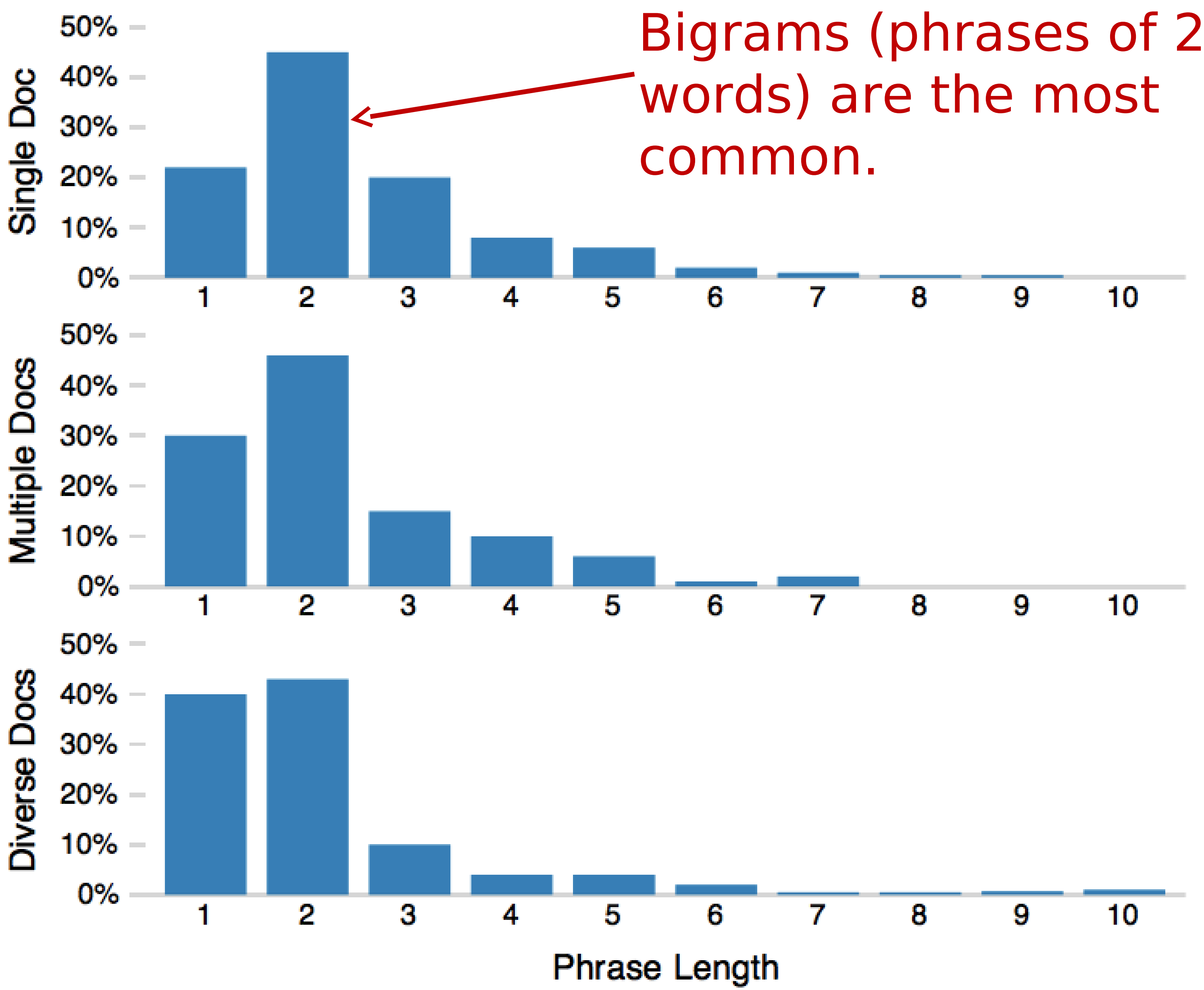
How do people describe text?

We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

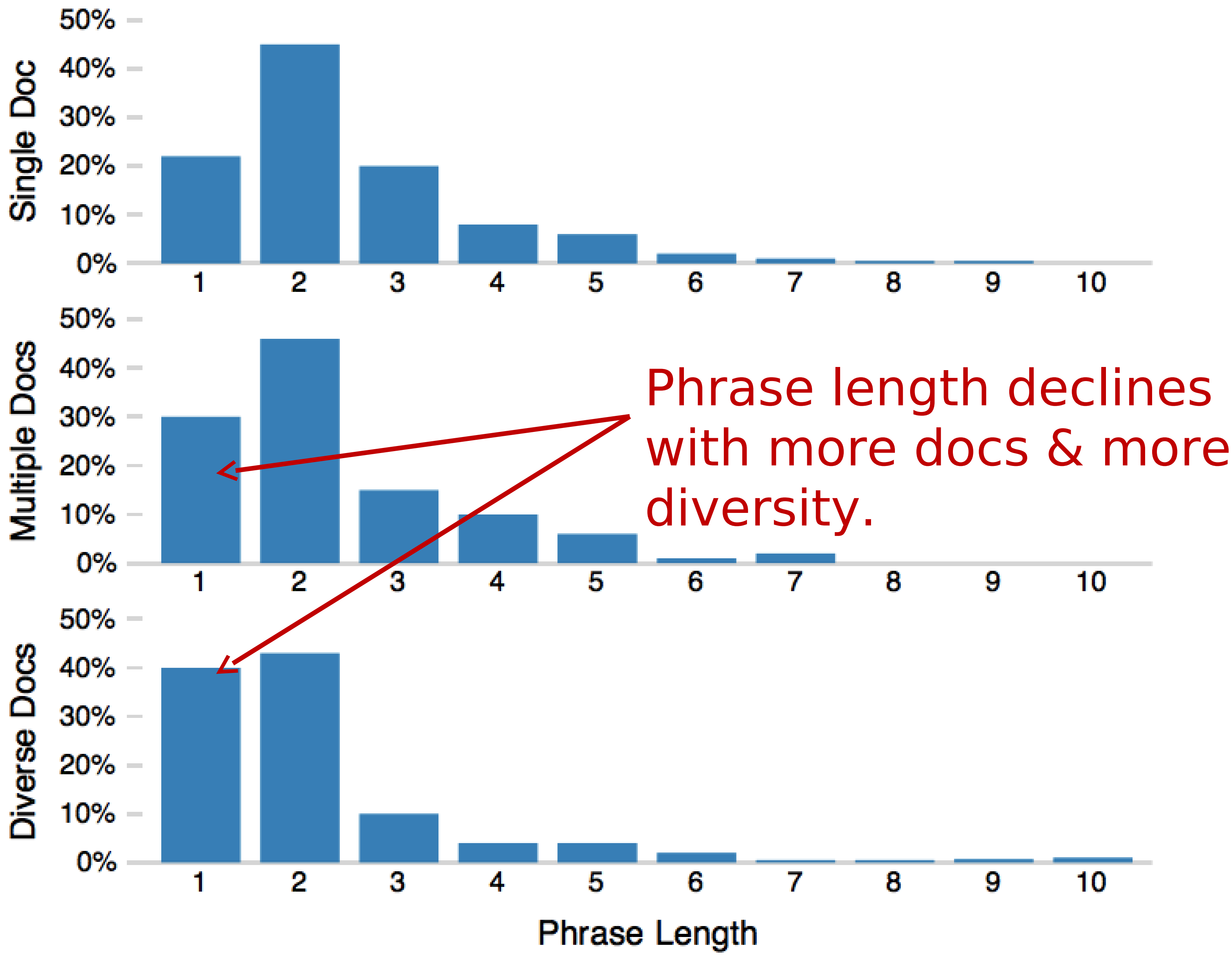
Students were given 3 documents in sequence; they then described the collection as a whole.

Students were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically.

[Chuang, Heer & Manning, 2010]



Bigrams (phrases of 2 words) are the most common.
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Phrase Length



Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Phrase Length
Phrase length declines with more docs & more diversity.

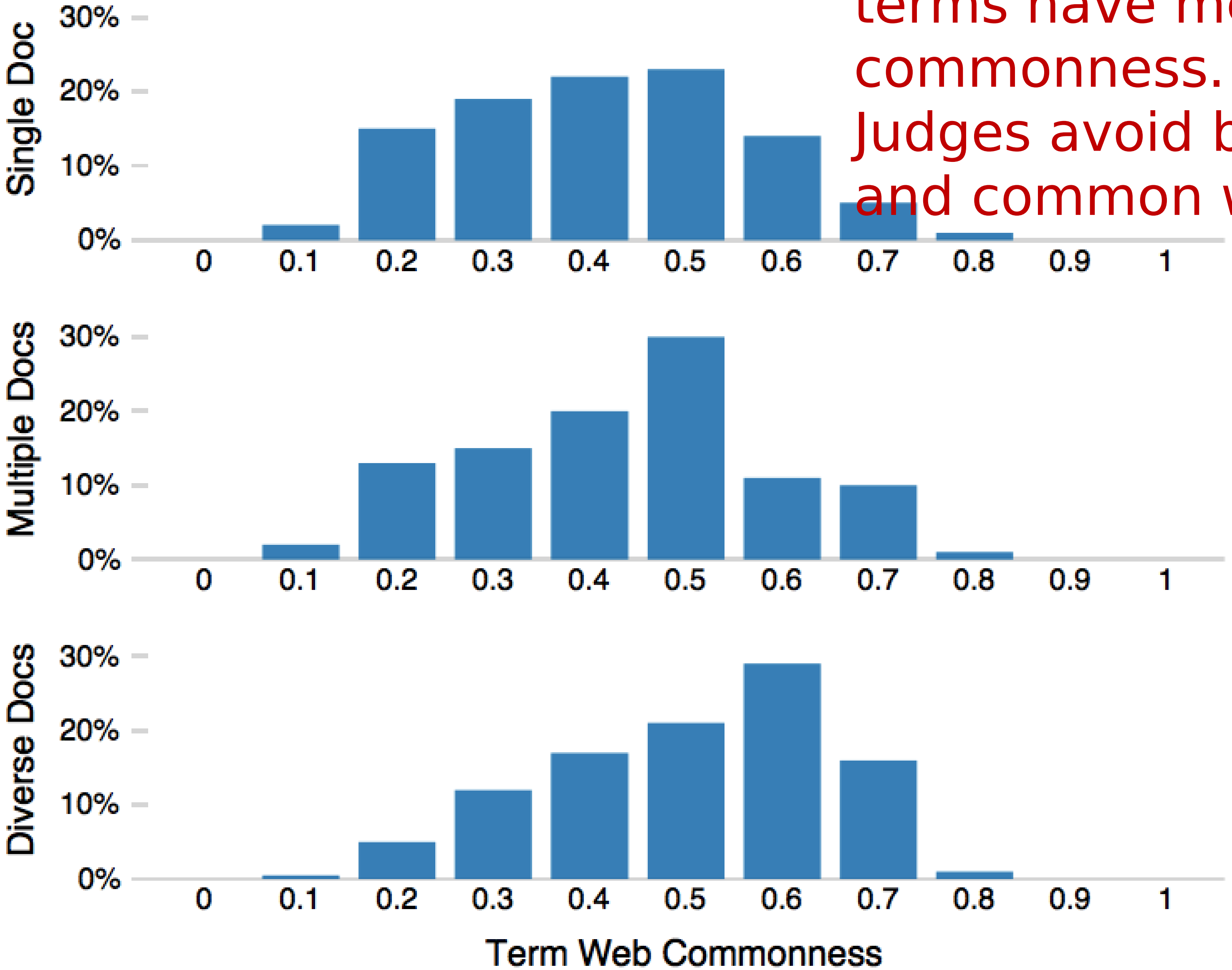
Term Commonness

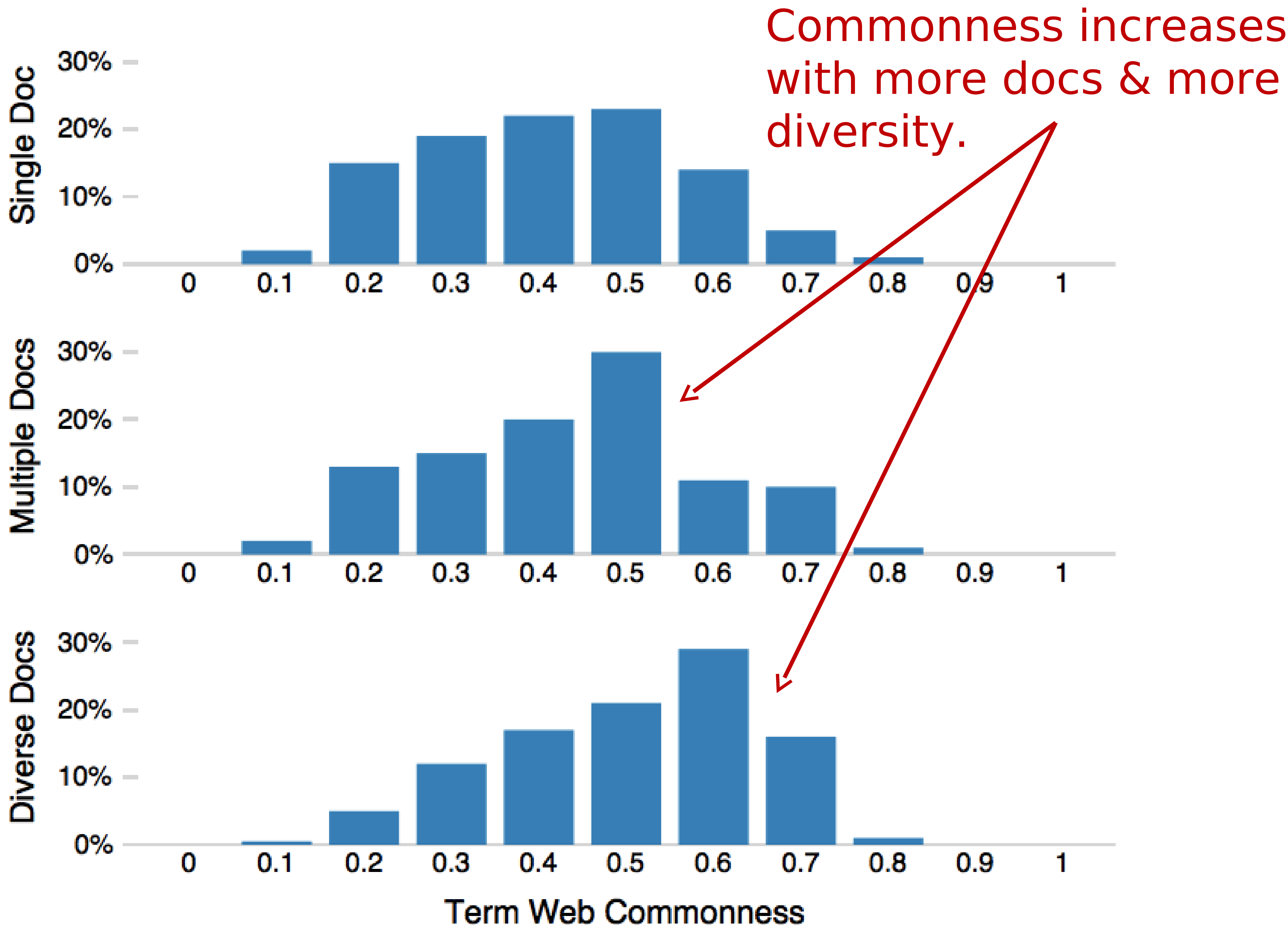
$\log(tf_w)$ / $\log(tf_{the})$

The normalized term frequency relative to the most frequent n-gram, e.g., the word “the”.

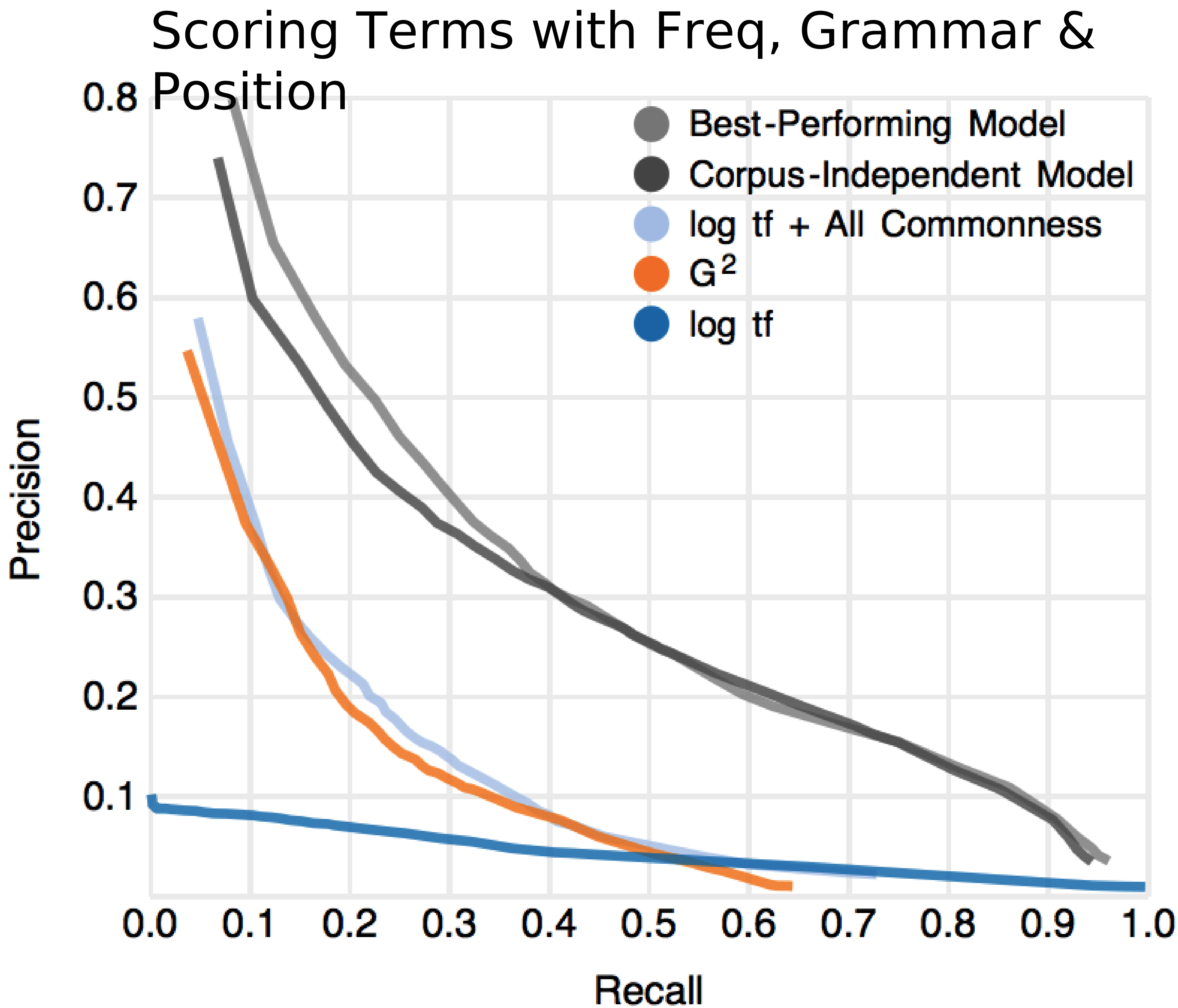
Measured across an entire corpus or across the entire English language (using Google n-grams)

Selected descriptive terms have medium commonness.
Judges avoid both rare and common words.





Commonness increases with more docs & more diversity.
Single Doc
Multiple Docs
Diverse Docs
30%
20%
10%
0%
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Term Web Commonness



Scoring Terms with Freq, Grammar & Position
Best-Performing Model
Corpus-Independent Model
log tf + All Commonness
G2
log tf
Precision
Recall
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

A fighter jet rain check

Story and video by Chamila Jayaweera

CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

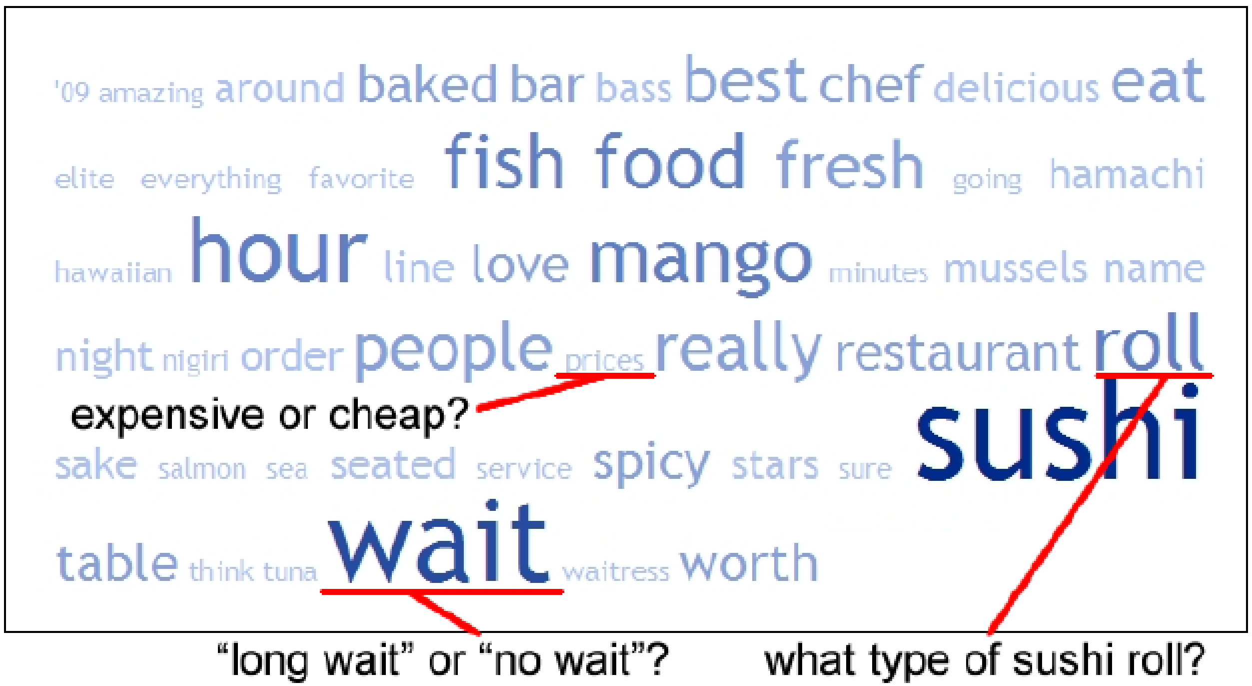
Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out moreabout how the process works and watch the action unfold, click above to see the video story.

G^2	Regression Model
fighter	Super Hornet
F/A	F/A -18
Hornet	fighter jet
Super	Boeing engineers
Boeing	special process
-18	rain check
rain	electronics suite
St.	Program experts
jet	simulated rain
Louis	ultimate customers
15-inch-per-hour	enormous hangar
douse	water-tight integrity
hangar	Rich Baxter
water-tight	15-inch-per-hour rate
Check	video story
Baxter	aircraft
sea-based	U.S. Navy fighter pilots
aircraft	Super Hornet final assembly manager
Rich	U.S.
seep	Navy fighter
click	fighter pilots
Navy	sea-based fighter
sure	
Water	
moisture	
watch	
enormous	
stay	
want	

Yelp: Review Spotlight [Yatani 2011]



Yelp: Review Spotlight [Yatani 2011]

Tips: Descriptive Keyphrases

Understand the limitations of your language model.

Bag of words

Easy to compute

Single words

Loss of word ordering

Select appropriate model and visualization

Generate longer, more meaningful phrases

Adjective-noun word pairs for reviews

Show keyphrases within source text

Thank You