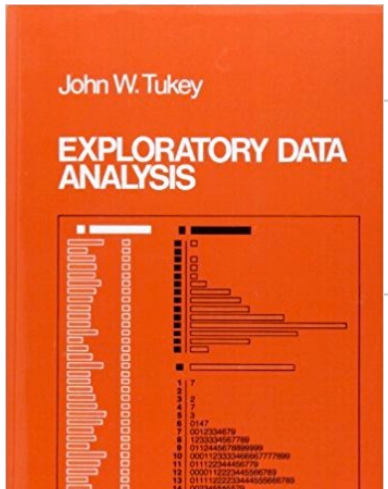
Word Association Norms, Mutual Information and Lexicography

Plan

- Summarize main points of paper
- Call out
  - some highlights of subsequent literature
  - suggestions for future work

A Big Tent

- My more successful papers don't tend to invent new ideas,
  - but rather borrow ideas from established fields and
  - connect the dots to novel applications in other fields
- I love to work with smart colleagues that know neat stuff
  - that's fun to hear about
- Interdisciplinary team
  - Gale (Statistics)
  - Hanks (Lexicography)
  - Hindle (Linguistics)



Title	Add	More	1–20	Cited by	Year
Word association norms, mutual information, and lexicography KW Church, P Hanks Computational linguistics 16 (1), 22-29				4075	1990
A stochastic parts program and noun phrase parser for unrestricted text KW Church Proceedings of the second conference on Applied natural language processing ...				1555	1988
A program for aligning sentences in bilingual corpora WA Gale, KW Church Computational linguistics 19 (1), 75-102				1454	1993
A method for disambiguating word senses in a large corpus WA Gale, KW Church, D Yarowsky Computers and the Humanities 26 (5), 415-439				761	1992
Using statistics in lexical analysis K Church, W Gale, P Hanks, D Hindle Lexical acquisition: exploiting on-line resources to build a lexicon 115, 164				654	1991
One sense per discourse WA Gale, KW Church, D Yarowsky Proceedings of the workshop on Speech and Natural Language, 233-237				602	1992
Introduction to the special issue on computational linguistics using large corpora KW Church, RL Mercer Computational linguistics 19 (1), 1-24				496	1993

Title | Add | More | 1–20 | Cited by

Word association norms, mutual information, and lexicography — 4075
KW Church, P Hanks
Computational linguistics 16 (1), 22-29

1555

1454

761

654

602

496

9/15/17

Table 3. Some interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Uninteresting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

There is no data like more data

Table 3. Some interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Uninteresting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

The Quote

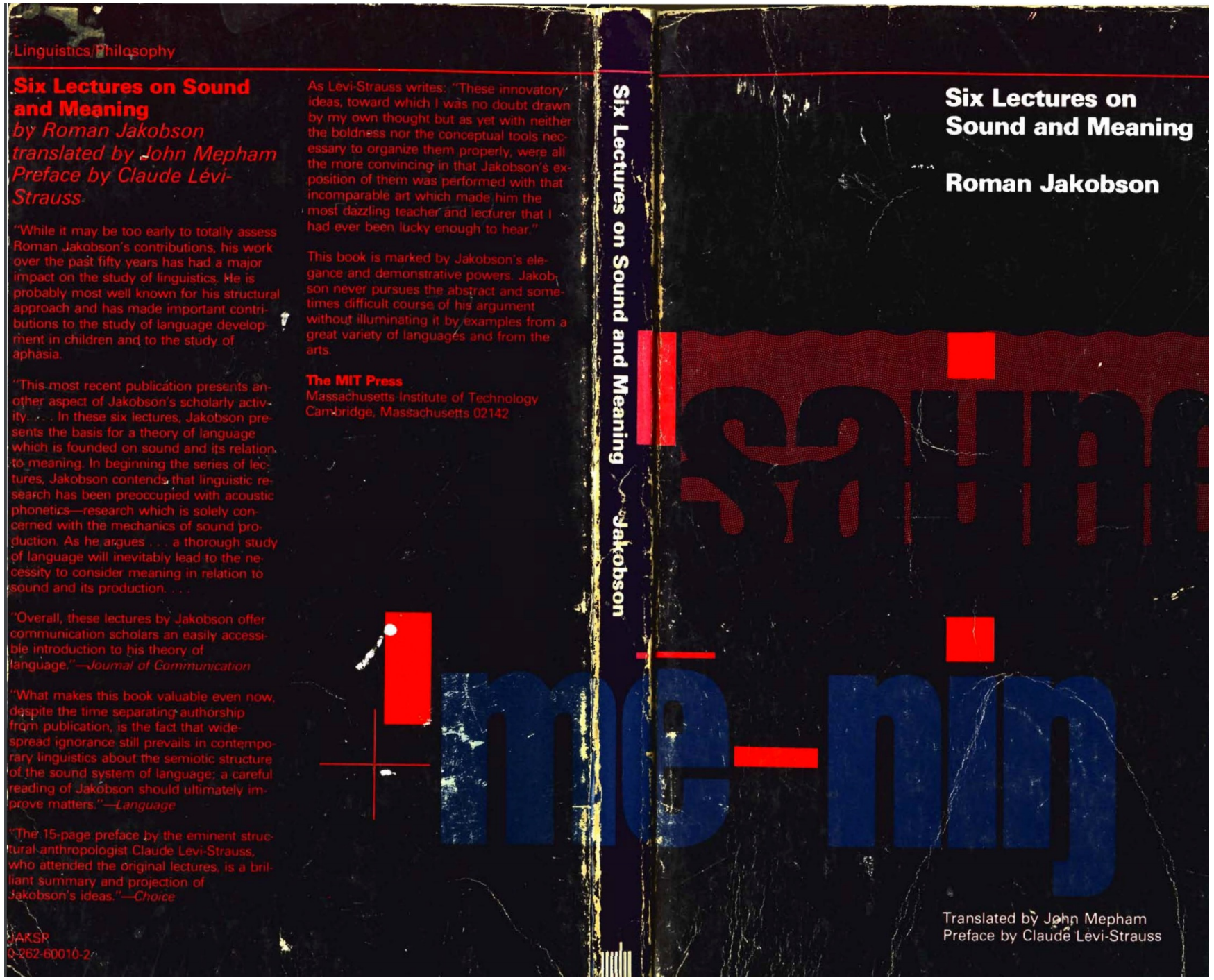
"Whenever I fire a linguist our system performance improves"

From my talk entitled:

Applying Information Theoretic Methods: Evaluation of Grammar Quality

Workshop on Evaluation of NLP Systems, Wayne PA, December 1988

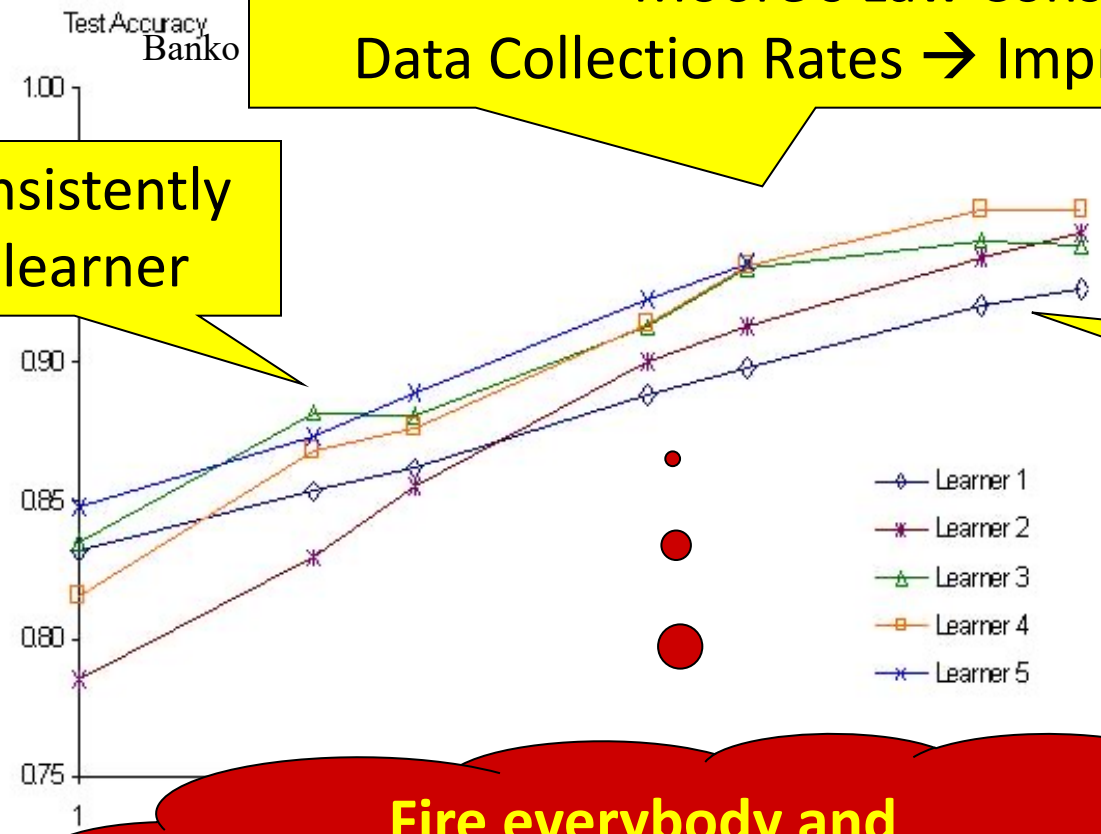
http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf


Linguistics/Philosophy
Six Lectures on Sound and Meaning
by Roman Jakobson
translated by John Mepham
Preface by Claude Lévi-Strauss
"While it may be too early to totally assess Roman Jakobson's contributions, his work over the past fifty years has had a major impact on the study of linguistics. He is probably most well known for his structural approach and has made important contributions to the study of language development in children and to the study of aphasia.
"This most recent publication presents another aspect of Jakobson's scholarly activity. . . . In these six lectures, Jakobson presents the basis for a theory of language which is founded on sound and its relation to meaning. In beginning the series of lectures, Jakobson contends that linguistic research has been preoccupied with acoustic phonetics—research which is solely concerned with the mechanics of sound production. As he argues . . . a thorough study of language will inevitably lead to the necessity to consider meaning in relation to sound and its production. . . .
"Overall, these lectures by Jakobson offer communication scholars an easily accessible introduction to his theory of language."—Journal of Communication
"What makes this book valuable even now, despite the time separating authorship from publication, is the fact that widespread ignorance still prevails in contemporary linguistics about the semiotic structure of the sound system of language; a careful reading of Jakobson should ultimately improve matters."—Language
"The 15-page preface by the eminent structural anthropologist Claude Levi-Strauss, who attended the original lectures, is a brilliant summary and projection of Jakobson's ideas."—Choice
JAKSP
0-262-60010-2
As Levi-Strauss writes: "These innovatory ideas, toward which I was no doubt drawn by my own thought but as yet with neither the boldness nor the conceptual tools necessary to organize them properly, were all the more convincing in that Jakobson's exposition of them was performed with that incomparable art which made him the most dazzling teacher and lecturer that I had ever been lucky enough to hear."
This book is marked by Jakobson's elegance and demonstrative powers. Jakobson never pursues the abstract and sometimes difficult course of his argument without illuminating it by examples from a great variety of languages and from the arts.
The MIT Press
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142
Six Lectures on Sound and Meaning
Jakobson
Six Lectures on Sound and Meaning
Roman Jakobson
Translated by John Mepham
Preface by Claude Lévi-Strauss

"It never pays to think until you've run out of data" – Eric Brill

Moore's Law Constant:
Data Collection Rates → Improvement Rates

No consistently best learner



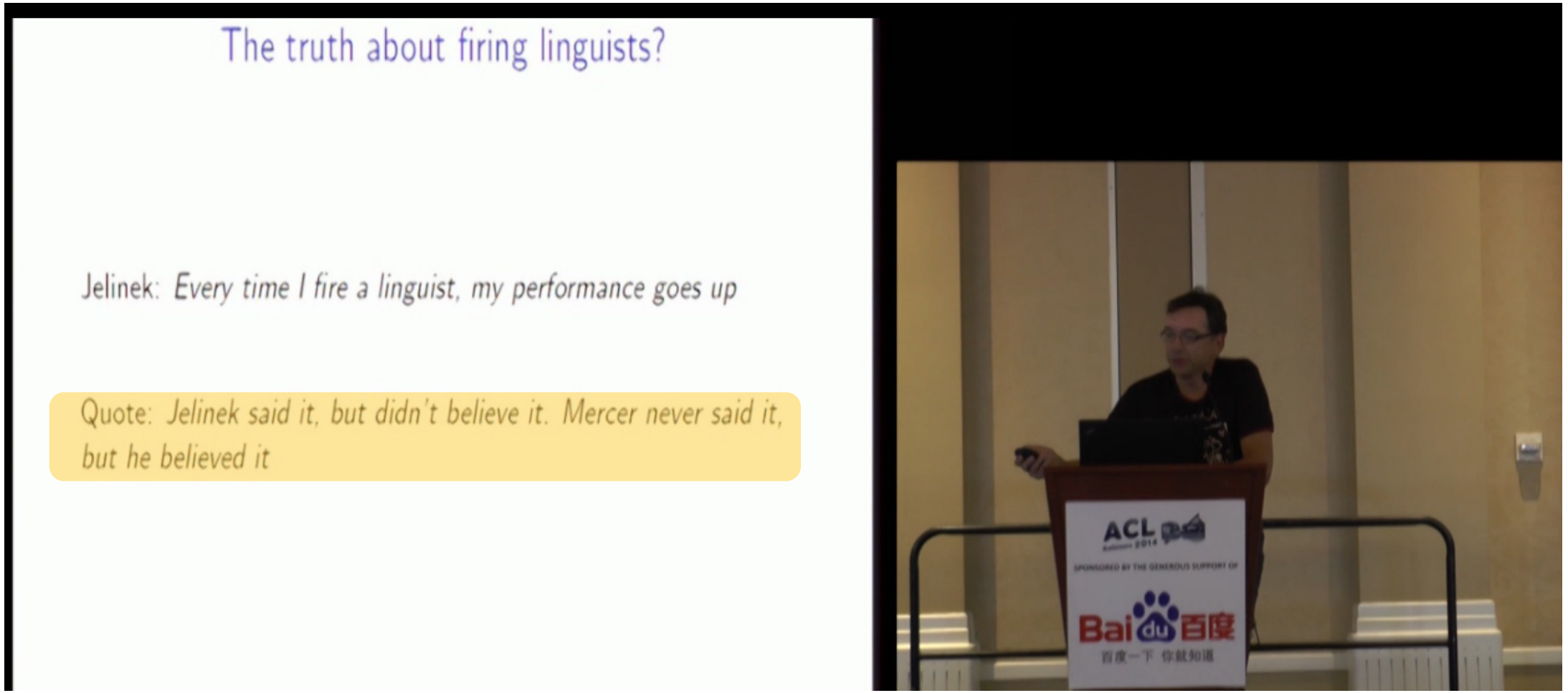
There is no data like more data

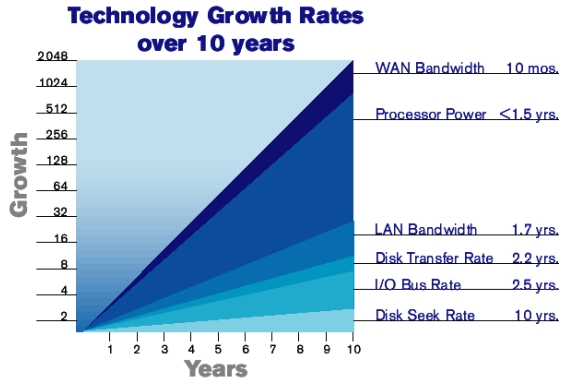
Fire everybody and spend the money on data

Quoted out of context

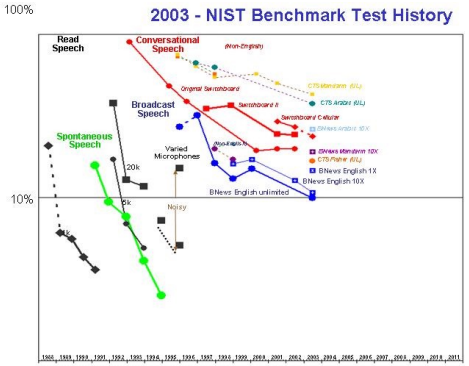
Robert Mercer
ACL Lifetime Achievement

http://techtalks.tv/talks/closing-session/60532/





The Disk Space Conjecture



- Improvements in Speech, Language (& more)
  - are indexed to improvements in disk capacities
  - because disk prices → size of web → training data
- 2003 Prediction:
  - Disks improve1000x per decade
  - 1TB: $1k (2003) → $1 (2013)
  - ***Missed by 30x (a TB is currently ~ $30 >> $1)***

Compare

Toshiba - Canvio 3TB External USB 3.0 Hard Drive - Black

Model: HDWC130XK3J1 **SKU:** 6914041

USB 3.0 interface, backward compatible with USB 2.0; NTI Backup Now EZ software; data transfer rates up to 5 Gbps; 32MB cache buffer

Customer Reviews: 4.2 of 5 (241 reviews)

Check Shipping & Availability ›

Sale: $104.99

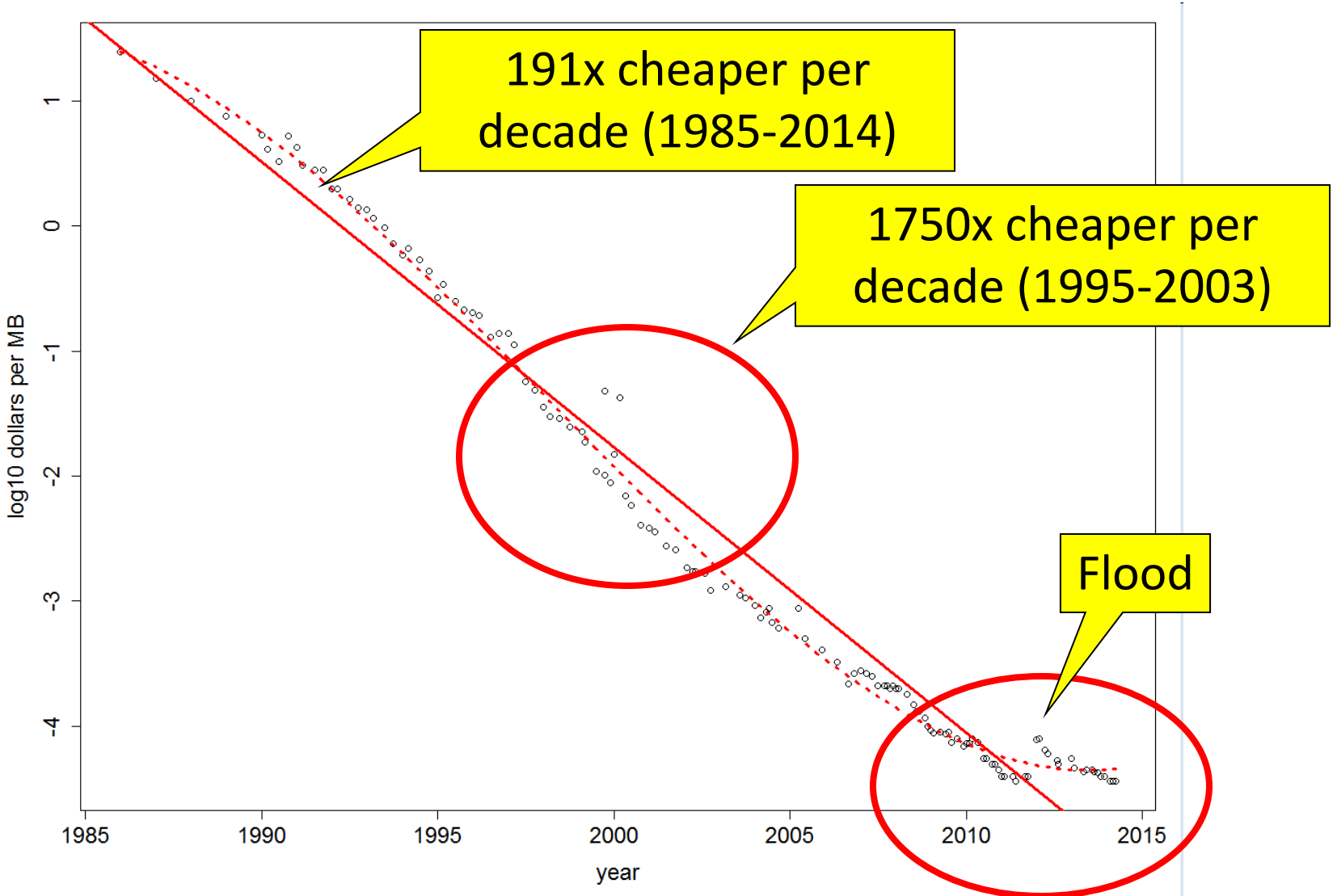
Reg. Price: $119.99
You Save: $15.00

- Free Shipping
- Get up to 6% Back in Rewards: See How

Add to Cart

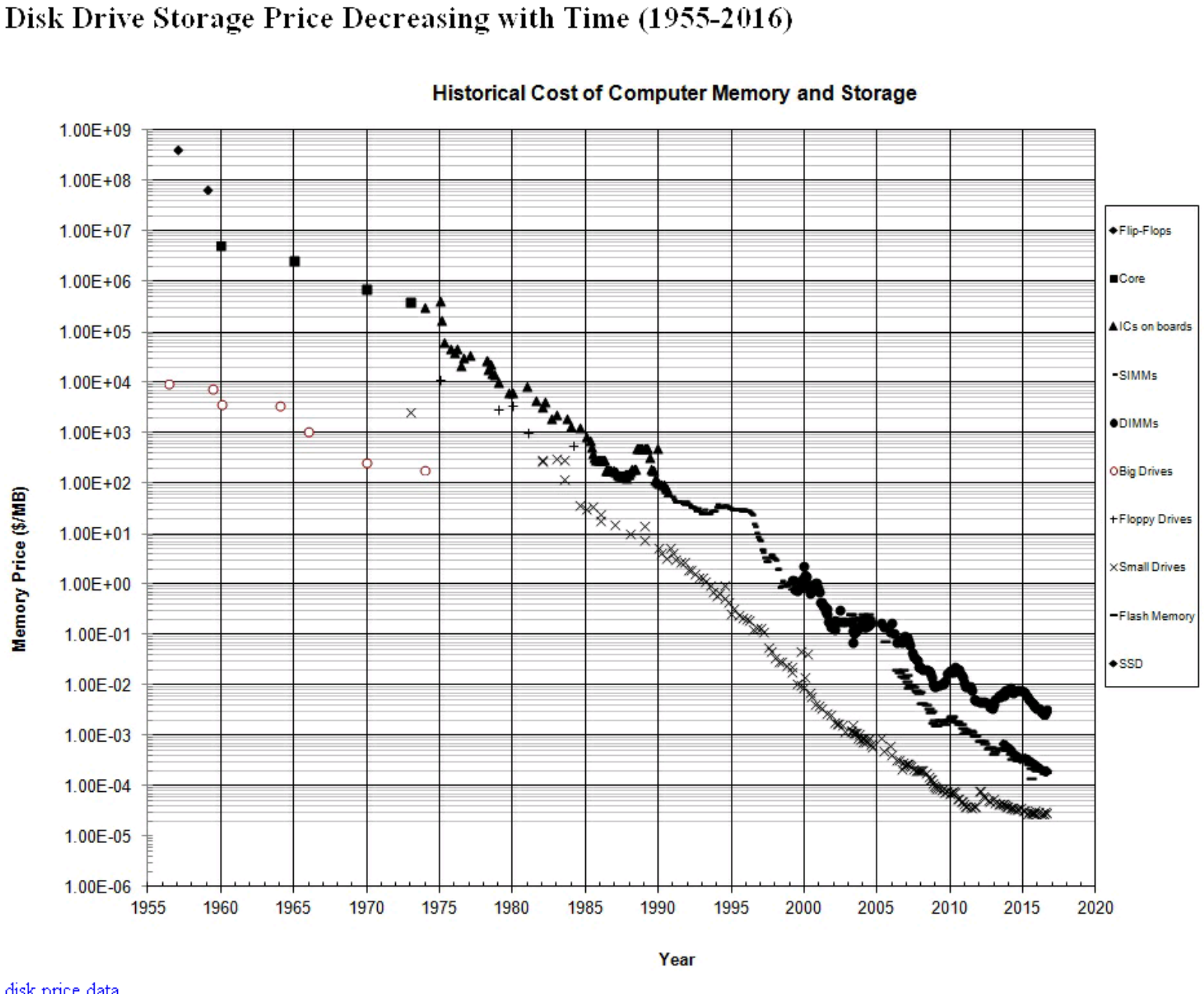
Disk Prices Over 30 Years

http://www.jcmit.com/diskprice.htm



Moore's Law: Different Time Constants for Different Things

https://archive.is/3pgVP



Word Association Norms: Doctor / Nurse

Table 3. Some interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Uninteresting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

Priming & Word Associations

Task: Subject is given two strings and responds “yes” if both are words

Journal of Experimental Psychology
1971, Vol. 90, No. 2, 227–234

FACILITATION IN RECOGNIZING PAIRS OF WORDS: EVIDENCE OF A DEPENDENCE BETWEEN RETRIEVAL OPERATIONS[1]

DAVID E. MEYER[2] AND ROGER W. SCHVANEVELDT
Bell Telephone Laboratories, Murray Hill, New Jersey — *University of Colorado*

EXPERIMENT I

Method

Subjects.—The *S*s were 12 high school students who served as paid volunteers.

Stimuli.—The following test stimuli were used: 48 pairs of associated words, e.g., BREAD-BUTTER and NURSE-DOCTOR, selected from the Connecticut Free Associational Norms (Bousfield, Cohen, & Whitmarsh, 1961); 48 pairs of unassociated words, e.g., BREAD-DOCTOR and NURSE-BUTTER, formed by randomly interchanging the response terms between the 48 pairs of associated words so that there were no obvious associations within the resulting pairs; 48 pairs of nonwords; and 96 pairs involving a word and a nonword. Within each pair of associated words, the second member was either the first or second most frequent free associate given in response to the first member. Within each pair of unassociated words, the second member was never the first or second most frequent free associate of the first member. The median length of strings in the pairs of associated words and pairs of unassociated words was 5 letters and ranged from 3 to 7 letters;

TABLE 1

MEAN REACTION TIMES (RTs) OF CORRECT RESPONSES AND MEAN PERCENT ERRORS IN THE YES–NO TASK

Type of stimulus pair		Correct response	Proportion of trials	Mean RT (msec.)	Mean % errors
Top string	Bottom string				
word	associated word	yes	.25	855	6.3
word	unassociated word	yes	.25	940	8.7
word	nonword	no	.167	1,087	27.6
nonword	word	no	.167	904	7.8
nonword	nonword	no	.167	884	2.6

Pointwise Mutual Information (PMI)

4 An Information Theoretic Measure

We propose an alternative measure, the *association ratio*, for measuring word association norms, based on the information theoretic concept of *mutual information*.[1] The proposed measure is more objective and less costly than the subjective method employed in Palermo and Jenkins (1964). The association ratio can be scaled up to provide robust estimates of word association norms for a large portion of the language. Using the association ratio measure, the five most associated words are, in order: *dentists, nurses, treating, treat,* and *hospitals*.

What is "mutual information?" According to Fano (1961), if two points (words), x and y, have probabilities $P(x)$ and $P(y)$, then their mutual information, $I(x,y)$, is defined to be

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Simple interpretation
  - PMI compares P(x,y) with chance
    - Chance = P(x) P(y)
  - If there is a genuine association
    - then P(x,y) >> P(x) P(y)
  - Uninteresting associations
    - P(x,y) ≈ P(x) P(y)
- Popular applications (lexicography)
  - more like hypothesis testing
    - focus on largest PMI
      - where we can reject null hypothesis
      - null hypo: uninteresting
  - less like language modeling for speech and machine translation

Frederick Jelinek

Born	Bedřich Jelínek November 18, 1932 Kladno, now Czech Republic
Died	September 14, 2010 (aged 77) Baltimore, United States
Citizenship	American
Fields	Information theory, natural language processing
Institutions	Cornell University, IBM Research, Johns Hopkins University
Alma mater	Massachusetts Institute of Technology
Doctoral advisor	Robert Fano
Notable students	Neil Sloane
Known for	Advancement of natural language processing techniques
Influences	Roman Jakobson
Notable awards	• James L. Flanagan Award (2005) • ACL Lifetime Achievement Award (2009)
Spouse	Milena Jelinek

Windows for computing P(x,y)

- Bigrams:
  - rectangular window with width of 1 word
- Ngrams
  - rectangular window with width of n-1 words
- More generally
  - Windows need not be rectangular
  - Or symmetric around 0
    - (Mutual Information is symmetric
    - but "Association Measure" is not)
- Convenient to assume windows sum to 1
- More interesting windows
  - Parse Trees / SVO

Table 5. What Can You Drink?

Verb	Object	Mutual Info	Joint Freq
drink/V	*martinis/O*	12.6	3
drink/V	*cup_water/O*	11.6	3
drink/V	*champagne/O*	10.9	3
drink/V	*beverage/O*	10.8	8
drink/V	*cup_coffee/O*	10.6	2
drink/V	*cognac/O*	10.6	2
drink/V	*beer/O*	9.9	29
drink/V	*cup/O*	9.7	6
drink/V	*coffee/O*	9.7	12
drink/V	*toast/O*	9.6	4
drink/V	*alcohol/O*	9.4	20
drink/V	*wine/O*	9.3	10
drink/V	*fluid/O*	9.0	5
drink/V	*liquor/O*	8.9	4
drink/V	*tea/O*	8.9	5
drink/V	*milk/O*	8.7	8
drink/V	*juice/O*	8.3	4
drink/V	*water/O*	7.2	43
drink/V	*quantity/O*	7.1	4

OCR Application

Consider the optical character recognizer (OCR) application. Suppose that we have an OCR device as in Kahan et al. (1987), and it has assigned about equal probability to having recognized *farm* and *form*, where the context is either: (1) *federal* __ *credit* or (2) *some* __ *of*.

- $federal \begin{pmatrix} farm \\ form \end{pmatrix} credit$
- $some \begin{pmatrix} farm \\ form \end{pmatrix} of$

The proposed association measure can make use of the fact that *farm* is much more likely in the first context and *form* is much more likely in the second to resolve the ambiguity. Note that alternative disambiguation methods based on syntactic constraints such as part of speech are unlikely to help in this case since both *form* and *farm* are commonly used as nouns.

Applications in Lexicography

rs Sunday, calling for greater economic reforms to	save China from poverty.
mmission asserted that " the Postal Service could	save enormous sums of money in contracting out individual c
Then, she said, the family hopes to	save enough for a down payment on a home.
e out-of-work steelworker, " because that doesn't	save jobs, that costs jobs. "
" We suspend reality when we say we'll	save money by spending $10,000 in wages for a public works
scientists has won the first round in an effort to	save one of Egypt's great treasures, the decaying tomb of R
about three children in a mining town who plot to	save the " pit ponies " doomed to be slaughtered.
GM executives say the shutdowns will	save the automaker $500 million a year in operating costs a
rtment as receiver, instructed officials to try to	save the company rather than liquidate it and then declared
The package, which is to	save the country nearly $2 billion, also includes a program
newly enhanced image as the moderate who moved to	save the country.
million offer from chairman Victor Posner to help	save the financially troubled company, but said Posner stil
after telling a delivery-room doctor not to try to	save the infant by inserting a tube in its throat to help i
h birthday Tuesday, cheered by those who fought to	save the majestic Beaux Arts architectural masterpiece.
at he had formed an alliance with Moslem rebels to	save the nation from communism.
" Basically we could	save the operating costs of the Pershings and ground-launch
We worked for a year to	save the site at enormous expense to us, " said Leveillee.
their expensive mirrors, just like in wartime, to	save them from drunken Yankee brawlers, " Tass said.
ard of many who risked their own lives in order to	save those who were passengers. "
We must increase the amount Americans	save. "

Figure 1 Short Sample of the Concordance to "save" from the AP 1987 Corpus.

The AP 1987 concordance to *save* is many pages long; there are 666 lines for the base form alone, and many more for the inflected forms *saved*, *saves*, *saving*, and *savings*. In the discussion that follows, we shall, for the sake of simplicity, not analyze the inflected forms and we shall only look at the patterns to the right of *save* (see Table 7).

It is hard to know what is important in such a concordance and what is not. For example, although it is easy to see from the concordance selection in Figure 1 that the word "to" often comes before "save" and the word "the" often comes after "save," it is hard to say from examination of a concordance alone whether either or both of these co-occurrences have any significance.

Two examples will illustrate how the association ratio measure helps make the analysis both quicker and more accurate.

Proper Place for Automation: Start with Drudgery

(Support our colleagues; don't talk too much about taking away jobs they love to do)

In point of fact, we actually developed these results in basically the reverse order. Concordance analysis is still extremely labor-intensive and prone to errors of omission. The ways that concordances are sorted don't adequately support current lexicographic practice. Despite the fact that a concordance is indexed by a single word, often lexicographers actually use a second word such as *from* or an equally common semantic concept such as a time adverbial to decide how to categorize concordance lines. In other words, they use two words to *triangulate in* on a word sense. This triangulation approach clusters concordance lines together into word senses based primarily on usage (distribu-

Computational Linguistics Volume 16, Number 1, March 1990

Some of my Best Friends are Linguists

(LREC 2004)

Frederick Jelinek
Johns Hopkins University

The Quote

"*Whenever I fire a linguist our system performance improves*"

From my talk entitled:

Applying Information Theoretic Methods: Evaluation of Grammar Quality

Workshop on Evaluation of NLP Systems, Wayne PA, December 1988

Patrick found tables like this very exciting

Table 7. Words Often Co-Occurring to the Right of "Save"

I(x, y)	f(x, y)	f(x)	x	f(y)	y
9.5	6	724	*save*	170	*forests*
9.4	6	724	*save*	180	*$1.2*
8.8	37	724	*save*	1697	*lives*
8.7	6	724	*save*	301	*enormous*
8.3	7	724	*save*	447	*annually*
7.7	20	724	*save*	2001	*jobs*
7.6	64	724	*save*	6776	*money*
7.2	36	724	*save*	4875	*life*
6.6	8	724	*save*	1668	*dollars*
6.4	7	724	*save*	1719	*costs*
6.4	6	724	*save*	1481	*thousands*
6.2	9	724	*save*	2590	*face*
5.7	6	724	*save*	2311	*son*
5.7	6	724	*save*	2387	*estimated*
5.5	7	724	*save*	3141	*your*
5.5	24	724	*save*	10880	*billion*
5.3	39	724	*save*	20846	*million*
5.2	8	724	*save*	4398	*us*
5.1	6	724	*save*	3513	*less*
5.0	7	724	*save*	4590	*own*
4.6	7	724	*save*	5798	*world*
4.6	7	724	*save*	6028	*my*
4.6	15	724	*save*	13010	*them*
4.5	8	724	*save*	7434	*country*
4.4	15	724	*save*	14296	*time*
4.4	64	724	*save*	61262	*from*
4.3	23	724	*save*	23258	*more*
4.2	25	724	*save*	27367	*their*
4.1	8	724	*save*	9249	*company*
4.1	6	724	*save*	7114	*month*

Save <good> from <bad>

save X from Y (65 concordance lines)

1 save PERSON from Y (23 concordance lines)

1.1 save PERSON from BAD (19 concordance lines)

(Robert DeNiro) to	save Indian tribes[PERSON] from genocide[DESTRUCT[BAD]] at the hands of
" We wanted to	save him[PERSON] from undue trouble[BAD] and loss[BAD] of money , "
Murphy was sacrificed to	save more powerful Democrats[PERSON] from harm[BAD] .
" God sent this man to	save my five children[PERSON] from being burned to death[DESTRUCT[BAD]] and
Pope John Paul II to "	save us[PERSON] from sin[BAD] . "

1.2 save PERSON from (BAD) LOC(ATION) (4 concordance lines)

rescuers who helped	save the toddler[PERSON] from an abandoned well[LOC] will be feted with a parade
while attempting to	save two drowning boys[PERSON] from a turbulent[BAD] creek[LOC] in Ohio[LOC]

2. save INST(ITUTION) from (ECON) BAD (27 concordance lines)

member states to help	save the EEC[INST] from possible bankruptcy[ECON][BAD] this year .
should be sought " to	save the company[CORP[INST]] from bankruptcy[ECON][BAD] .
law was necessary to	save the country[NATION[INST]] from disaster[BAD] .
operation " to	save the nation[NATION[INST]] from Communism[BAD][POLITICAL] .
were not needed to	save the system from bankruptcy[ECON][BAD] .
his efforts to	save the world[INST] from the likes of Lothar and the Spider Woman

3. save ANIMAL from DESTRUCT(ION) (5 concordance lines)

give them the money to	save the dogs[ANIMAL] from being destroyed[DESTRUCT] ,
program intended to	save the giant birds[ANIMAL] from extinction[DESTRUCT] ,

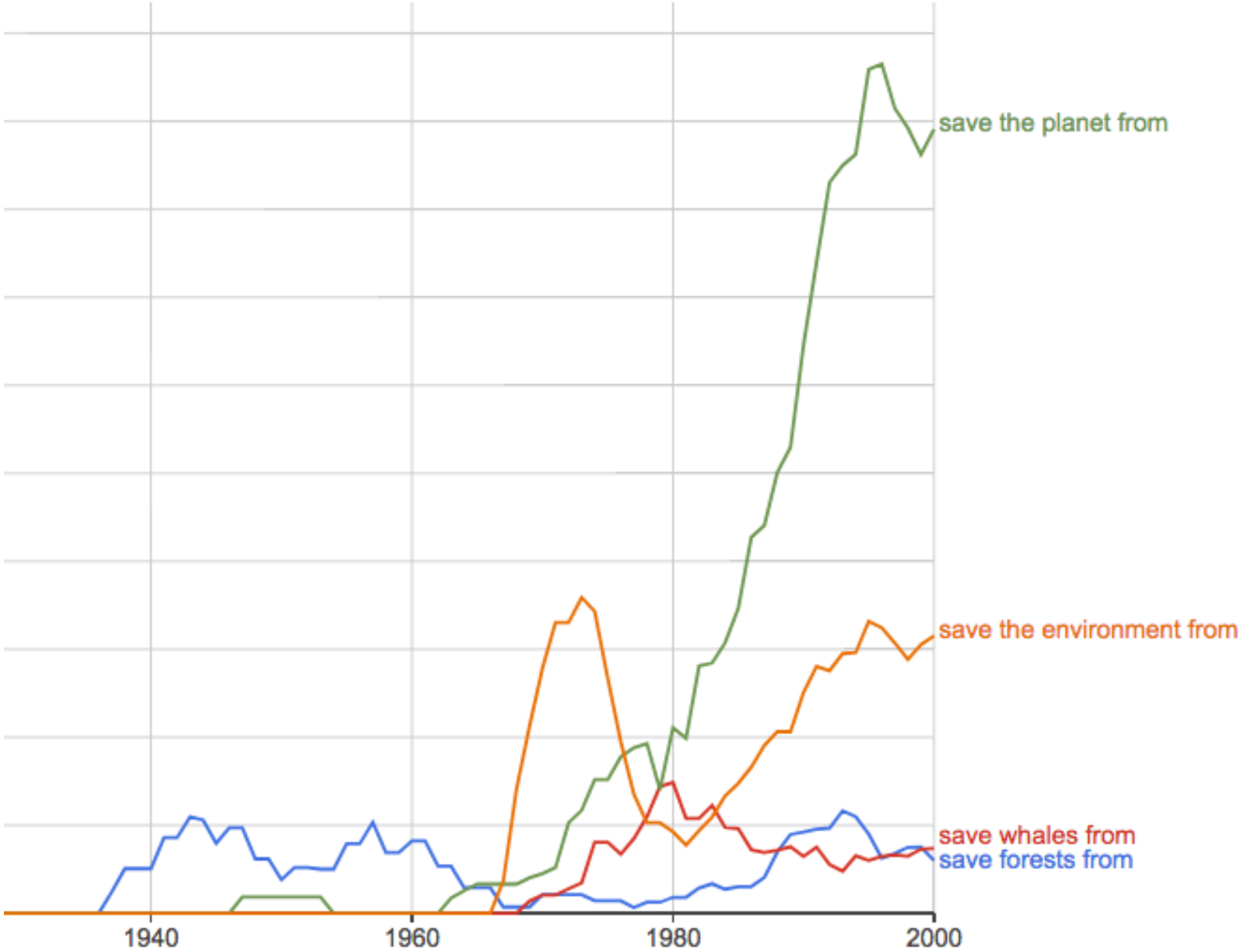
Save good shoppers from their evil $$

UNCLASSIFIED (10 concordance lines)

walnut and ash trees to	save them	from the axes and saws of a logging company .
after the attack to	save the ship	from a terrible[BAD] fire , Navy reports concluded Thursday .
certificates that would	save shoppers[PERSON]	anywhere from $50[MONEY] [NUMBER] to $500[MONEY

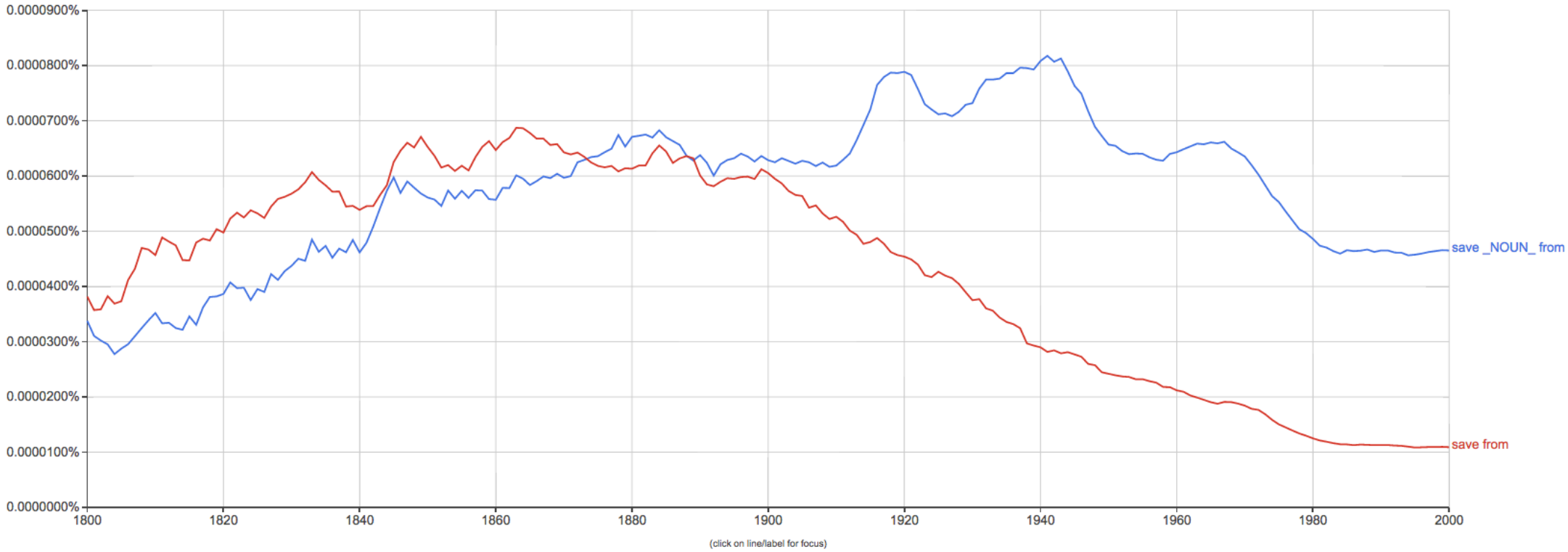
Patrick wanted me to “fix” my bug

Google Ngrams

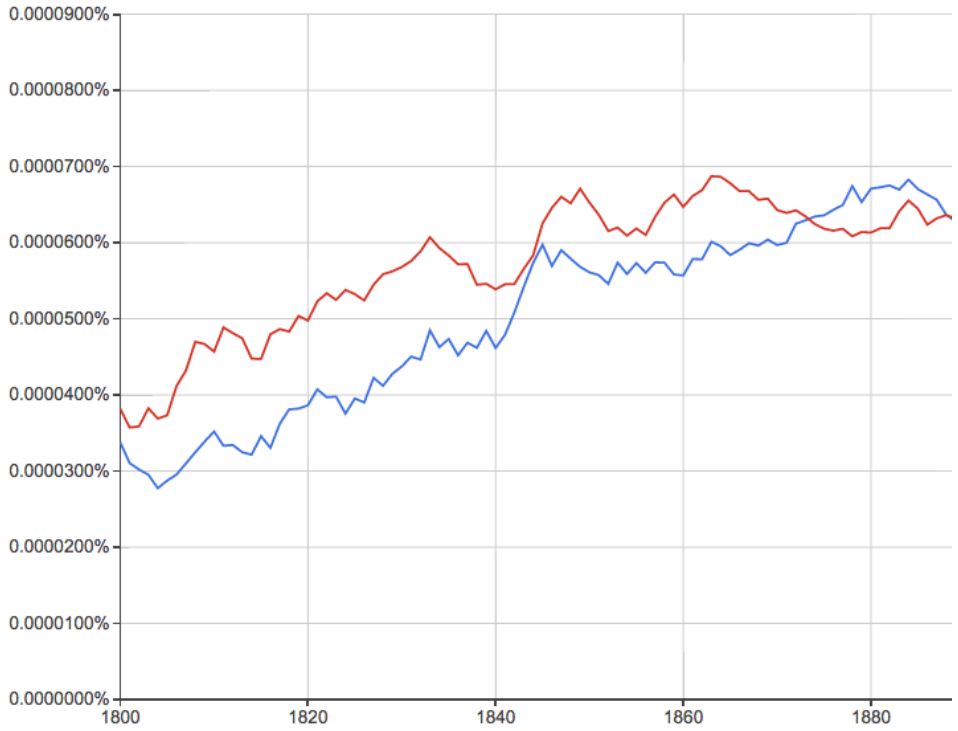


https://books.google.com/ngrams/graph?content=save+forests+from%2Csave+whales+from%2Csave+the+planet+from%2Csave+the+environment+from&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Csave%20forests%20from%3B%2Cc0%3B.t1%3B%2Csave%20whales%20from%3B%2Cc0%3B.t1%3B%2Csave%20the%20planet%20from%3B%2Cc0%3B.t1%3B%2Csave%20the%20environment%20from%3B%2Cc0

save from ≠ save X from Y



save from ≠
save X from Y



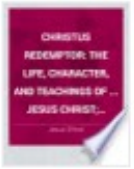
L'Africaine, opera in five acts, etc. [Translated from the French.]

https://books.google.com/books?id=3M5ZAAAAcAAJ

Augustin Eugène SCRIBE - 1871 - Read - More editions

Vaseo. Dally not, or all soon must perish, nor chance of safety more be found. Don Pedro. Is't for me, indeed, thou'rt thus moved, or is it for Ines? , Vaseo. 'Tis true! for her, my beloved, for Ines long adored, whom I must **save from** yawning death ...

Christus redemptor: the life, character, and teachings of ... Jesus ...



https://books.google.com/books?id=JAIDAAAAQAAJ

Henry Southgate - 1874 - Read

... if He did not pluck up the very roots of sinne. He saves us from the guilt, from the power, from the filthi- ness, yea, from the very being of sinne. His salvation is a compleat salvation. It is to save the whole man — to **save from** all evil to all good.

The Living Age ... - Volume 123 - Page 706

https://books.google.com/books?id=F6E_AQAAMAAJ

1874 - Read - More editions

THOU, who dost dwell alone - Thou, who dost know thine own — Thou to whom all are known From the cradle to the grave— Save, oh, **save ! From** the world's temptations, From tribulations; From that fierce anguish Wherein we languish; From ...

Poems of the inner life, selected chiefly from modern authors [by ...

https://books.google.com/books?id=gXQCAAAAQAAJ

Poems, Robert Crompton Jones - 1872 - Read - More editions

... And, when she fain would soar, Makes idols to adore ; Changing the pure emotion Of her high devotion To a skin-deep sense Of her own eloquence : Strong to deceive, strong to enslave— Save, oh, **save ! From** the ingrained fashion Of this ...

Theological Discussion Held at Des Moines, June 22, 1868 - Page 96

https://books.google.com/books?id=VqBDAQAAMAAJ

W. W. King, Alvin Ingals Hobbs - 1868 - Read

To **save from**, or to pardon sin, is to free from punishment due the sinner. Universalism says, " To **save from** sin is to **save from** sinning ; that is, to save me from my friends is to save me from being friendly; to save me from my debts, is to save ...

9 CONCLUSIONS

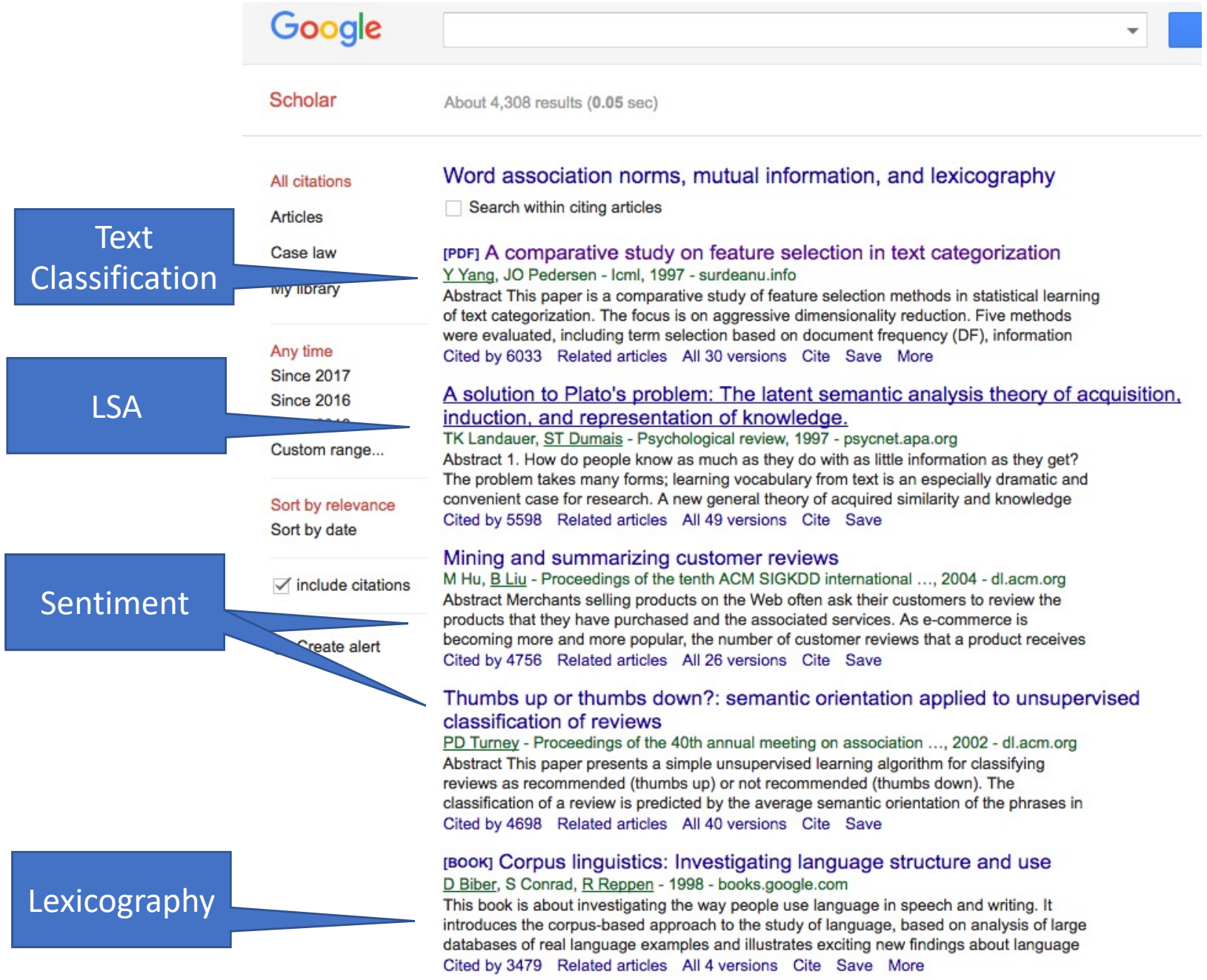
We began this paper with the psycholinguistic notion of word association norm, and extended that concept toward the information theoretic definition of mutual information. This provided a precise statistical calculation that could be applied to a very large corpus of text to produce a table of associations for tens of thousands of words. We were then able to show that the table encoded a number of very interesting patterns ranging from *doctor . . . nurse* to *save . . . from*. We finally concluded by showing how the patterns in the association ratio table might help a lexicographer organize a concordance.

Plan

✓Summarize main points of paper

- Call out
  - ➢some highlights of subsequent literature
  - suggestions for future work

Using Google Scholar to find subsequent work to call out



Levy & Goldberg (NIPS-2014)

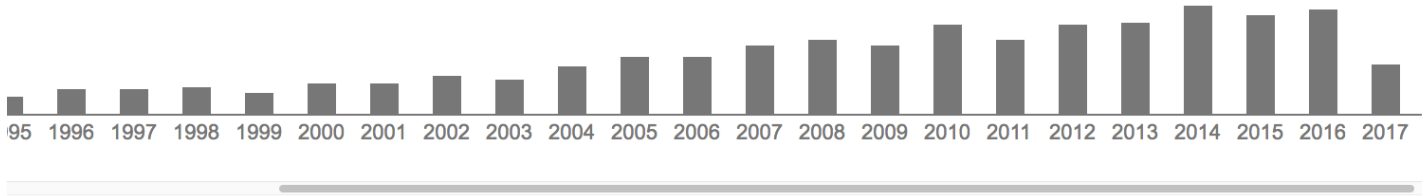
Word2Vec ≈ PMI (Pointwise Mutual Info)

$$sim(x, y) = \cos\big(vec(x), vec(y)\big) \approx PMI(x, y)$$

Word association norms, mutual information, and lexicography [PDF] from aclweb.org

Authors	Kenneth Ward Church, Patrick Hanks
Publication date	1990/3/1
Journal	Computational linguistics
Volume	16
Issue	1
Pages	22-29
Publisher	MIT Press
Description	Abstract The term word association is used in a very particular sense in the psycholinguistic literature.(Generally speaking, subjects respond quicker than normal to the word nurse if it follows a highly associated word such as doctor.) We will extend the term to provide the basis for a statistical description of a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence constraints between verbs and prepositions (content word/ ...
Total citations	Cited by 4269

What happened in 2014?



Scholar articles
Word association norms, mutual information, and lexicography
KW Church, P Hanks - Computational linguistics, 1990
Cited by 4269 - Related articles - All 42 versions

Neural Word Embedding as Implicit Matrix Factorization

Omer Levy
Department of Computer Science
Bar-Ilan University

Yoav Goldberg
Department of Computer Science
Bar-Ilan University

Word2vec is popular (massively cited)

- Word2vec is not first, last or best to discuss
  - Vector spaces, embeddings, analogies, similarity metrics, etc.
- But word2vec is simple and accessible
  - Anyone can download the code and use it in their next paper.
  - Any many do (for better and for worse)
- Available downloads
  - Pre-computed vectors (no training required)
  - Code for training your own vectors on your own corpora

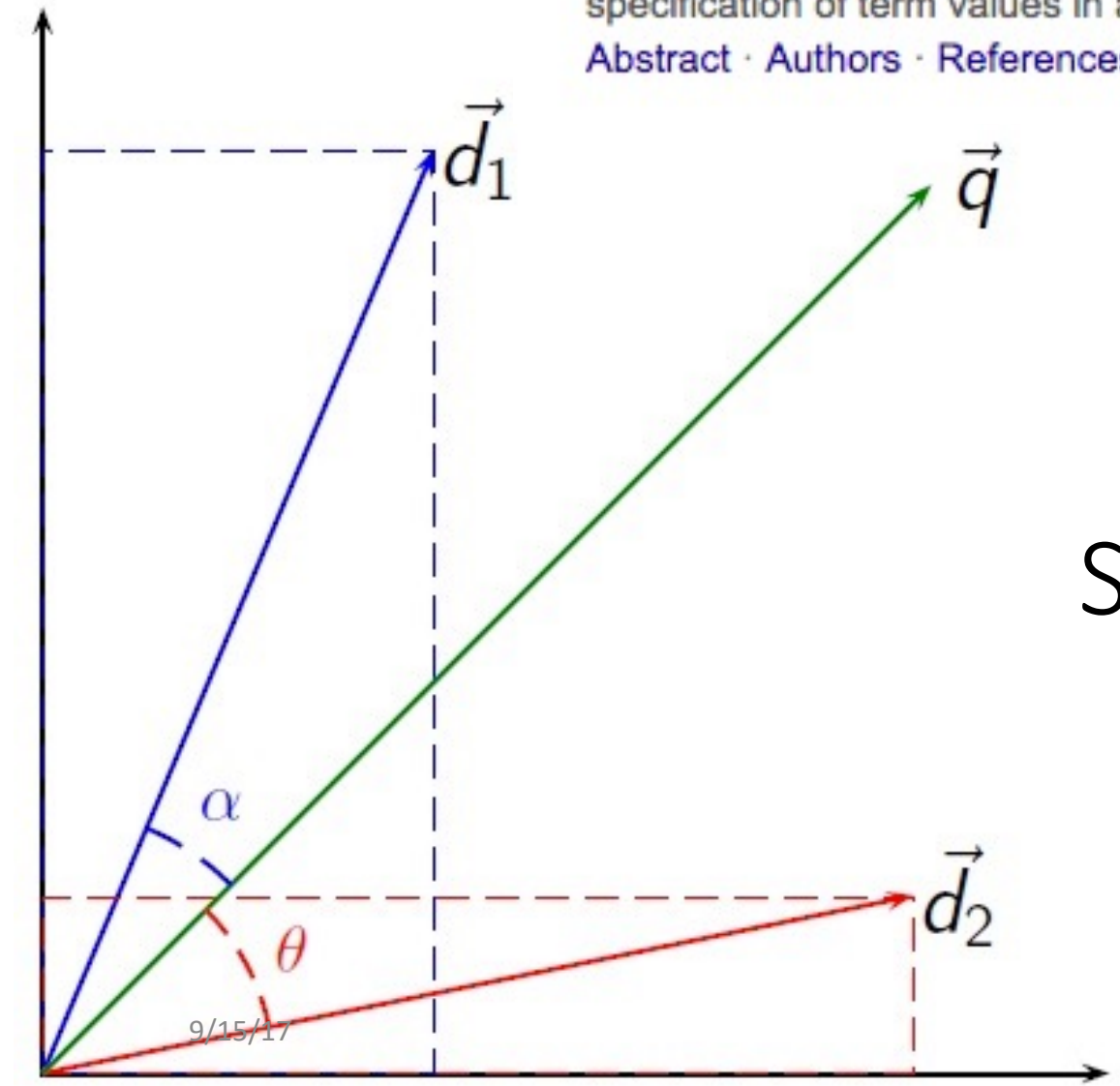
A vector space model for automatic indexing - ACM Digital Library

dl.acm.org/citation.cfm?id=361220 ▾

by G Salton - 1975 - Cited by 7464 - Related articles

A **vector space model** for automatic indexing, Published by ACM .... **Salton**, G., and Yang, C.S. On the specification of term values in automatic indexing.

Abstract · Authors · References · Cited By



Salton's Vector Space Model

Embeddings:
Similarity of docs/words ≈ cos (dot product)

Applications [edit]

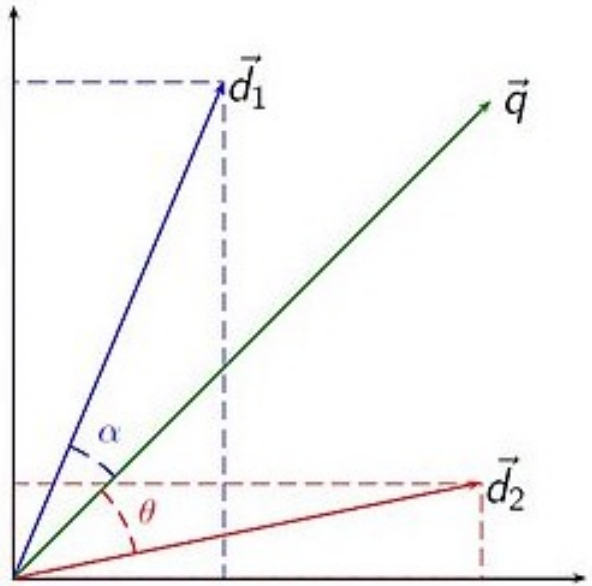
Relevance rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents.

In practice, it is easier to calculate the cosine of the angle between the vectors, instead of the angle itself:

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

Where $\mathbf{d_2} \cdot \mathbf{q}$ is the intersection (i.e. the dot product) of the document (d_2 in the figure to the right) and the query (q in the figure) vectors, $\|\mathbf{d_2}\|$ is the norm of vector d_2, and $\|\mathbf{q}\|$ is the norm of vector q. The norm of a vector is calculated as such:

$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$



As all vectors under consideration by this model are elementwise nonnegative, a cosine value of zero means that the query and document vector are orthogonal and have no match (i.e. the query term does not exist in the document being considered). See cosine similarity for further information.

Information Retrieval (IR) notation
Term Weighting: tf * IDF

- t: term
- d: document
- D: # of documents in library

- tf(t,d): term frequency
  - # of times that t appears in d
- df(t): document frequency
  - # of documents that contain t
  - (at least once)
- IDF(t): inverse doc frequency
  - $IDF(t) = -log_2 \frac{df(t)}{D}$
- tf * IDF weighting
  - Assumes (too much) indep

Example: tf-idf weights [edit]

In the classic vector space model proposed by Salton, Wong and Yang [1] the term-specific weights in the document vectors are products of local and global parameters. The model is known as term frequency-inverse document frequency model. The weight vector for document *d* is $\mathbf{v}_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$, where

$$w_{t,d} = \mathrm{tf}_{t,d} \cdot \log \frac{|D|}{|\{d' \in D \,|\, t \in d'\}|}$$

and

- $\mathrm{tf}_{t,d}$ is term frequency of term *t* in document *d* (a local parameter)
- $\log \frac{|D|}{|\{d' \in D \,|\, t \in d'\}|}$ is inverse document frequency (a global parameter). $|D|$ is the total number of documents in the document set; $|\{d' \in D \,|\, t \in d'\}|$ is the number of documents containing the term *t*.

Using the cosine the similarity between document d_j and query *q* can be calculated as:

$$\mathrm{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

LSA/LSI: Latent Semantic Analysis/Indexing
SVD: Singular Value Decomposition
PCA: Principal Component Analysis

- M: term by doc matrix
  - $M \approx UDV^T$
  - Dimension Reduction
  - s = svd(m)
  - m2 = s$u[,1:2] %*% diag(s$d[1:2]) %*% t(s$v[,1:2])
  - dimnames(m2)=dimnames(m)
- Bellcore Example

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

[PDF] An Introduction to Latent Semantic Analysis - (LSA) at Colorado
lsa.colorado.edu/papers/dp1.LSAintro.pdf
by TK Landauer - Cited by 4471 - Related articles
Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations ...

About 1,090,000,000 results (1.05 seconds)

R: The R Project for Statistical Computing

https://www.r-project.org/

R is a free software environment for statistical computing and graphics. It compiles ... To **download R**, please choose your preferred CRAN mirror. If you have ...

You've visited this page many times. Last visit: 7/29/17

Results from r-project.org

Windows

R-3.4.1 for Windows (32/64 bit).
Download R 3.4.1 for Windows ...

R for Mac OS X

R for Mac OS X. This directory contains binaries for a base ...

R-3.2.4 for Windows (32/64 bit)

R-3.2.4 for Windows (32/64 bit).
Download R 3.2.4 for Windows ...

R-3.3.2 for Windows (32/64 bit)

R-3.3.2 for Windows (32/64 bit).
Download R 3.3.2 for Windows ...

R 3.2.2

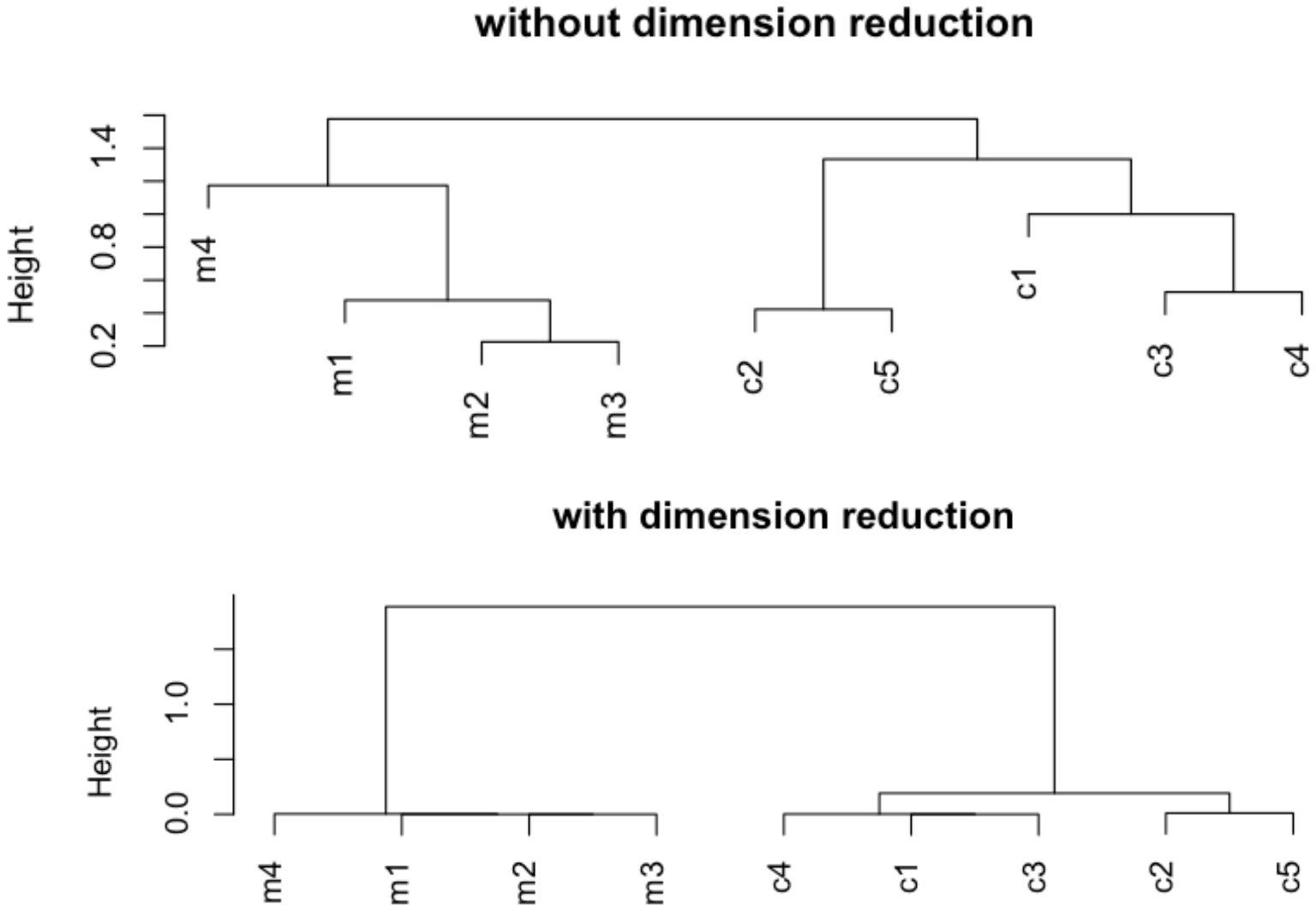
R-3.2.2 for Windows (32/64 bit).
Download R 3.2.2 for Windows ...

R-2.15.2 for Windows

R-2.15.2 for Windows (32/64 bit).
Download R 2.15.2 for ...

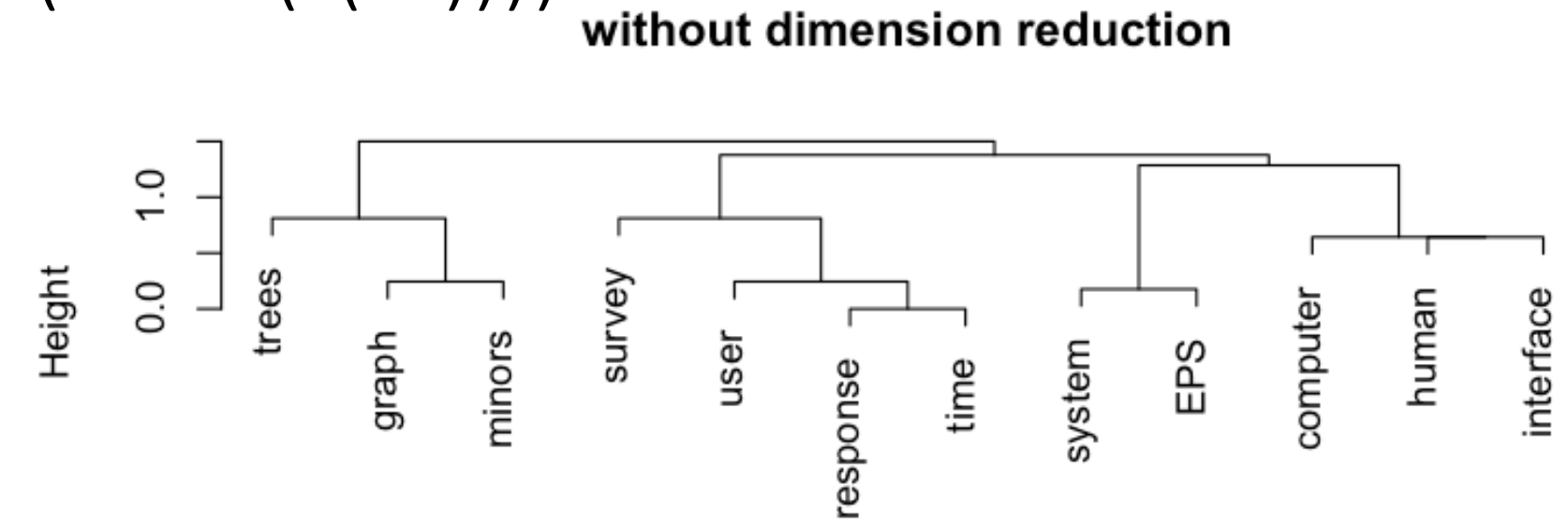
plot(hclust(as.dist(1-cor(m))))

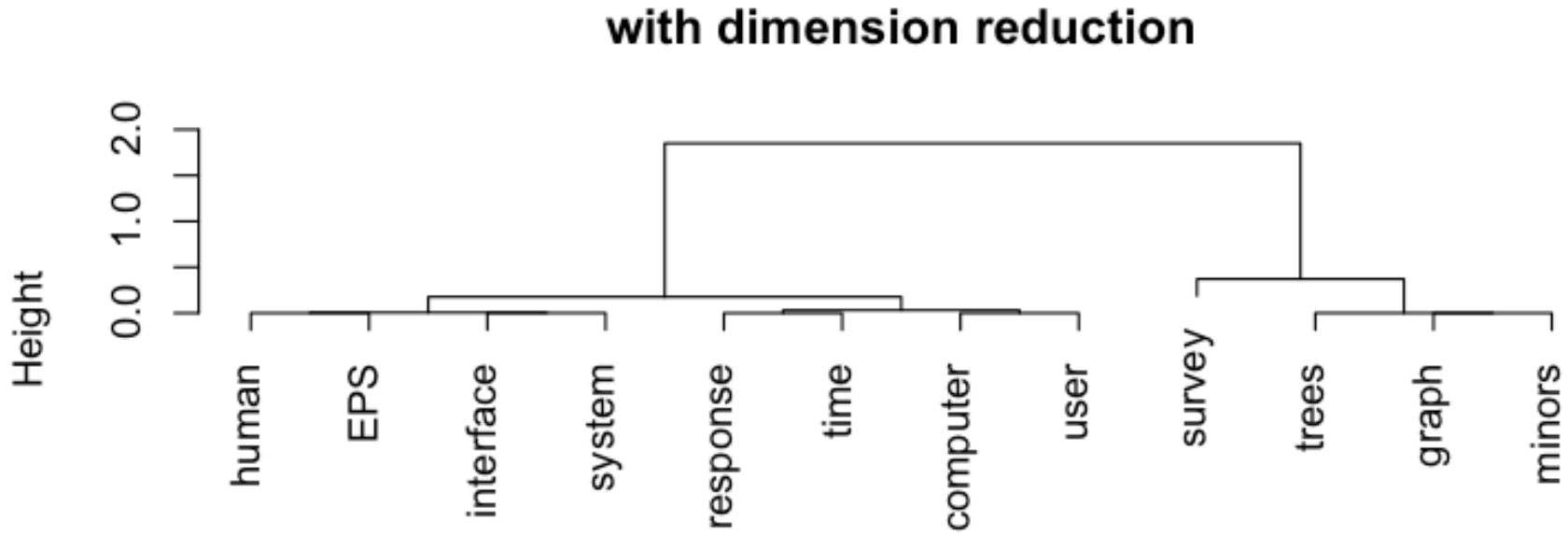
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1



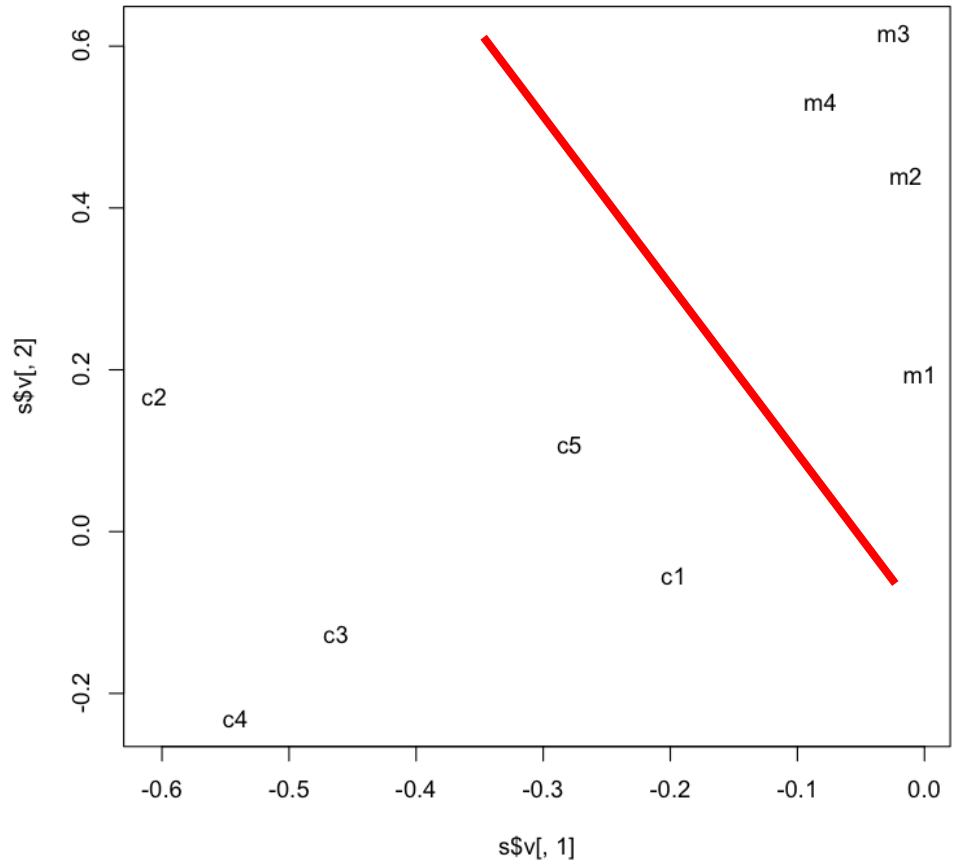
plot(hclust(as.dist(1-cor(t(m)))))

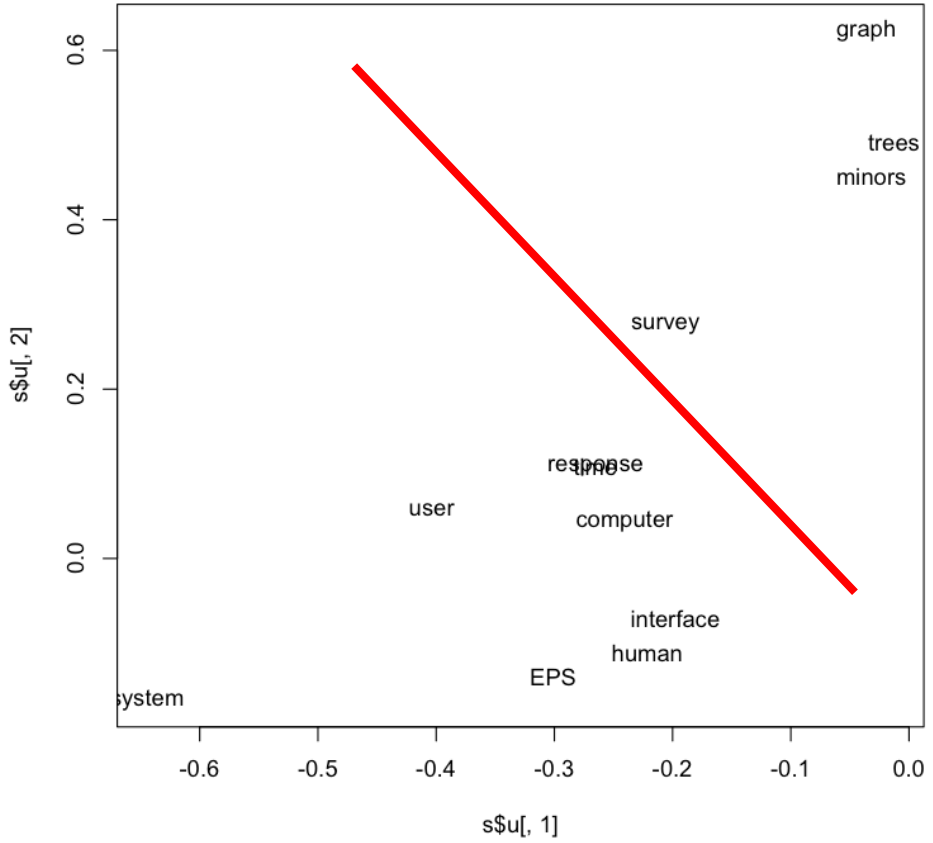
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1





LSA maps terms & docs into internal dimensions



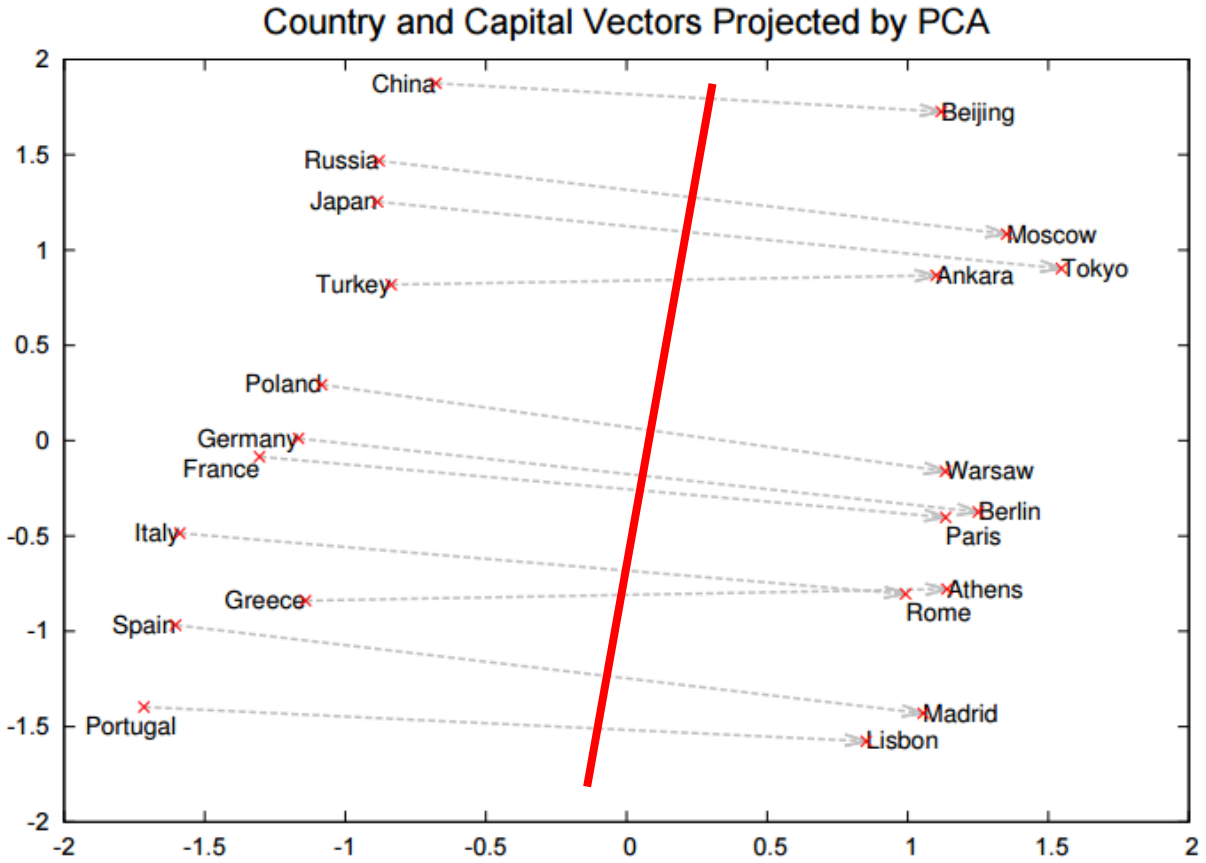


Word2vec is popular (massively cited)

- **Word2vec** is not first, last or best to discuss
  - Vector spaces, embeddings, **analogies**, similarity metrics, etc.
- But word2vec is simple and accessible
  - Anyone can download the code and use it in their next paper.
  - Any many do (for better and for worse)
- Available downloads
  - Pre-computed vectors (no training required)
  - Code for training your own vectors on your own corpora

Word2Vec: $sim(x, y) = \cos(vec(x), vec(y)) \approx PMI(x, y)$

https://code.google.com/archive/p/word2vec/



Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.

Ilya Sutskever
Google Inc.

Kai Chen
Google Inc.

- Linguistic generalizations
  - Word associations (distance in plot)
  - Features (red line)
    - Countries & Capitals
- Analogies:
  - Man : Woman :: King : *x*
    - *x* → queen
  - Athens : Greece :: Bangkok: *x*
    - *x* → Thailand
- Vector Space (Salton)
  - Addition & subtraction
  - Clustering, PCA
- Convenient for Neural Networks

man : woman :: king : ***x***

- Vector addition & subtraction
  - $vec(king + woman\ - man) = vec(king) + vec(woman) - vec(man)$
- Analogies
  - $\hat{x} = \underset{x \in V}{\text{ARGMAX}}\, sim(x', king + woman\ - man)$

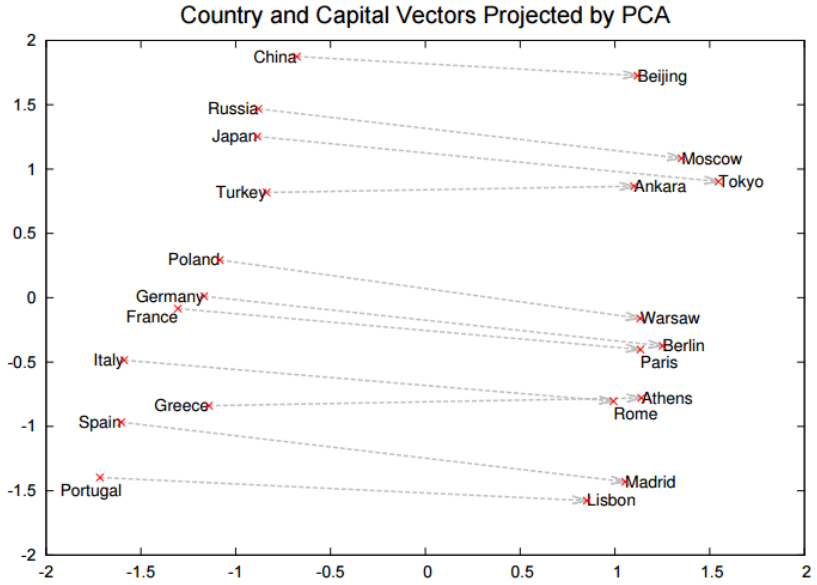
x	Gender	Number
Queen	f	sg
Monarch	m	sg
Princess	f	sg
Crown˙prince	m	sg
Prince	m	sg
Kings	m	pl
Queen˙Consort	m	sg
Queens	f	pl
Sultan	m	sg
Monarchy	m	sg

Some analogies are easier than others

- Tweets
  - RT @tallinzen: sure, king:queen etc, but did you know word2vec gets real SAT analogies right just 1% of the time?
  - 15 copies of this tweet
    - Some by NLP experts
- Resources Debate
  - WordNet &
  - British National Corpora

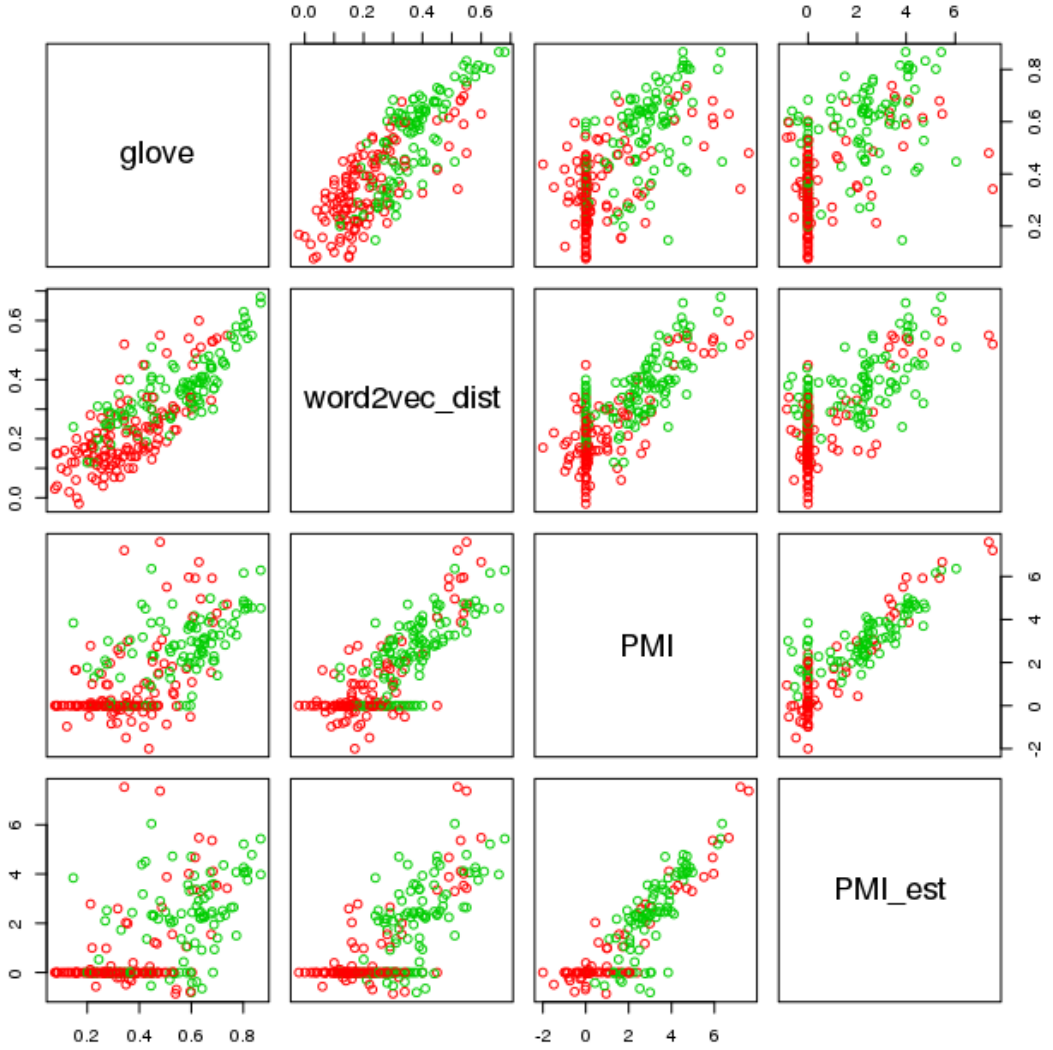
Table 2. *Some types of analogies are easier than others, as indicated by accuracies for top choice (A_1), as well as top 2 (A_2), top 10 (A_{10}) and top 20 (A_{20}). The rows are sorted by A_1. These analogies and the type classification come from the questions-words test set, except for the last row, SAT questions. SAT questions are harder than questions-words*

A_1	A_2	A_{10}	A_{20}	N	Analogy type	Example
0.91	0.95	0.98	0.99	1,332	Comparative	young : younger :: wide : wider
0.90	0.94	0.97	0.98	1,599	Nationality-adjective	Ukraine : Ukrainian :: Switzerland : Swiss
0.90	0.93	0.97	0.98	1,332	Plural	woman : women :: snake : snakes
0.87	0.94	1.00	1.00	1,122	Superlative	young : youngest :: wide : widest
0.85	0.90	0.97	1.00	506	Family	uncle : aunt :: stepson : stepdaughter
0.83	0.89	0.97	0.98	335	Capital-countries	Tokyo : Japan :: Tehran : Iran
0.79	0.86	0.94	0.96	4,695	Capital-world	Zagreb : Croatia :: Dublin : Ireland
0.78	0.84	0.98	0.99	1,056	Present-participle	write : writing :: walk : walking
0.71	0.79	0.90	0.92	2,467	City-in-state	Worcester : Massachusetts :: Cincinnati : Ohio
0.68	0.78	0.93	0.95	870	Plural-verbs	write : writes :: work : works
0.66	0.82	0.97	0.98	1,560	Past-tense	writing : wrote :: walking : walked
0.43	0.48	0.64	0.69	812	Opposite	tasteful : distasteful :: sure : unsure
0.35	0.42	0.57	0.62	866	Currency	Vietnam : dong :: USA : dollar
0.29	0.37	0.63	0.73	992	Adjective-to-adverb	usual : usually :: unfortunate : unfortunately
0.01	0.02	0.08	0.10	190	SAT questions	audacious : boldness :: sanctimonious : hypocrisy



Levy & Goldberg (NIPS-2014)

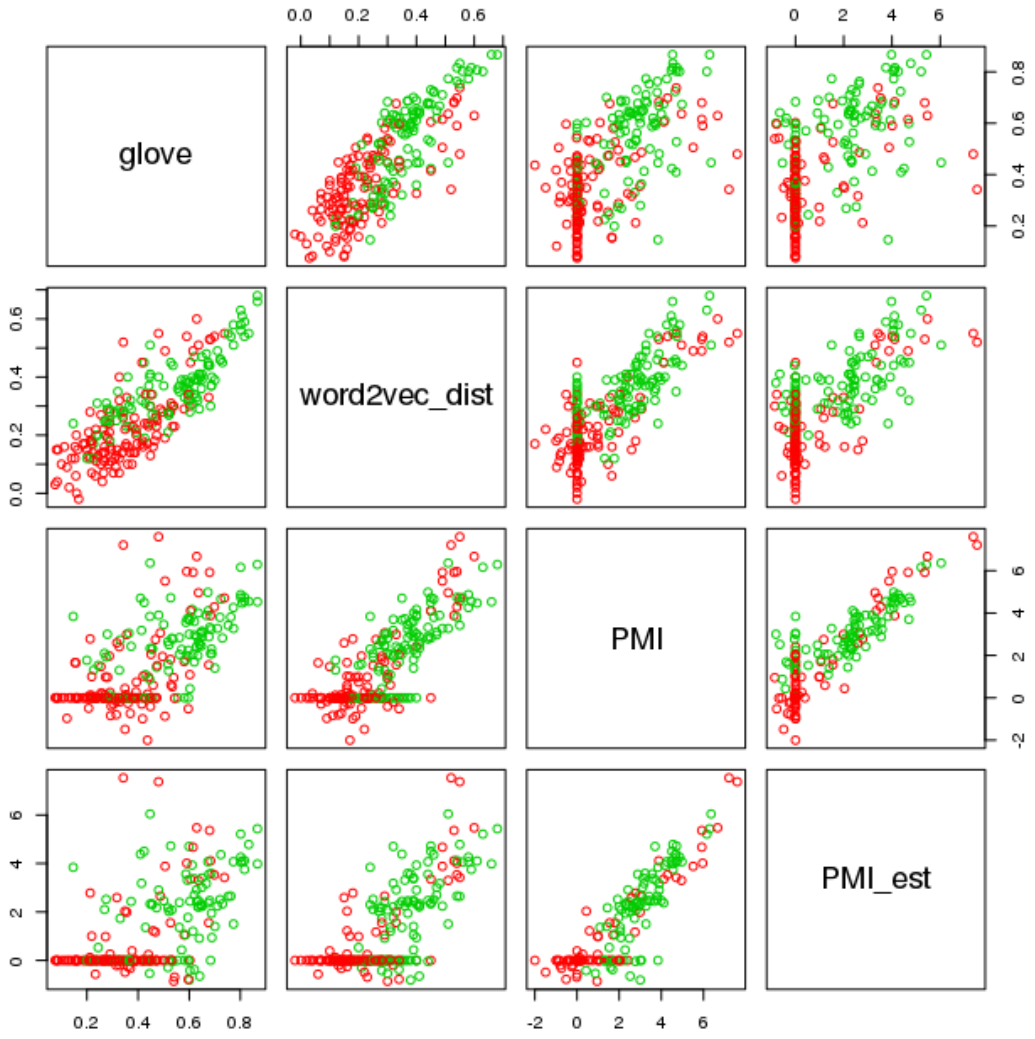
Word2Vec ≈ PMI (Pointwise Mutual Info)

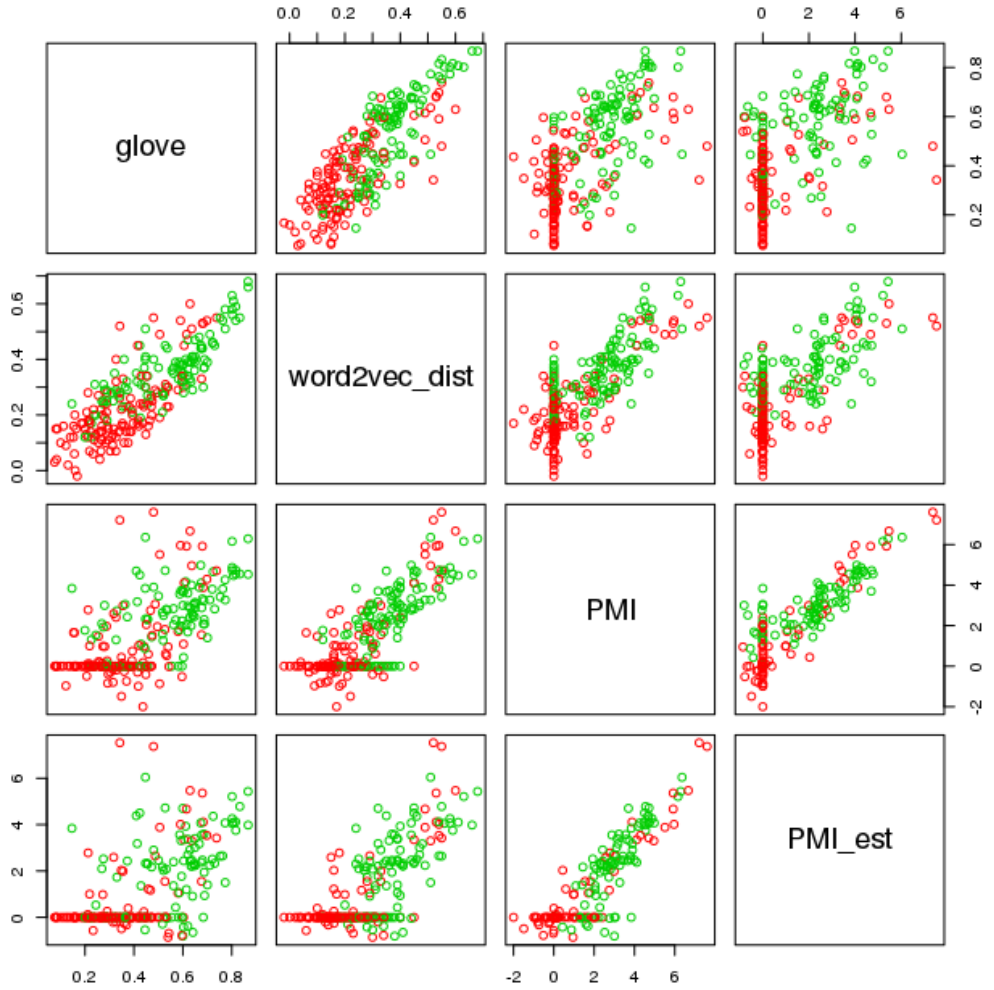


- Levy & Goldberg (NIPS-2014) is a theoretical arg
  - Plots → correlations are large, but far from perfect
- Materials:
  - N = 22 words (11 cities + 11 countries)
  - N (N-1)/2 = 231 pairs of words (points)
  - type in {city, country}
- Color:
  - **Green** → type match
  - **Red** → type mismatch

GloVe: https://nlp.stanford.edu/projects/glove/

- Alternative to Word2vec:
  - Similar performance
    - Both Word2vec & GloVe ≈ PMI
  - More interpretable:
    - Neural net → SGD minimization of a sensible loss
  - More generous license
- Both methods output embedding M
  - M: V by K matrix where $M \times M^T \approx P$
  - V: size of vocabulary
  - K: # of internal dimensions
  - P: V by V matrix of PMI values
- Space: GloVe materializes a V^2 matrix while computing M (unlike Word2vec)

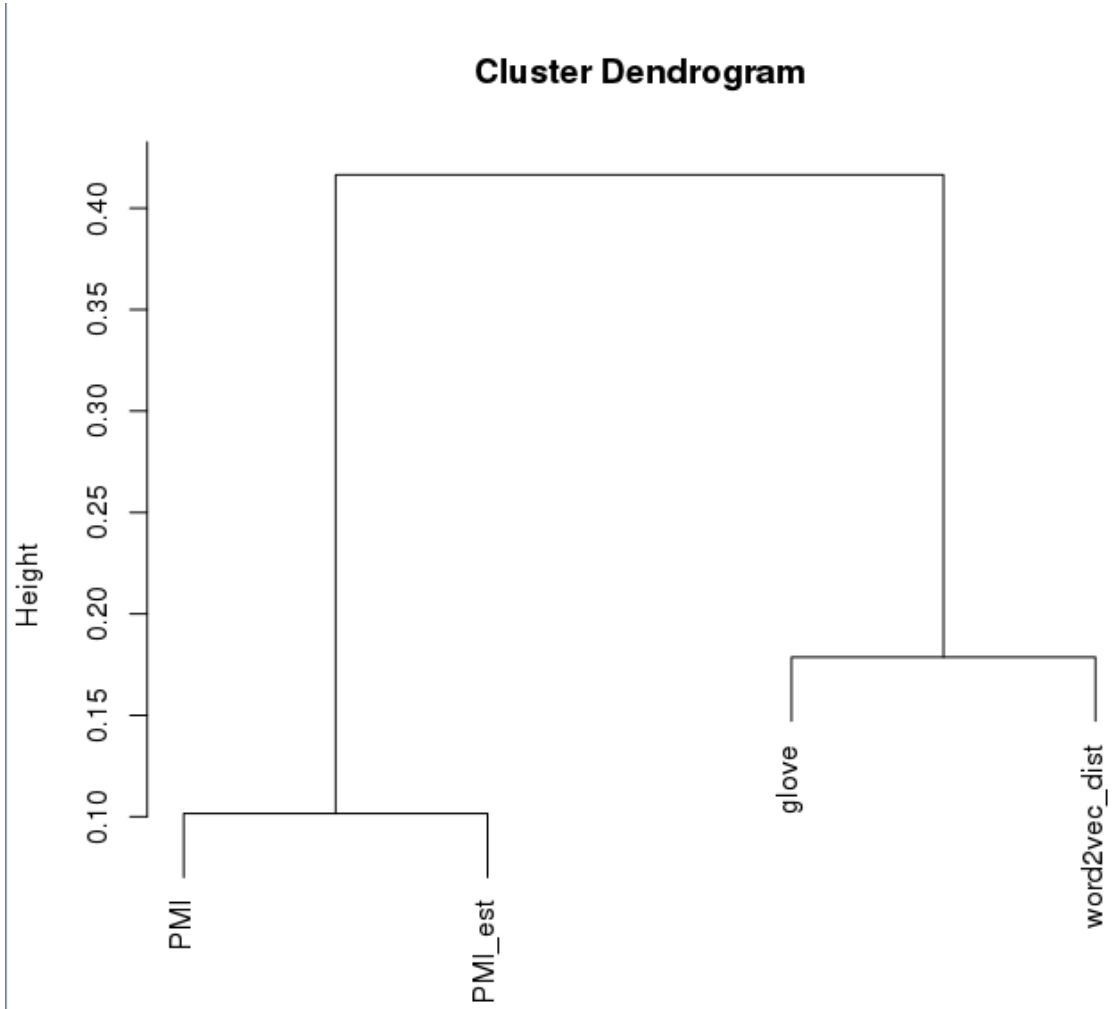




Glass is half-full/empty

Word2vec is similar to PMI (but different)

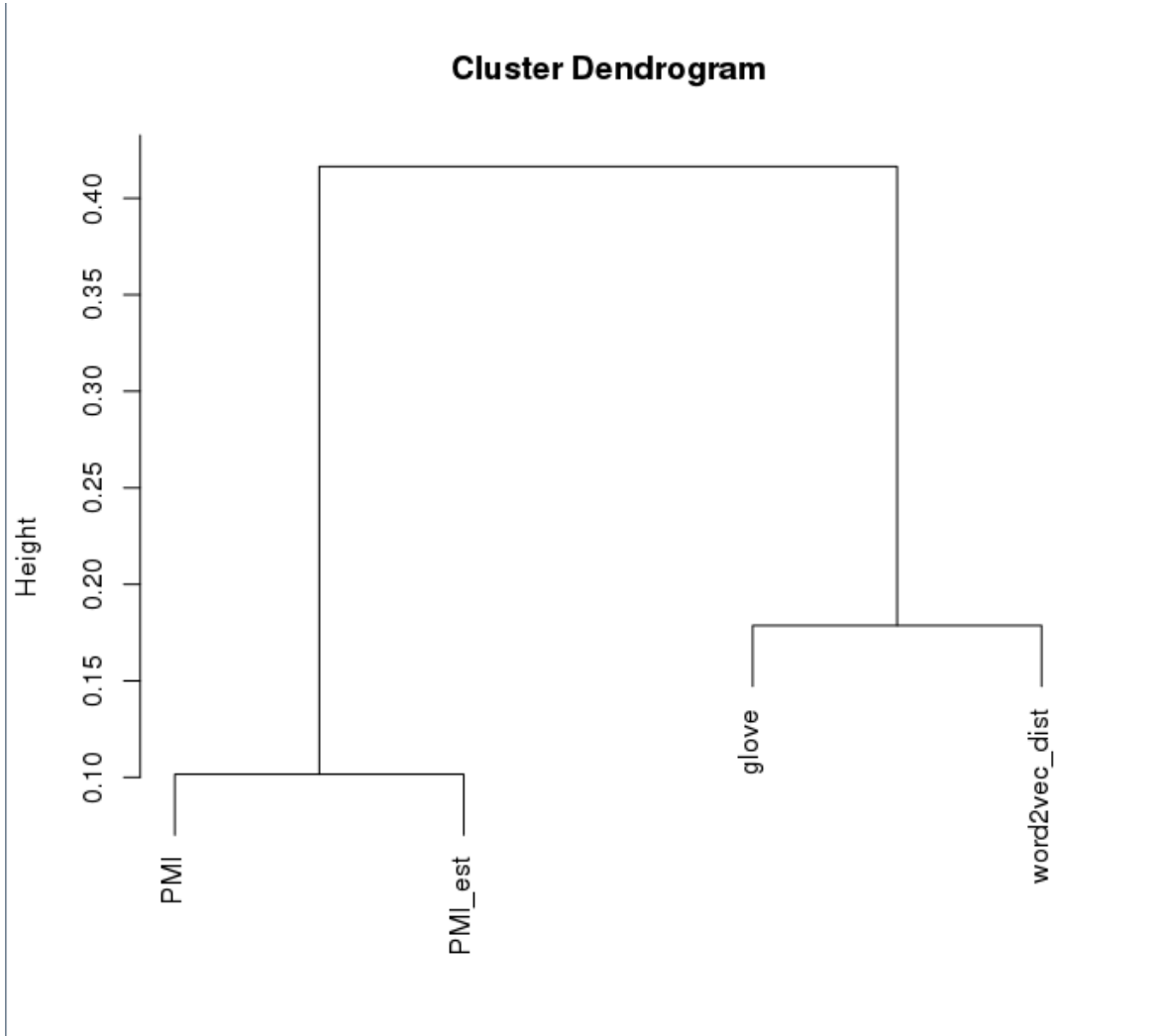
	glove	word2vec	PMI	PMI_est
glove	1.00	0.82	0.65	0.58
word2vec	0.82	1.00	0.80	0.73
PMI	0.65	0.80	1.00	0.90
PMI_est	0.58	0.73	0.90	1.00

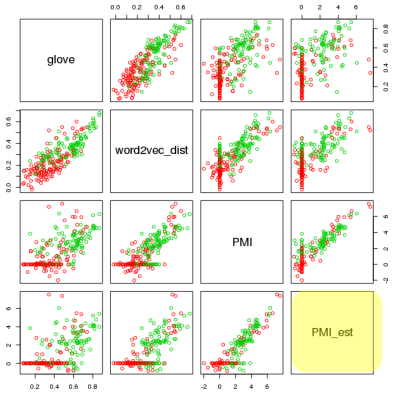


IMHO: Word2vec & GloVe >> PMI

- Glass is half full
  - Levy & Goldberg are correct in pointing out similarities
- But half empty
  - Word2vec & GloVe do better in practice
  - Because of “tricks” for small counts, etc.
  - Code is not well understood
    - Even though there isn’t much there
    - 702 lines in word2vec.c; 446 in glove.c
- Dimension Reduction
  - Matrix completion (?!?)
  - Obvious suggestion, SVD on PMIs
    - Disappointing in practice
    - May not want to model PMI too well

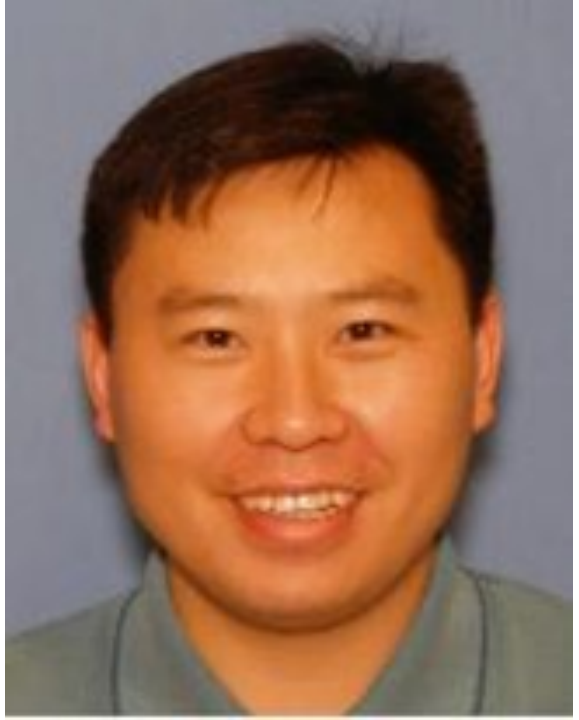
	glove	word2vec	PMI	PMI_est
glove	1.00	0.82	0.65	0.58
word2vec	0.82	1.00	0.80	0.73
PMI	0.65	0.80	1.00	0.90
PMI_est	0.58	0.73	0.90	1.00





Estimating PMI (and Contingency Tables)

Li & Church, Computational Linguistics (2007)



http://stat.rutgers.edu/home/pingli/

- Compute Sample Contingency Table
- MLE Inference:
  - What is the most likely original table
  - Given sample and margins?
- Evaluation

Sample contingency table

	W_2	$\sim W_2$
W_1	a_s	b_s
$\sim W_1$	c_s	d_s

Original contingency table

	W_2	$\sim W_2$
W_1	a	b
$\sim W_1$	c	d

Prior Art: Removing (Near) Duplicate Web Pages

Reasons for duplicate filtering

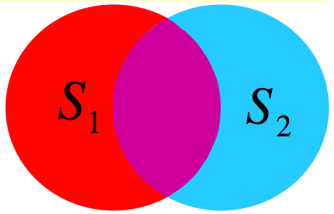
- Proliferation of almost but not quite equal documents on the Web:
  - Legitimate: Mirrors, local copies, updates, etc.
  - Malicious: Spammers, spider traps, dynamic URLs, "cookie crumbs"
  - Mistaken: Spider errors
- Costs:
  - RAM and disks
  - Unhappy users
- Approximately 30% of the pages on the web are (near) duplicates. [B,Glassman,Manasse & Zweig '97, Shivakumar & Garcia-Molina '98]
- In enterprise search even larger amount of duplication.

Sampling minima

- Apply a random permutation σ to the set $[0..2^{64}]$
- Crucial fact

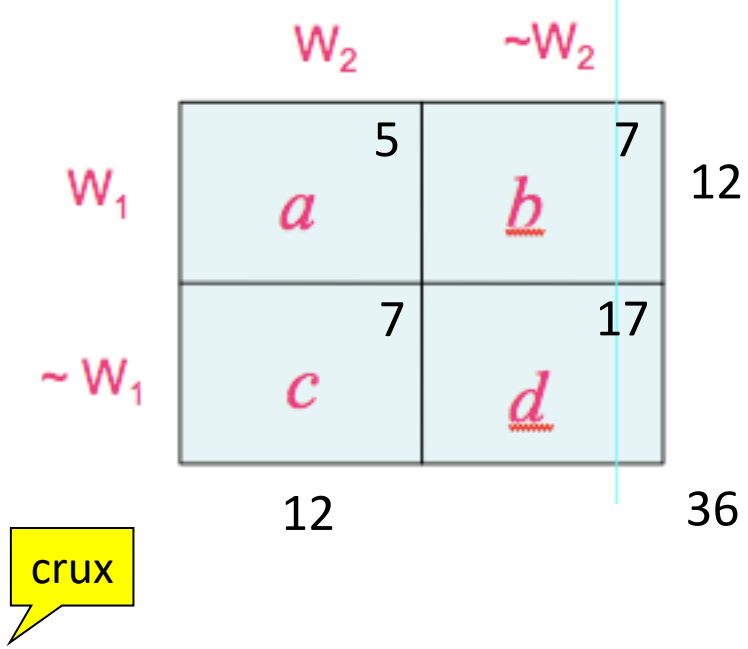
Let $\alpha = \sigma^{-1}(\min(\sigma(S_1)))$ $\beta = \sigma^{-1}(\min(\sigma(S_2)))$

$$\Pr(\alpha = \beta) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$



- More generally, we look at the k smallest elements in $S_1 \cup S_2$ and check how many are in common.

Computing Contingency Tables (Brute Force)

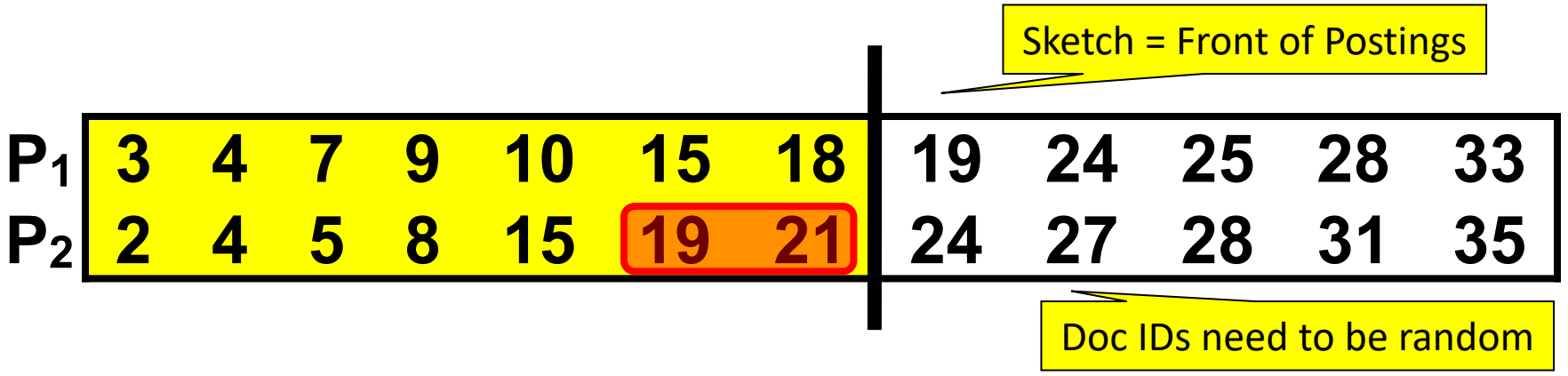


P_1	3	4	7	9	10	15	18	19	24	25	28	33
P_2	2	4	5	8	15	19	21	24	27	28	31	35

Computing Sample Contingency Table (Using Sketches, An Approximation: Yellow)

- Key: throw out red {19, 21}
  - D_s = min(max(sketch(w_1)), max(sketch(w_2))) = 18
  - Can't tell that 19 is in intersection and 21 is not
    - Without looking outside the sketch

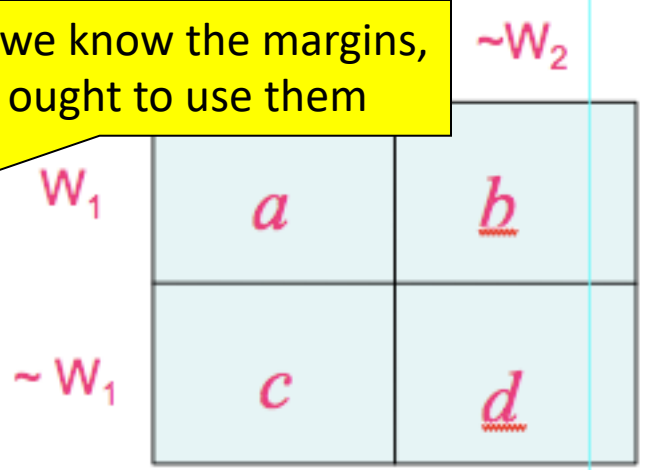
D_s = Sample Size



MLE Inference:

What is the most likely table
Given sample and margins?

When we know the margins,
We ought to use them



D_s/D = Sampling Rate

$$\hat{a}_{ME} = \frac{D}{D_s} a_s$$

- Consider all possible contingency tables:
  - *a, b, c & d*
- Select the table that maximizes the probability of observations
  - a_s, b_s, c_s & d_s

$$\hat{a}_{MLE} = \arg\max_{a} P(a_s, b_s, c_s, d_s \mid D_s; a)$$

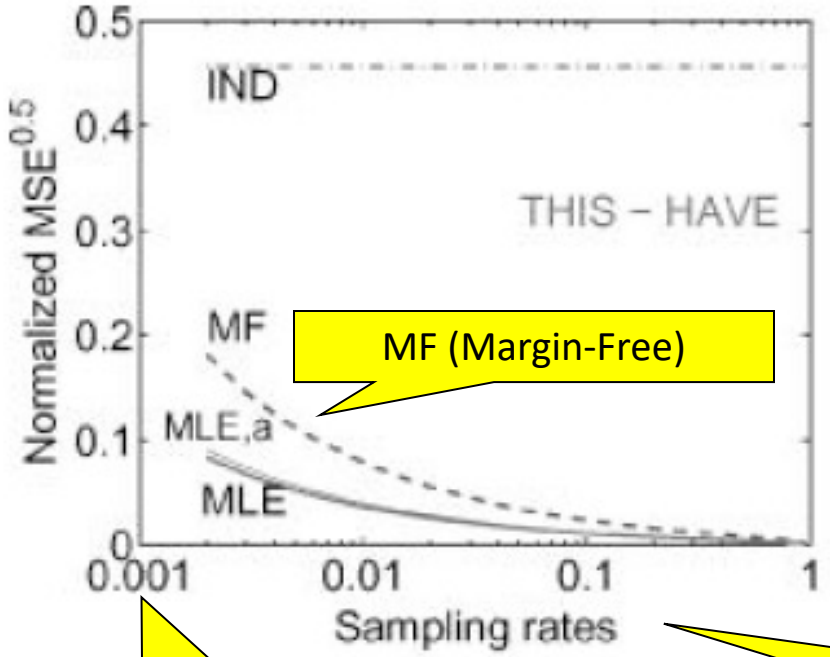
$$P(a_s, b_s, c_s, d_s \mid D_s; a)$$

$$= \binom{a}{a_s}\binom{b}{b_s}\binom{c}{c_s}\binom{d}{d_s} \Big/ \binom{D}{D_s}$$

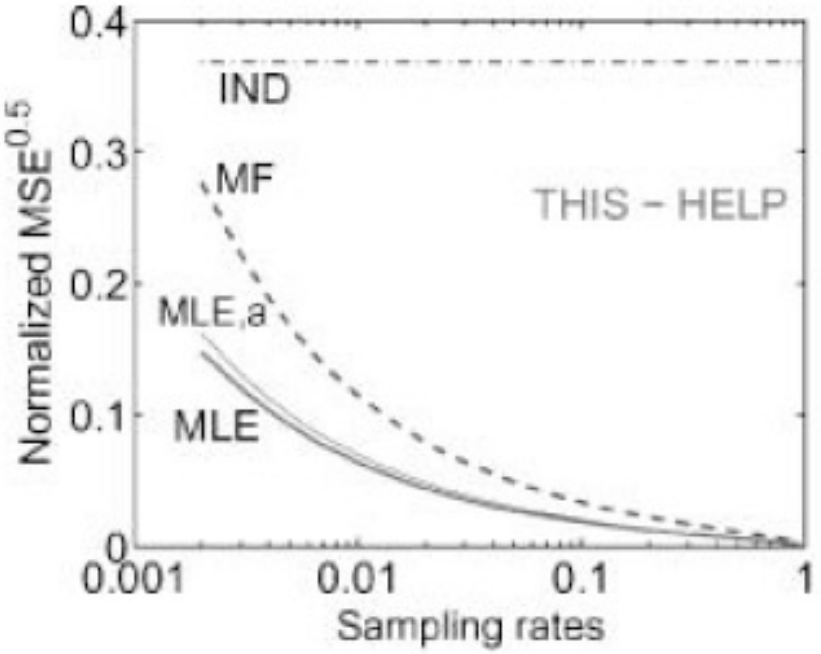
$$\hat{a}_{MLE,i} = \frac{f_x(2a_s + c_s) + f_y(2a_s + b_s) - \sqrt{\left(f_x(2a_s + c_s) + f_y(2a_s + b_s)\right)^2 - 8 f_x f_y a_s (2a_s + b_s + c_s)}}{2(2a_s + b_s + c_s)}$$

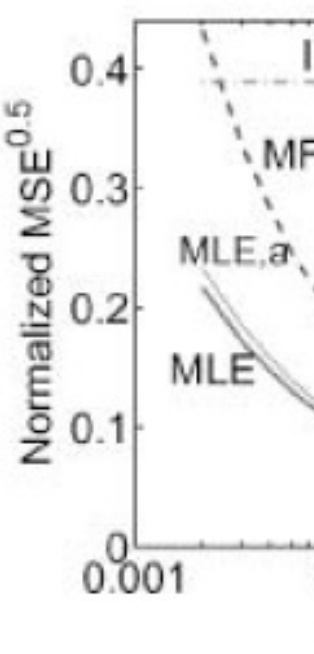
Empirical Evaluation
MLE >> MF >> Independence

IND (assume independence): too much error



MF (Margin-Free)



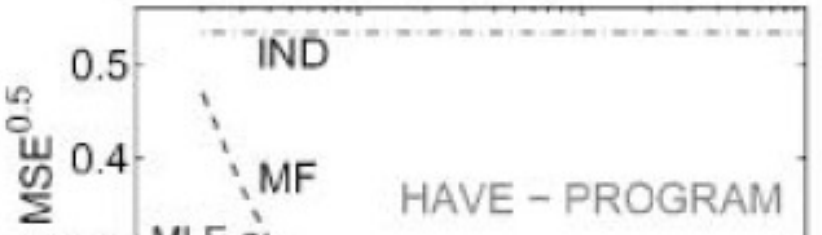


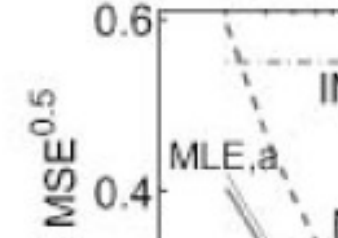
Target: minimize effort & error

(a) Case 2-1

D_s/D = Sampling Rate (effort)

Case 2-2





Non-Empirical Evaluation
(Closed Form)

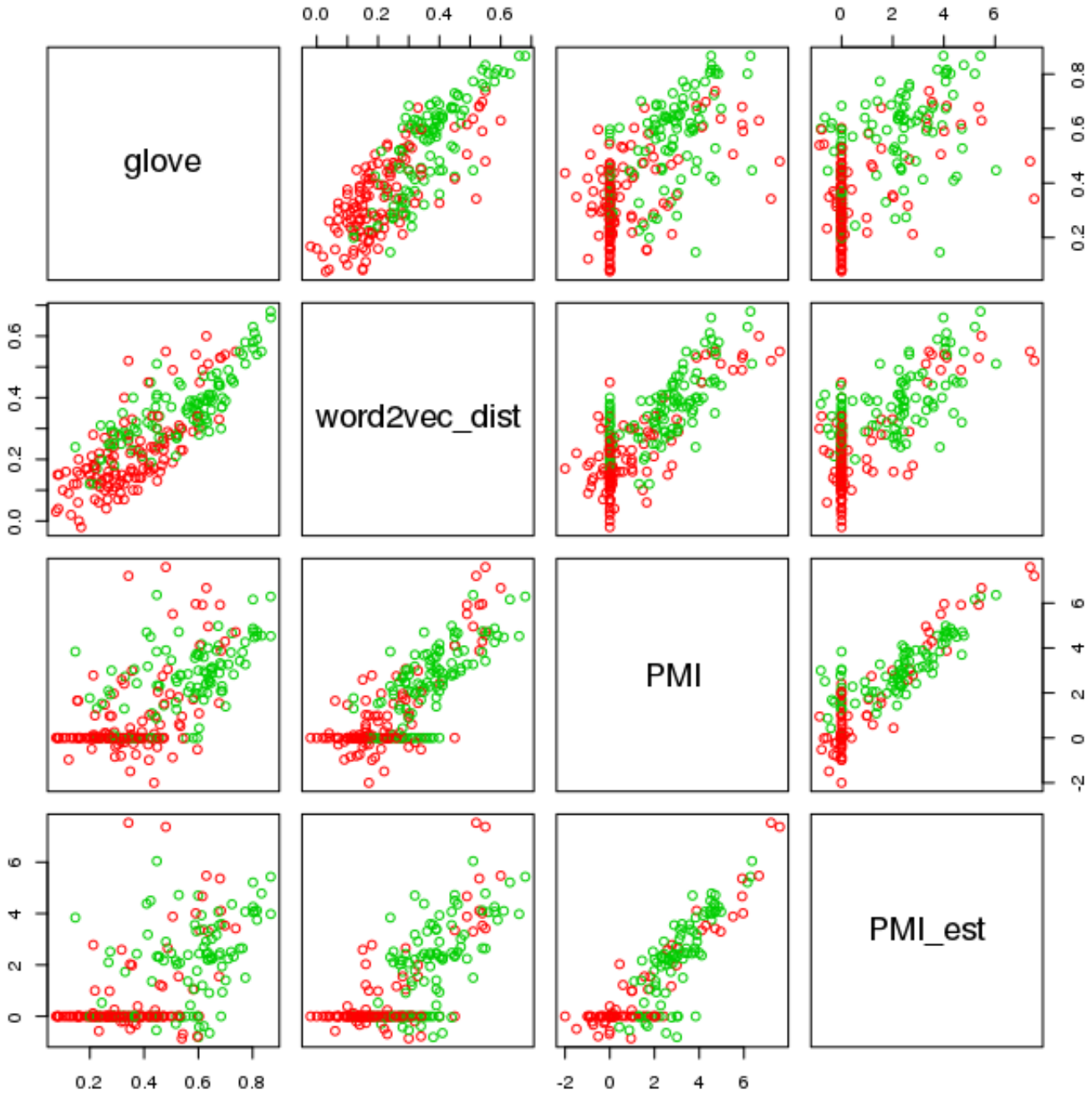
$$\mathrm{Var}\left(\hat{a}_{MLE}\right) \approx \frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_x - a} + \frac{1}{f_y - a} + \frac{1}{D - f_x - f_y + a}}$$

$$\mathrm{Var}(\hat{a}_{MF}) = \frac{D}{D_s} \frac{1}{\frac{1}{a} + \frac{1}{D-a}} \frac{D - D_s}{D - 1}$$

- Non-Empirical Evaluation:
  - MLE has smaller variance than Margin Free (MF) baseline
- Confirms Empirical Evaluation:
  - MLE has less error than MF
- When we have margins, we should use them
  - Improves accuracy as well as consistency

Sketches vs. Word2Vec (as estimates of PMI)

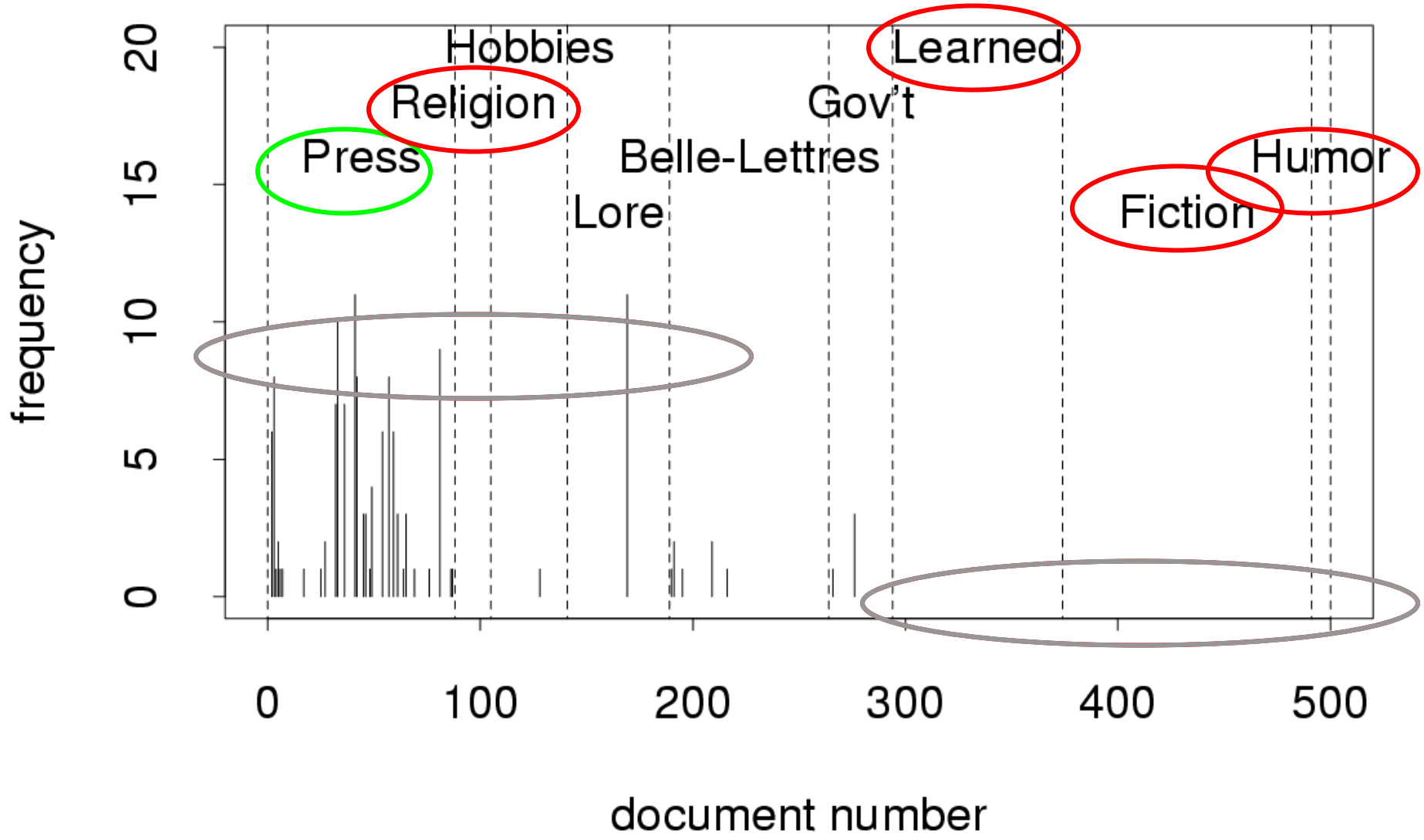
- It is said that word2vec is fast
  - But sketches are also fast
  - and closer to PMI
- But is closer to PMI better?

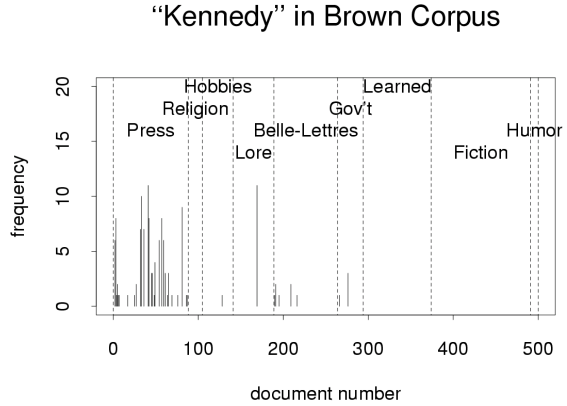


Interestingness Metrics: Deviations from Independence

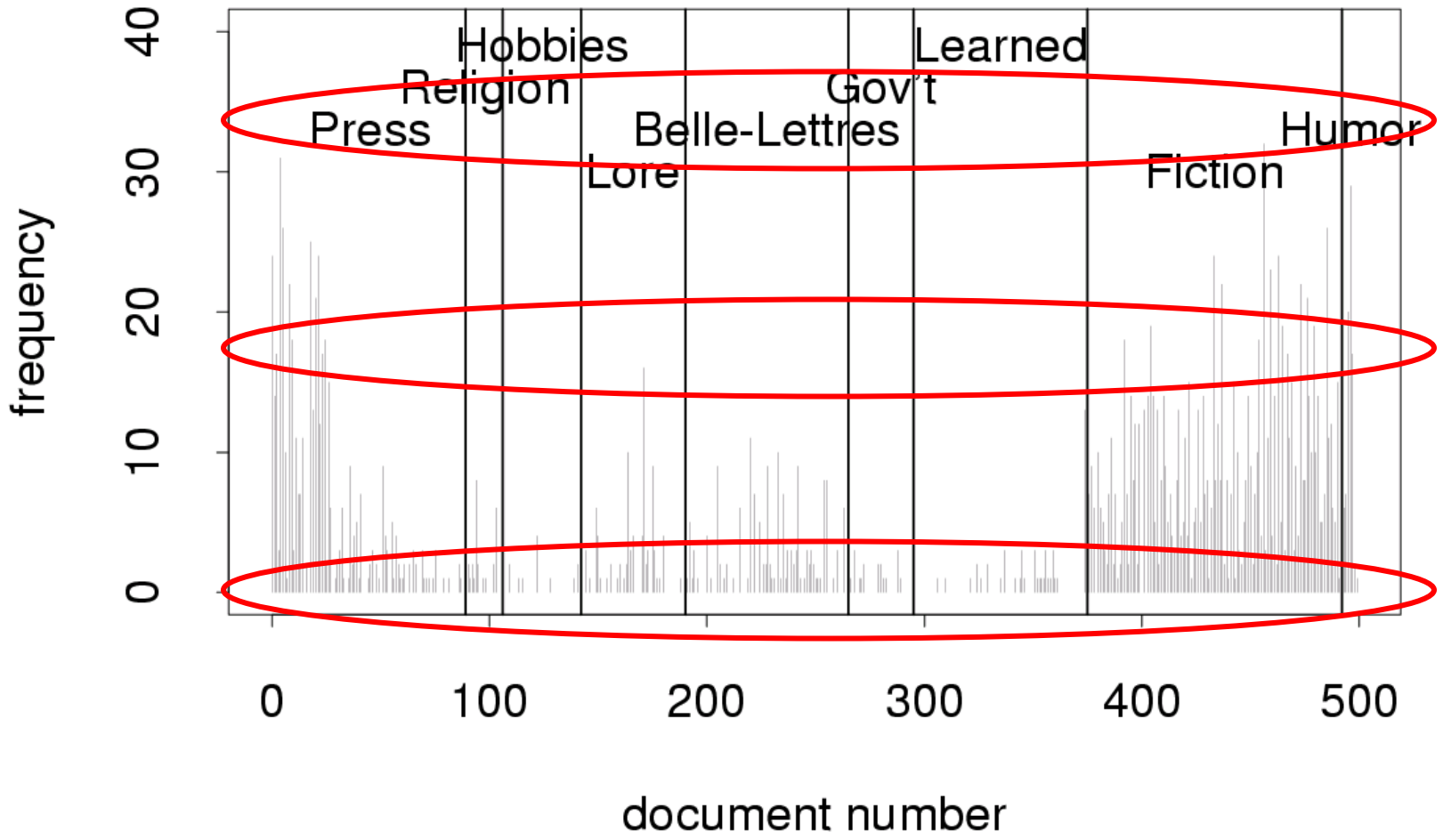
- Poisson (and other independence assumptions)
  - Not bad for meaningless random strings
- Deviations from Poisson are clues for hidden variables
  - Meaning, content, genre, topic, author, etc.
- Analogous to mutual information (Hanks)
  - Pr(doctor…nurse) >> Pr(doctor) Pr(nurse)

“Kennedy” in Brown Corpus

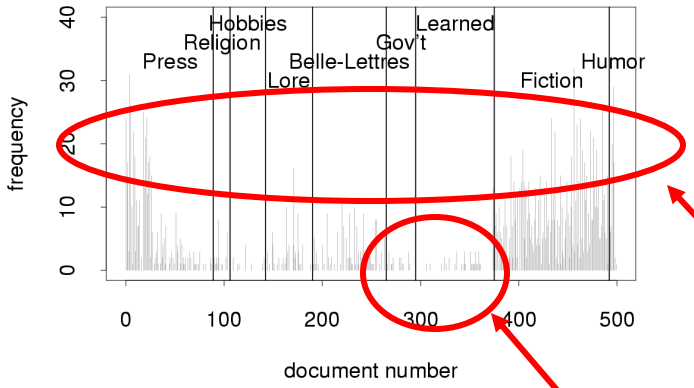




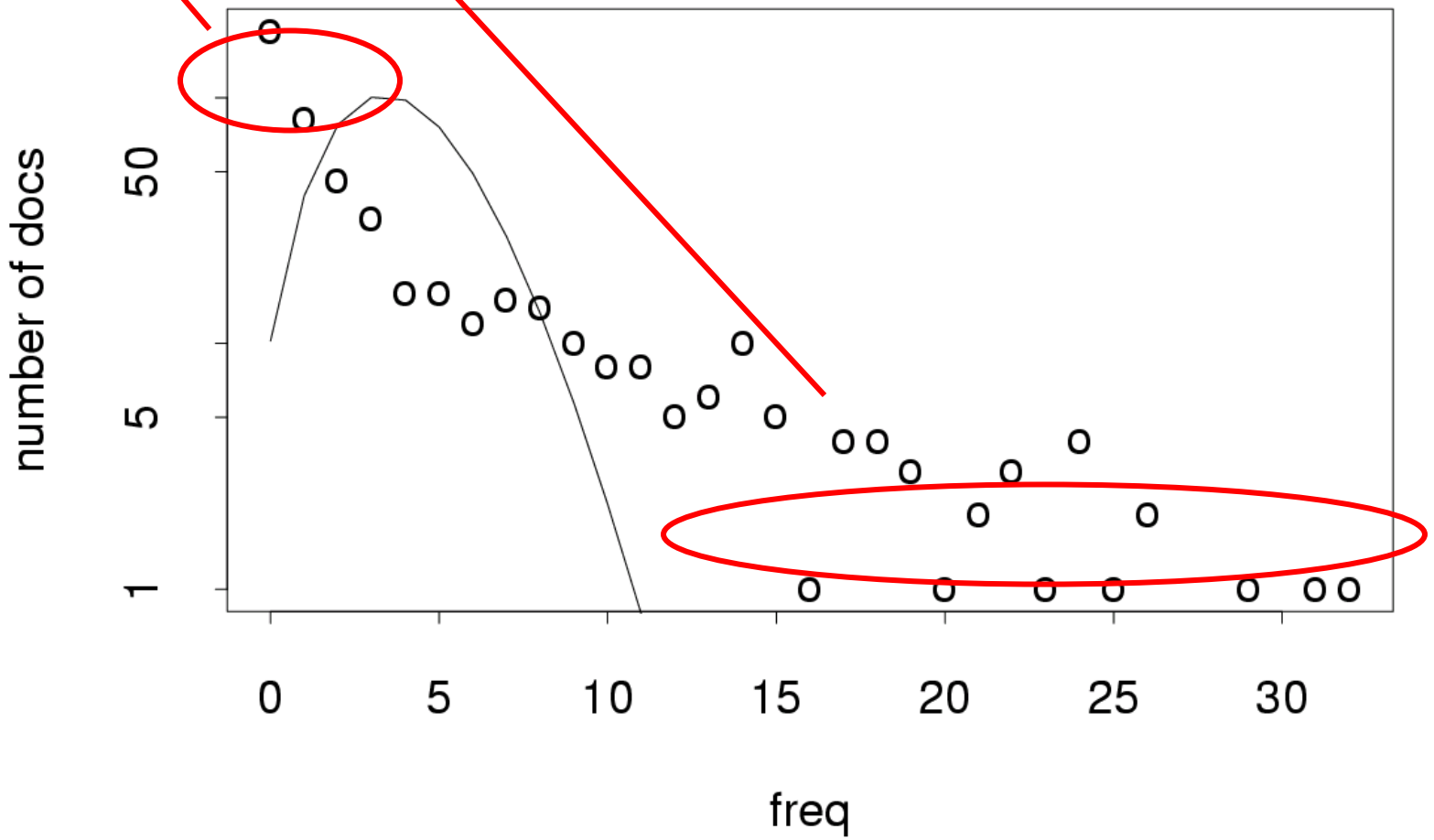
"said" in Brown Corpus



"said" in Brown Corpus



Poisson Doesn't Fit



Adaptation: Three Approaches

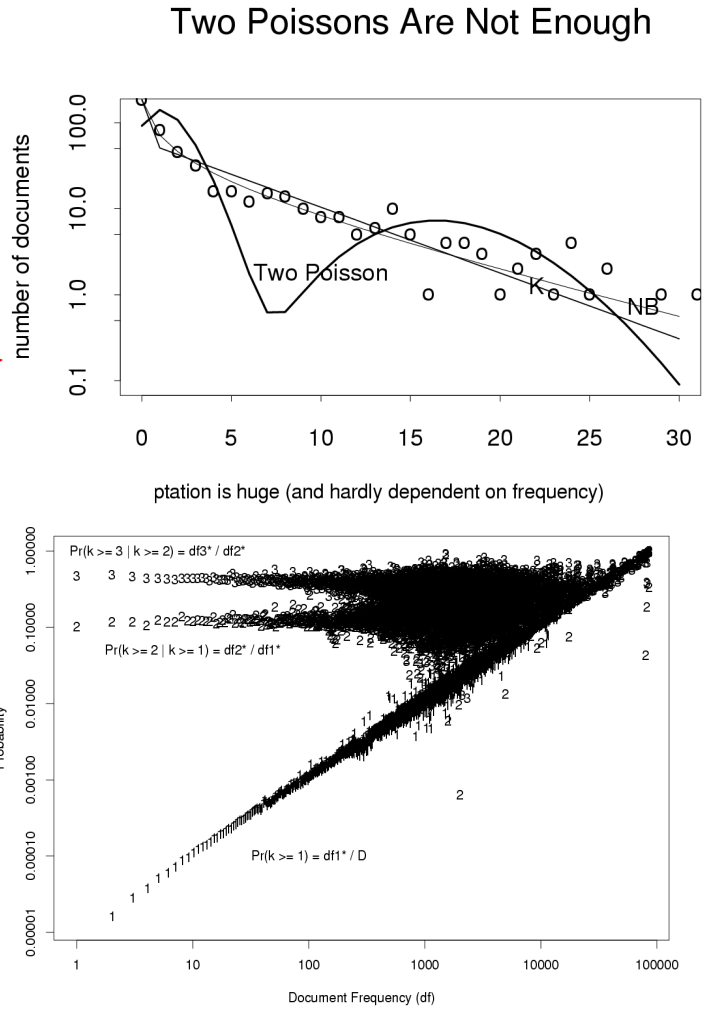
1. Cache-based adaptation

$$\Pr(w|...) = \lambda \Pr_{local}(w|...) + (1-\lambda)\Pr_{global}(w|...)$$

2. Parametric Models
   - Poisson, Two Poisson, Mixtures (neg binomial)

$$\Pr(k \geq 2 \,|\, k \geq 1) = \frac{1-\Pr(1)-\Pr(0)}{1-\Pr(0)}$$

3. Non-parametric
   - $\Pr(+\text{adapt}_1) \equiv \Pr(\text{test}|\text{hist})$
   - $\Pr(+\text{adapt}_2) \equiv \Pr(k \geq 2 | k \geq 1)$



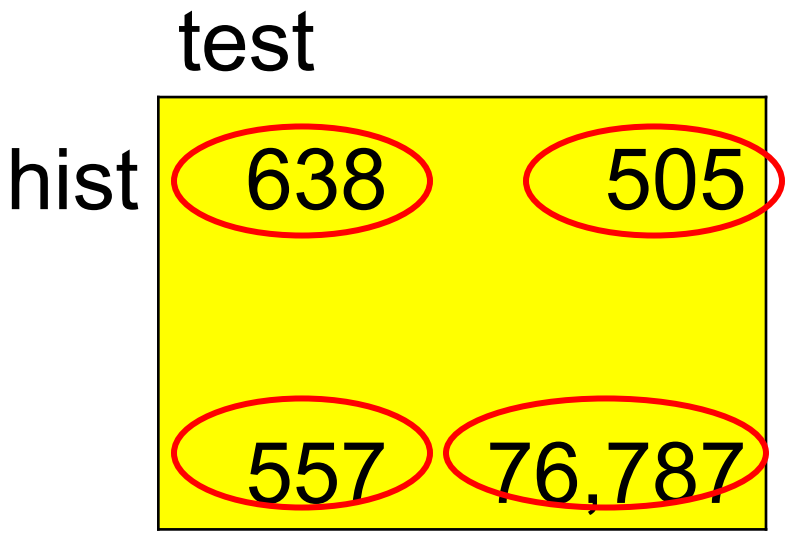
Positive & Negative Adaptation

- Adaptation:
  - How do probabilities change as we read a doc?
- Intuition: If a word *w* has been seen recently
  1. +adapt: prob of *w* (and its friends) goes way up
  2. −adapt: prob of many other words goes down a little
- Pr(+adapt) >> Pr(prior) > Pr(−adapt)

Adaptation: Method 1

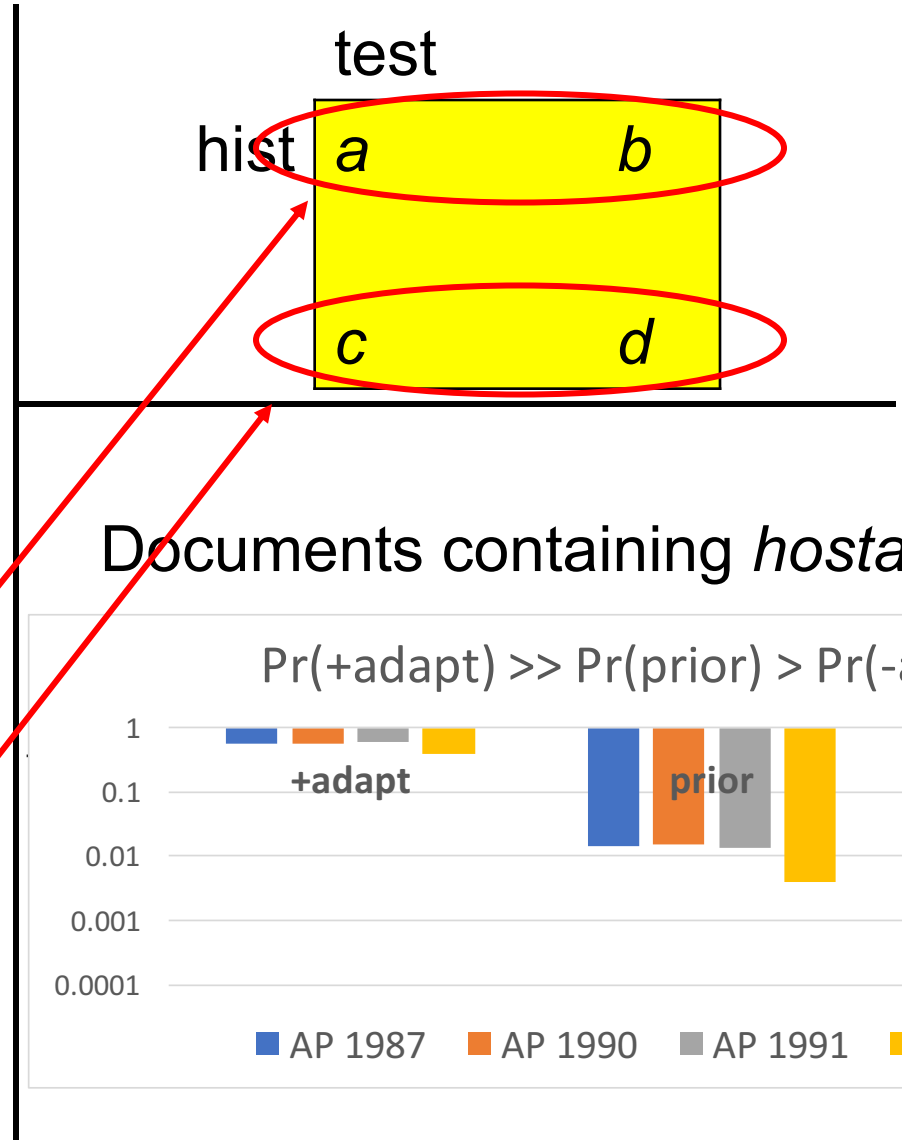
- Split each document into two equal pieces:
  - Hist: 1st half of doc
  - Test: 2nd half of doc
- Task:
  - Given hist
  - Predict test
- Compute contingency table for each word

Documents containing *hostages* in 1990 AP News



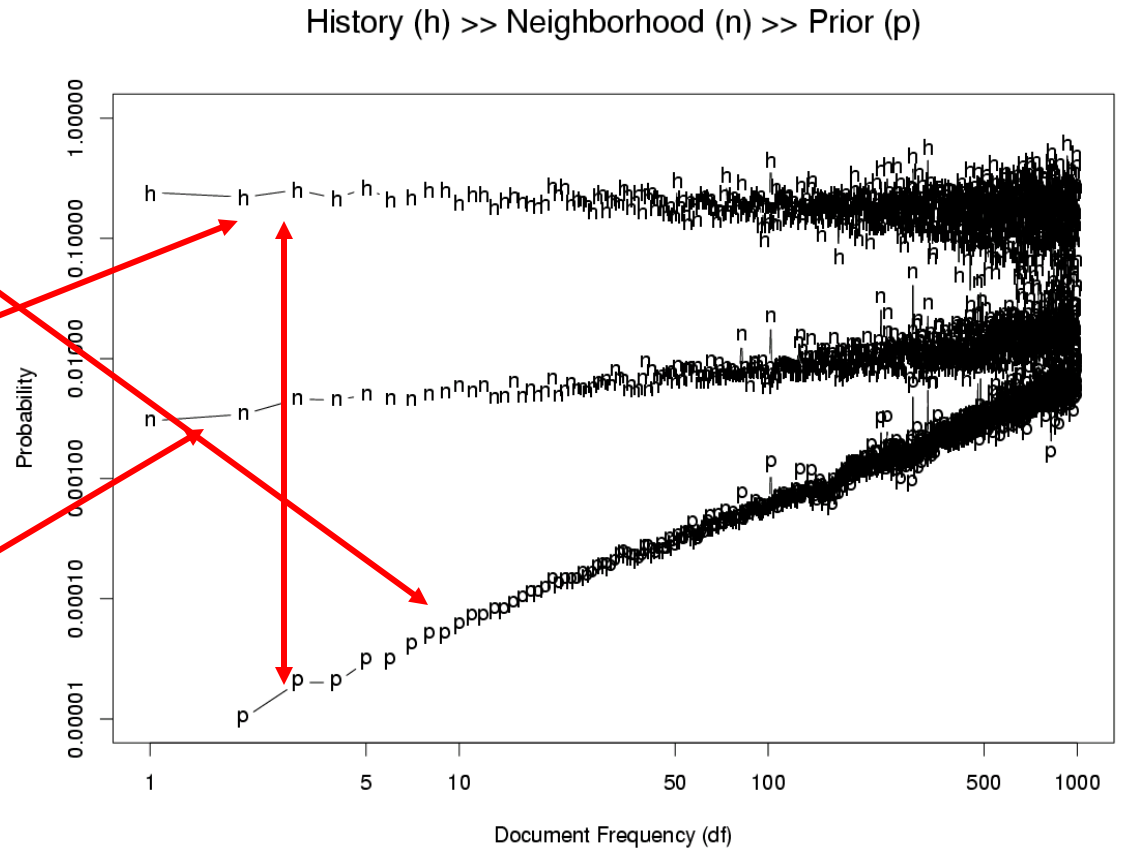
Adaptation: Method 1

- Notation
  - D = a+b+c+d (library)
  - df = a+b+c (doc freq)
- Prior: $\Pr(w \in test) = \frac{a+c}{D}$
- +adapt $\Pr(w \in test \mid w \in hist) = \frac{a}{a+b}$
- −adapt $\Pr(w \in test \mid w \notin hist) = \frac{c}{c+d}$



Adaptation: Hist >> Near >> Prior

- Magnitude is huge
- Shape: Given/new
  - 1st mention: marked
    - Surprising (low prob)
    - Depends on freq
  - 2nd: unmarked
    - Less surprising
    - Independent of freq
  - Priming:
    - “a little bit” marked

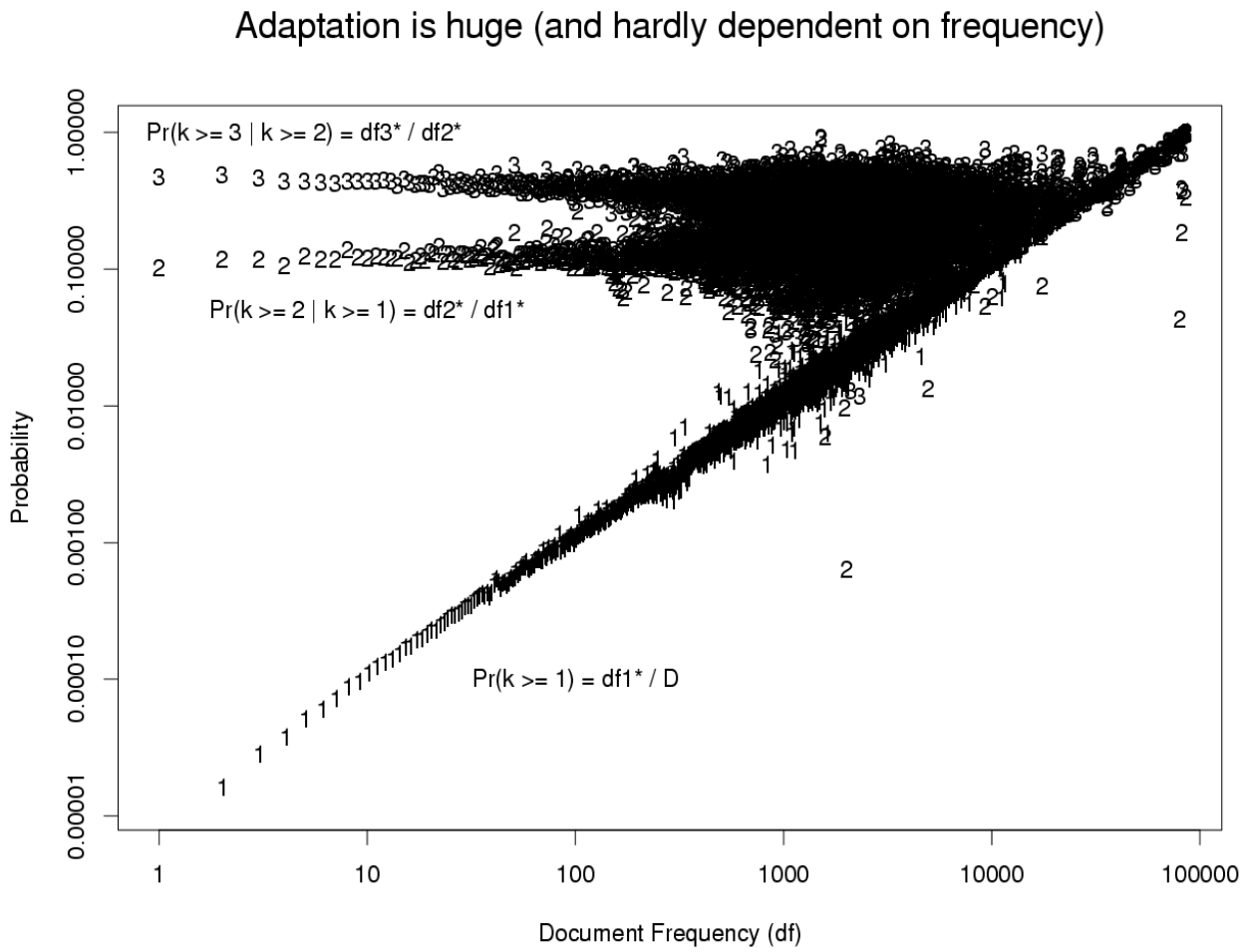


Adaptation: Method 2

- Pr(+adapt$_2$)

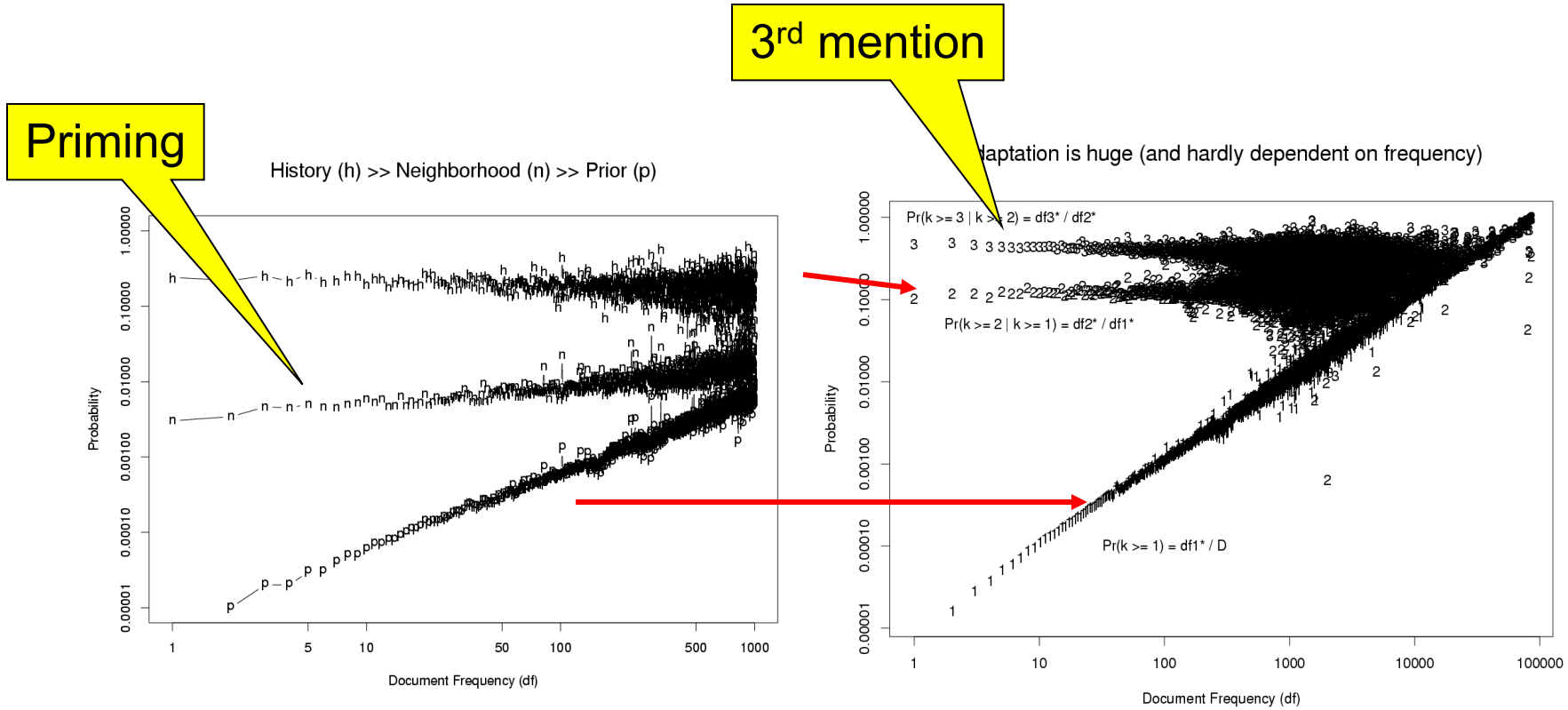
$$\Pr(k \geq 2 \mid k \geq 1) = \frac{df_2}{df_1}$$

- $df_k(w)$ ≡ number of docs that
  - mention word *w*
  - at least *k* times
- $df_1(w)$ ≡ standard def of document freq (df)



$Pr(+adapt_1) \approx Pr(+adapt_2)$

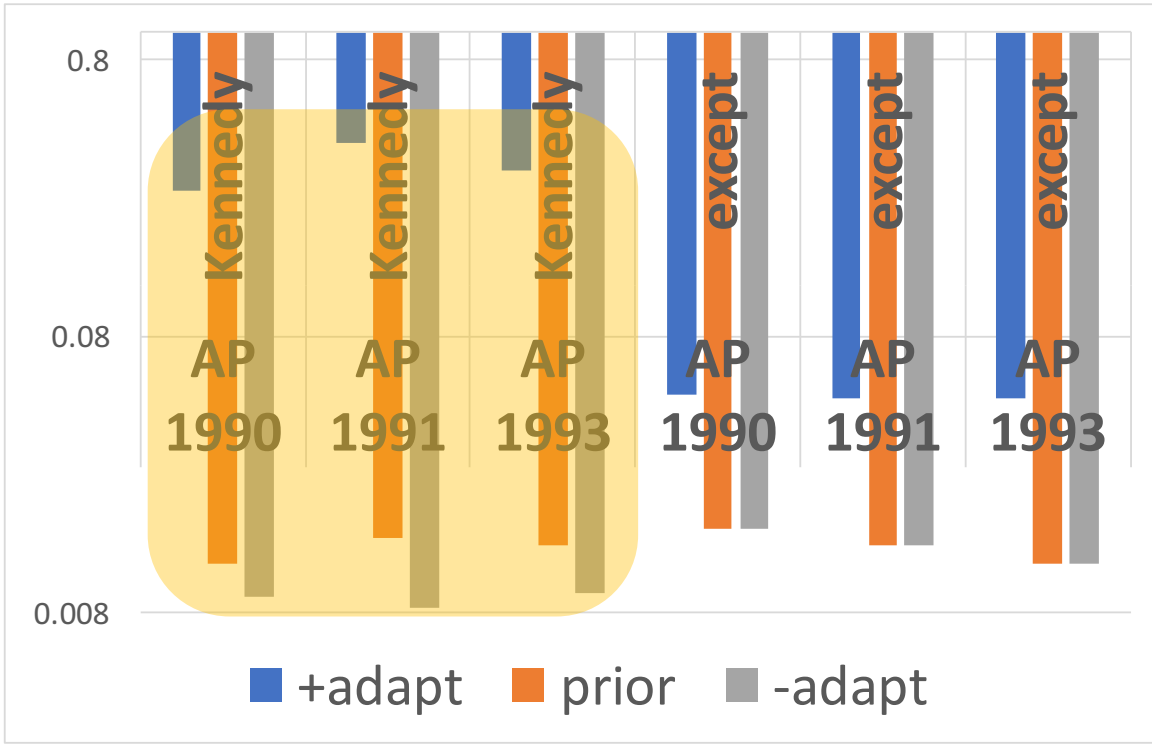
Within factors of 2-3 (as opposed to 10-1000)



Adaptation is Lexical

- Lexical: adaptation is
  - Stronger for good keywords (*Kennedy*)
  - Than random strings, function words (*except*), etc.
- Content ≠ low frequency

+adapt	prior	−adapt	source	word
0.27	0.012	0.0091	AP90	*Kennedy*
0.40	0.015	0.0084	AP91	*Kennedy*
0.32	0.014	0.0094	AP93	*Kennedy*
0.049	0.016	0.016	AP90	*except*
0.048	0.014	0.014	AP91	*except*
0.048	0.012	0.012	AP93	*except*



Adaptation Conclusions

1. Large magnitude ($p/2 >> p^2$); *big* quantity discounts
2. Distinctive shape
   - 1st mention depends on freq
   - 2nd does not
   - Priming: between 1st mention and 2nd
3. Lexical:
   – Independence assumptions aren't bad for meaningless random strings, function words, common first names, etc.
   – More adaptation for content words (good keywords, OOV)

Conclusions / Plan

- Summarize main points of paper
- Call out
  - some highlights of subsequent literature
    - Word2vec ≈ PMI (with factoring)
    - Word2vec ≈ LSA / SVD (similar factoring)
    - Are there better ways to approximate PMI? Sketches?
  - suggestions for future work
    - If the matrix that we are factoring isn't positive definite,
      - then the optimization might be looking for a solution that may not exist.
      - It might be useful to generalize the optimization to consider solutions where the left eigenvector need not be the same as the right eigenvector, or where the solution can make use of imaginary numbers.
    - It is common for embeddings to use K=300 dimensions,
      - but most words don't appear K times in the corpus.
      - It is hard to justify K parameters for a word that doesn't appear K times.

9 CONCLUSIONS

We began this paper with the psycholinguistic notion of word association norm, and extended that concept toward the information theoretic definition of mutual information. This provided a precise statistical calculation that could be applied to a very large corpus of text to produce a table of associations for tens of thousands of words. We were then able to show that the table encoded a number of very interesting patterns ranging from *doctor . . . nurse* to *save . . . from*. We finally concluded by showing how the patterns in the association ratio table might help a lexicographer organize a concordance.

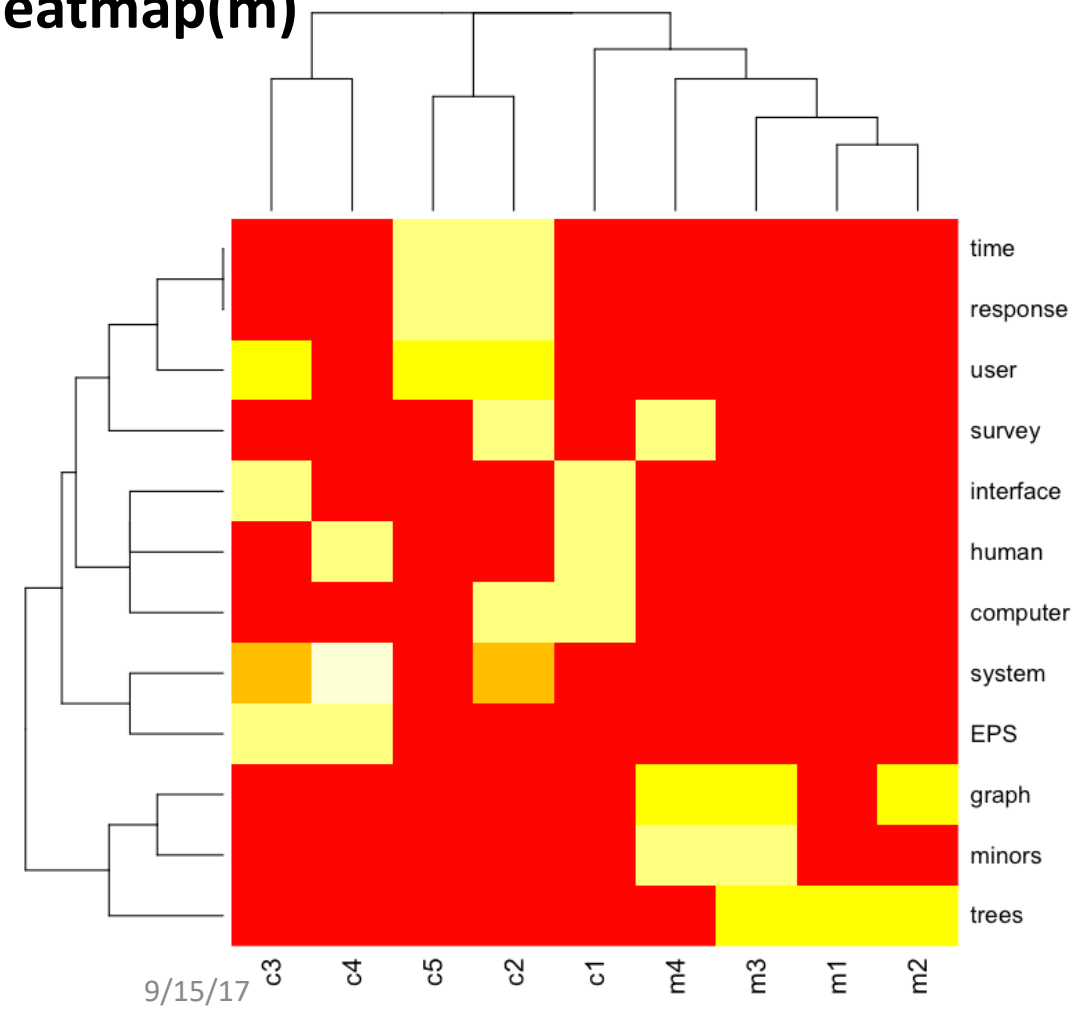
IMHO: Word2vec & GloVe >> PMI

- Glass is half full
  - Levy & Goldberg are correct in pointing out similarities
- But half empty
  - Word2vec & GloVe do better in practice
  - Because of "tricks" for small counts, etc.
  - Code is not well understood
    - Even though there isn't much there
    - 702 lines in word2vec.c; 446 in glove.c
- Dimension Reduction
  - Matrix completion (?!?)
  - Obvious suggestion, SVD on PMIs
    - Disappointing in practice
    - May not want to model PMI too well

backup

```
s=svd(m)
m2 = s$u[,1:2] %*% diag(s$d[1:2]) %*% t(s$v[,1:2])
dimnames(m2) = dimnames(m)
```

heatmap(m)



heatmap(m2)