CS6120: Lecture 3

Kenneth Church

https://kwchurch.github.io

Agenda

- Homework
  - Assignment 1: Better Together
  - Assignment 2: HuggingFace Pipelines
- Background Material
  - Python
    - numpy, matplotlib, requests, json
    - sklearn, scipy
    - requests: APIs (Semantic Scholar)
  - Linear Algebra
  - Graph Algorithms
  - Probability
  - Machine Learning
- Old Business
  - (Nearly) everything → Vector
    - Word2vec
    - Doc2vec
  - Similarity → Cosine
  - Approximate Nearest Neighbors
- New Business
  - Colab
  - Deep Nets: Inference
    - Classification & Regression
    - Anything → Vector
    - Machine Translation
    - Fill Mask

Graphs

- $G = (V, E)$
  - V: vertices (nodes)
  - E: edges
- Sizes
  - $|V| = N$
  - $|E| \leq N^2$
- Represent graph, G, as matrix, M
  - Sparse Matrices
  - scipy.sparse

Compressed sparse graph routines (`scipy.sparse.csgraph`)

Fast graph algorithms based on sparse matrix representations.

Contents

`connected_components`(csgraph[, directed, ...])	Analyze the connected components of a sparse graph
`laplacian`(csgraph[, normed, return_diag, ...])	Return the Laplacian of a directed graph.
`shortest_path`(csgraph[, method, directed, ...])	Perform a shortest-path graph search on a positive directed or undirected graph.
`dijkstra`(csgraph[, directed, indices, ...])	Dijkstra algorithm using Fibonacci Heaps
`floyd_warshall`(csgraph[, directed, ...])	Compute the shortest path lengths using the Floyd-Warshall algorithm
`bellman_ford`(csgraph[, directed, indices, ...])	Compute the shortest path lengths using the Bellman-Ford algorithm.
`johnson`(csgraph[, directed, indices, ...])	Compute the shortest path lengths using Johnson's algorithm.
`breadth_first_order`(csgraph, i_start[, ...])	Return a breadth-first ordering starting with specified node.

Graphs, Transitive Closure & Random Walks

- $G = (V, E)$
  - V: vertices (nodes)
  - E: edges
- Sizes
  - $|V| = N$
  - $|E| \leq N^2$
- Represent graph, G, as matrix, M
  - Sparse Matrices
  - scipy.sparse

- M: paths of length 1
- M^2: paths of length 2
- $M + M^2$: paths of length 1 or 2
- $\sum_{i=0}^{i=N} M^i$: paths of length 0 to N
- $\sum x^i = \frac{1}{1-x}$
- Laplacian
- Random Walks
  - M: $\Pr(w_j | w_i)$
  - M^2: paths of length 2
  - M^i: paths of length i

Relations: $R \in \{=, \neq, <\}$

- Equivalence Relations: $R \rightarrow =$
  - Reflexive:
    - $a = a$
  - Symmetric:
    - $a = b \rightarrow b = a$
  - Transitive:
    - $a = b \,\&\, b = c \rightarrow a = c$
- Partial Order: $<$
  - Transitive, but antisymmetric

- Lexical Semantics
  - Synonyms: $good = nice$
  - Antonyms: $good \neq bad$
  - is-a: $car < vehicle$
- Challenges:
  - Is symmetry desirable?
    - $\cos$ is symmetric
    - (unlike antonyms, is-a)
  - Is transitivity desirable?
- Ontologies: WordNet
  - https://www.nltk.org/howto/wordnet.html

Probability Theory

- Urn Models
- Events:
  - A corpus is a sample of a population
  - Picking the next word is like a coin toss
  - Let p be the probability of heads
    - The next word is "Kennedy"
  - Let q be the probability of tails
    - where $p + q = 1$
  - $(p+q)^n = \sum_{k=0}^{k=n} \binom{n}{k} p^k q^{n-k}$
- Binomial is one of many models
  - Binomial is related to logistic regression
  - Multinomial is related to softmax

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

Create account Log in

Contents [hide]

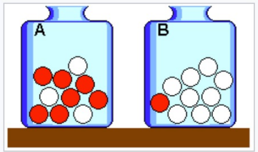
Urn problem

7 languages

Article Talk   Read Edit View history Tools

From Wikipedia, the free encyclopedia

In probability and statistics, an **urn problem** is an idealized mental exercise in which some objects of real interest (such as atoms, people, cars, etc.) are represented as colored balls in an urn or other container. One pretends to remove one or more balls from the urn; the goal is to determine the probability of drawing one color or another, or some other properties. A number of important variations are described below.



Two urns containing white and red balls.

An **urn model** is either a set of probabilities that describe events within an urn problem, or it is a probability distribution, or a family of such distributions, of random variables associated with urn problems.[1]

History [edit]

In *Ars Conjectandi* (1713), Jacob Bernoulli considered the problem of determining, given a number of pebbles drawn from an urn, the proportions of different colored pebbles within the urn. This problem was known as the *inverse probability* problem, and was a topic of research in the eighteenth century, attracting the attention of Abraham de Moivre and Thomas Bayes.

Bernoulli used the Latin word *urna*, which primarily means a clay vessel, but is also the term used in ancient Rome for a vessel of any kind for collecting ballots or lots; the present-day Italian word for ballot box is still *urna*. Bernoulli's inspiration may have been lotteries, elections, or games of chance which involved drawing balls from a container, and it has been asserted that elections in medieval and renaissance Venice, including that of the doge, often included the choice of electors by lot, using balls of different colors drawn from an urn.[2]

Basic urn model [edit]

In this basic urn model in probability theory, the urn contains x white and y black balls, well-mixed together. One ball is drawn randomly from the urn and its color observed; it is then placed back in the urn (or not), and the selection process is repeated.[3]

Possible questions that can be answered in this model are:

- Can I infer the proportion of white and black balls from n observations? With what degree of confidence?
- Knowing x and y, what is the probability of drawing a specific sequence (e.g. one white followed by one black)?
- If I only observe n balls, how sure can I be that there are no black balls? (A variation both on the first and the second question)

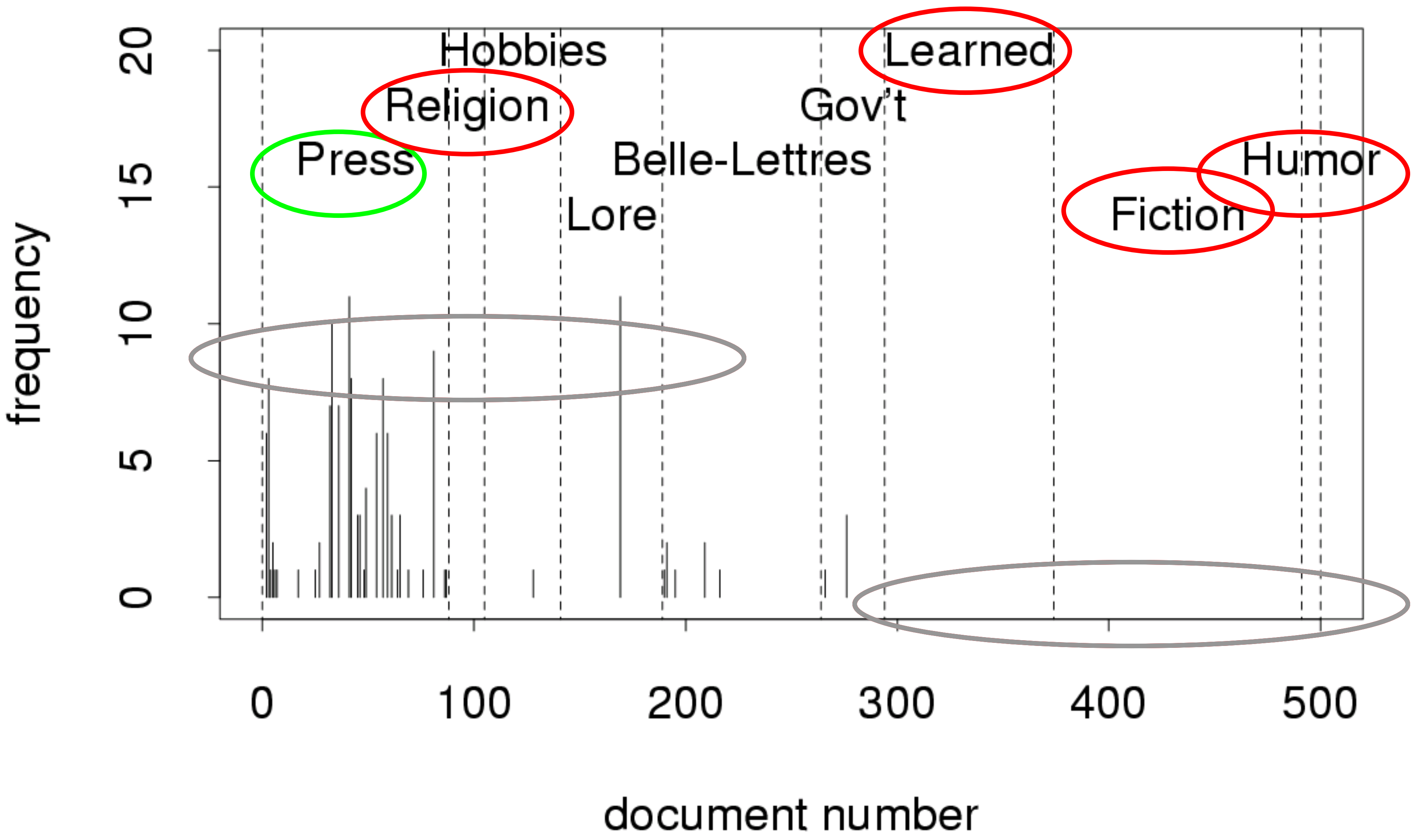
Statistics:
Combining models with observations

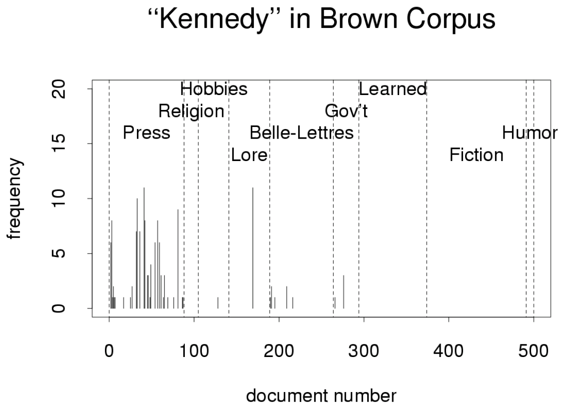
- Models (from Probability)
  - Binomial
  - Multinomial
  - Normal
  - Poisson
  - Exponential
- Observations
  - Corpora
  - Data tables
- Assumptions:
  - IID: https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables
- Example:
  - What is the probability of finding
    - exactly k instances of "Kennedy"
    - in a sample of n words?
  - Model: binomial
    - $\Pr(k) = \binom{n}{k} p^k q^{n-k}$
  - Observations: Brown Corpus
    - Sample size: $N = 1$M words
    - freq("Kennedy") $= 104$
  - Fitting the model
    - $p = 104/10^6 \approx 10^{-4}$
    - $q = 1 - p$
  - Challenge(s): IID assumption

Interestingness Metrics: Deviations from Independence

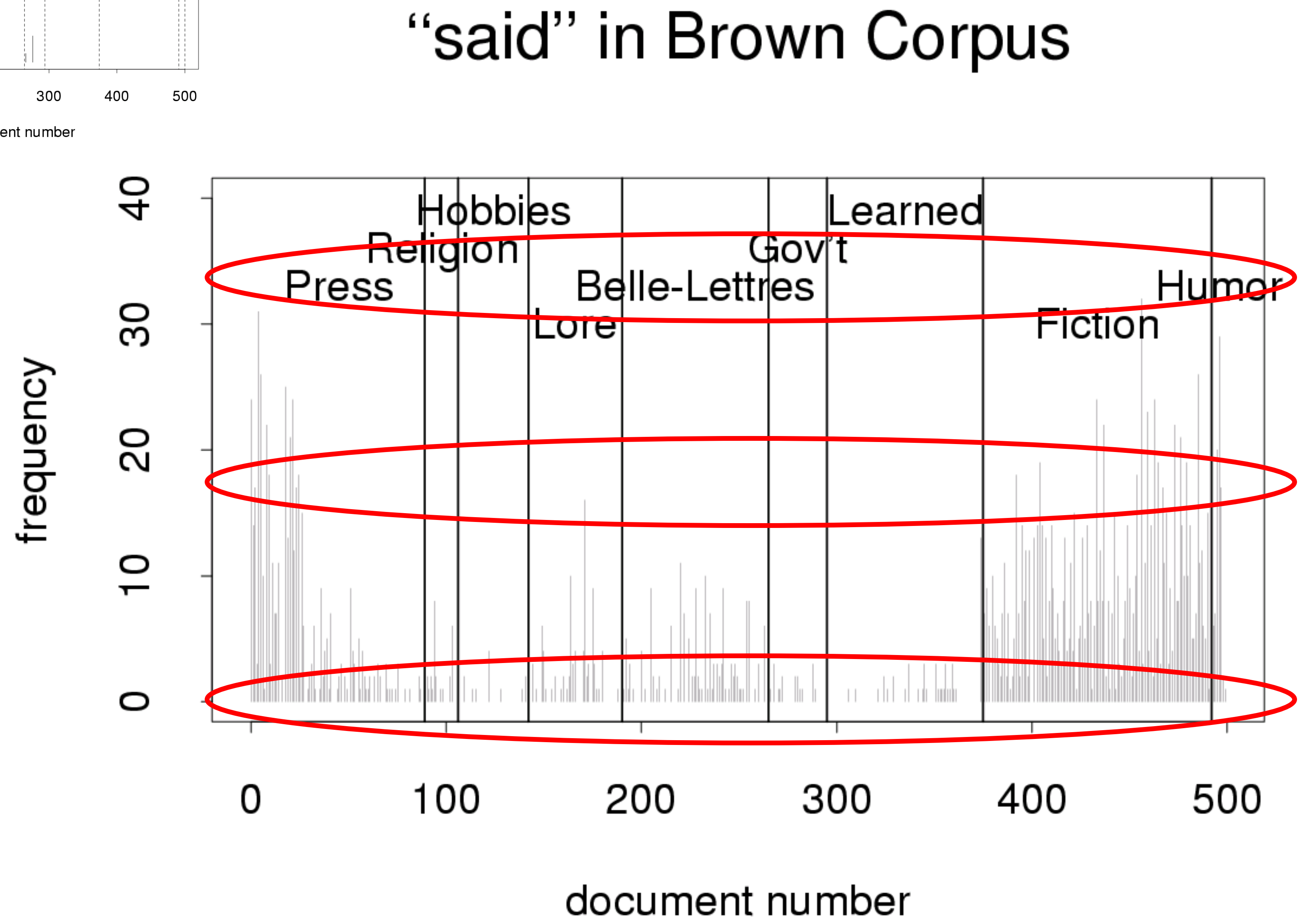
- Poisson (and other independence assumptions)
  - Not bad for meaningless random strings
- Deviations from Poisson are clues for hidden variables
  - Meaning, content, genre, topic, author, etc.
- Analogous to pointwise mutual information (Hanks)
  - Pr(doctor…nurse) >> Pr(doctor) Pr(nurse)

"Kennedy" in Brown Corpus



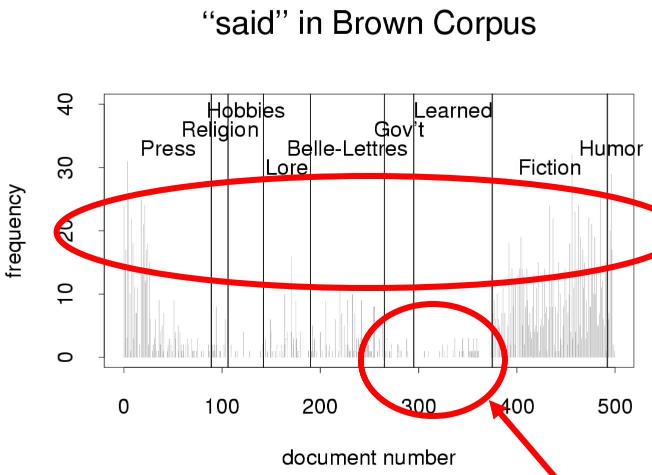


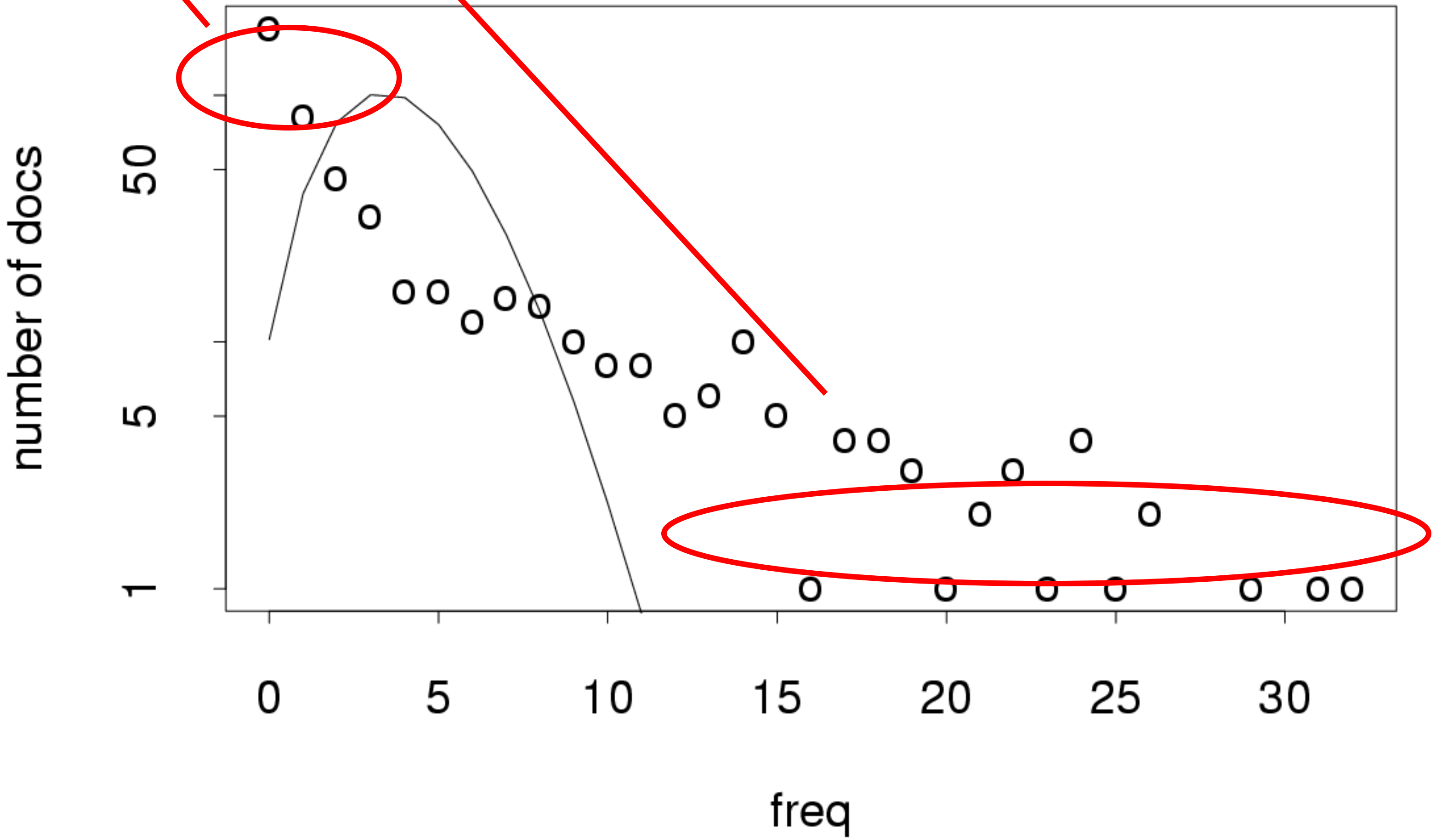
“Kennedy” in Brown Corpus
frequency
document number
Press
Religion
Hobbies
Lore
Belle-Lettres
Gov’t
Learned
Fiction
Humor
0
5
10
15
20
0
100
200
300
400
500



“said” in Brown Corpus
frequency
document number
Press
Religion
Hobbies
Lore
Belle-Lettres
Gov’t
Learned
Fiction
Humor
0
10
20
30
40
0
100
200
300
400
500

Poisson Doesn't Fit

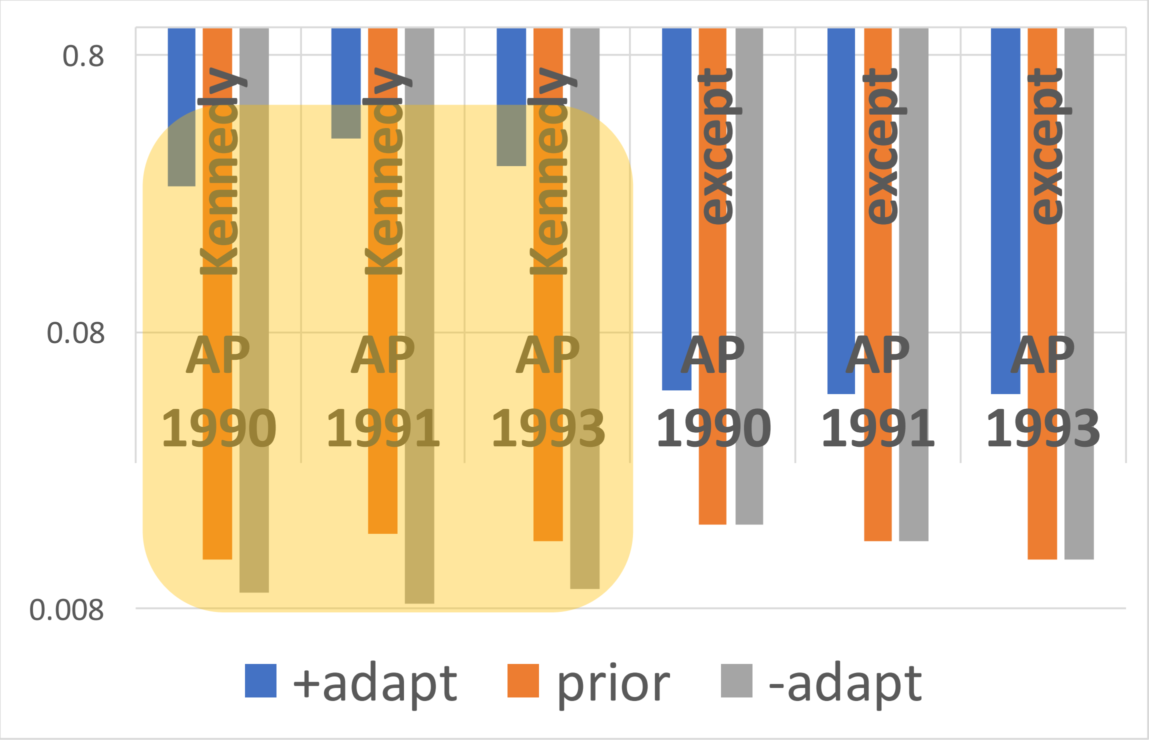




Adaptation is Lexical

- 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)

Word Association Norms, Mutual Information and Lexicography

Title | Add | More | 1–20 | Cited by

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

1555
1454
761
654
602
496

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*

Computational Linguistics Volume 16, Number 1, March 1990

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*

Computational Linguistics Volume 16, Number 1, March 1990

Counts are growing 1000x per decade (same as disks)

Rising Tide of Data Lifts All Boats

The Quote

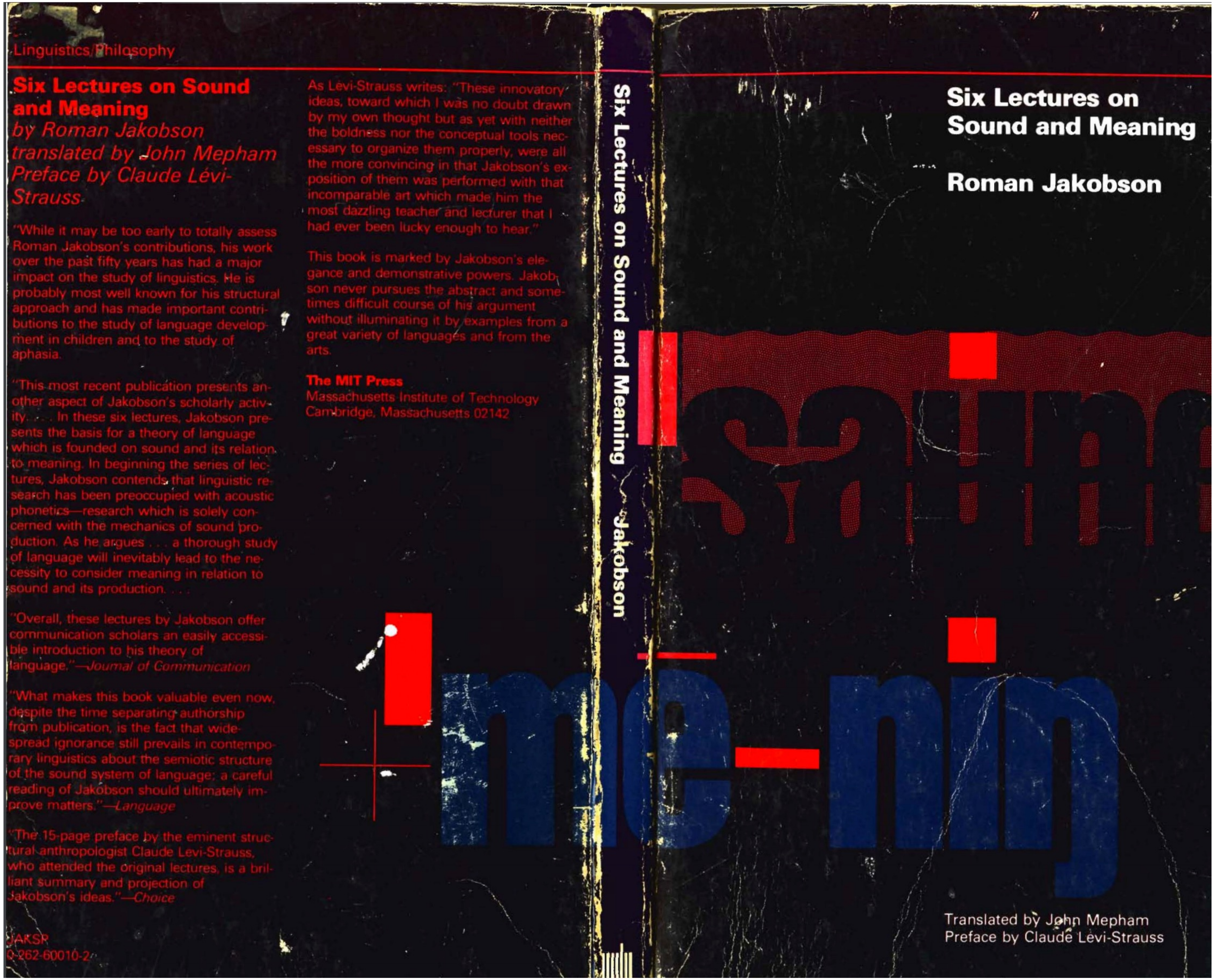
“Whenever I fire a linguist our system performance improves”

From my talk entitled:

Applying Information Theoretic Methods: Evaluation of Grammar Quality

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

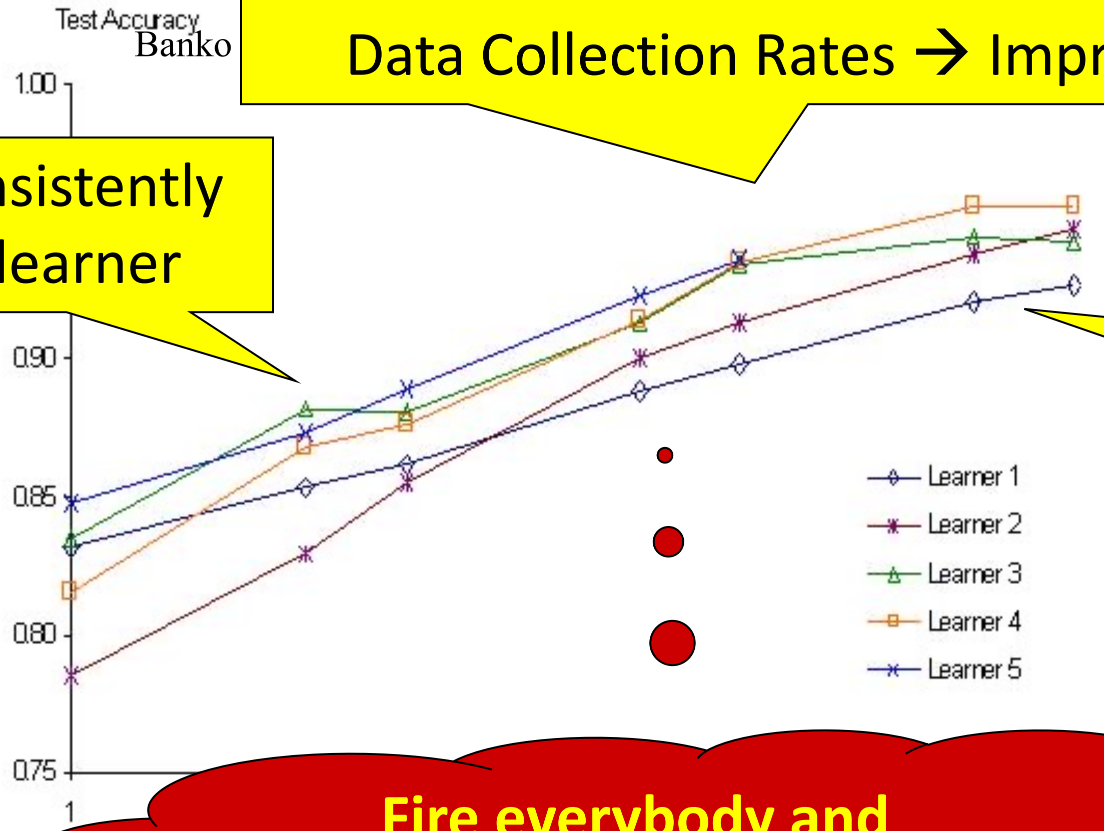
http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf


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



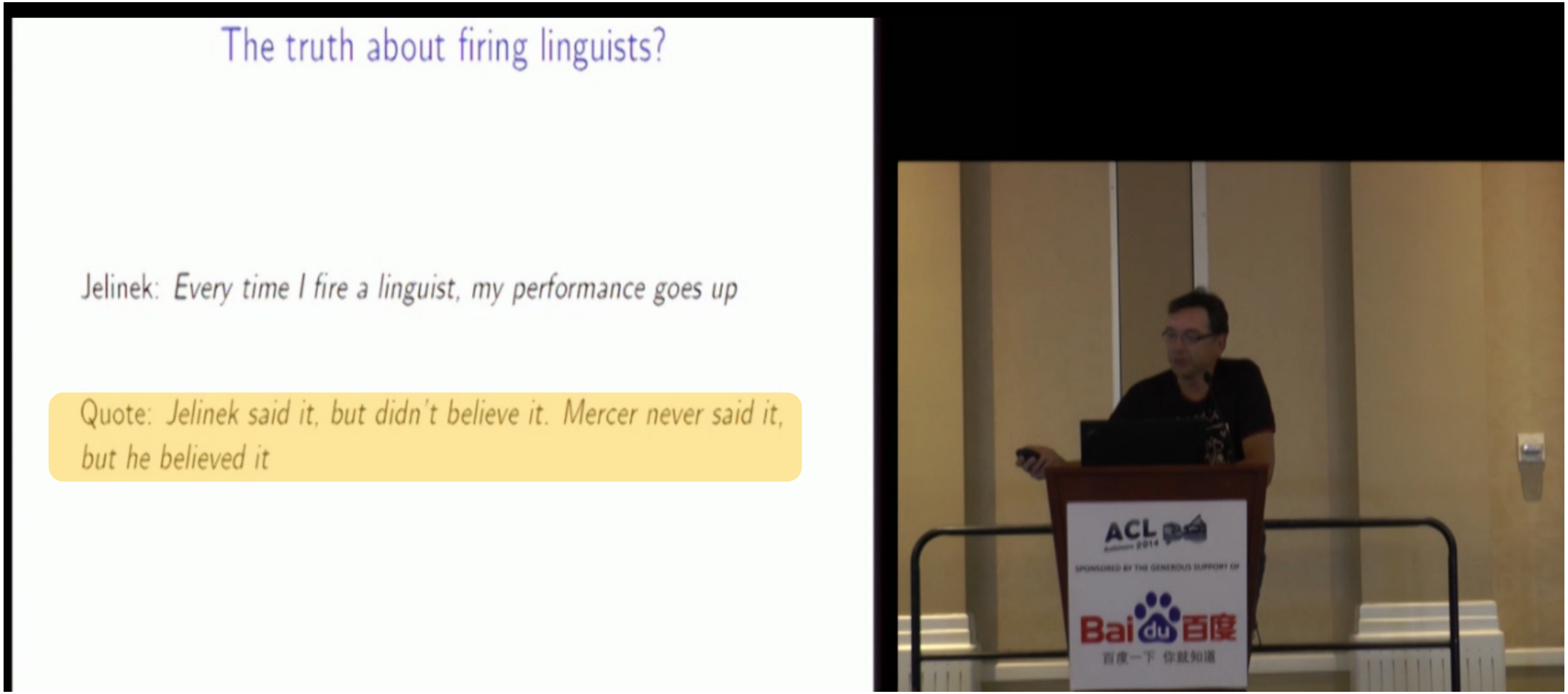
There is no data like more data

Fire everybody and spend the money on data

Quoted out of context

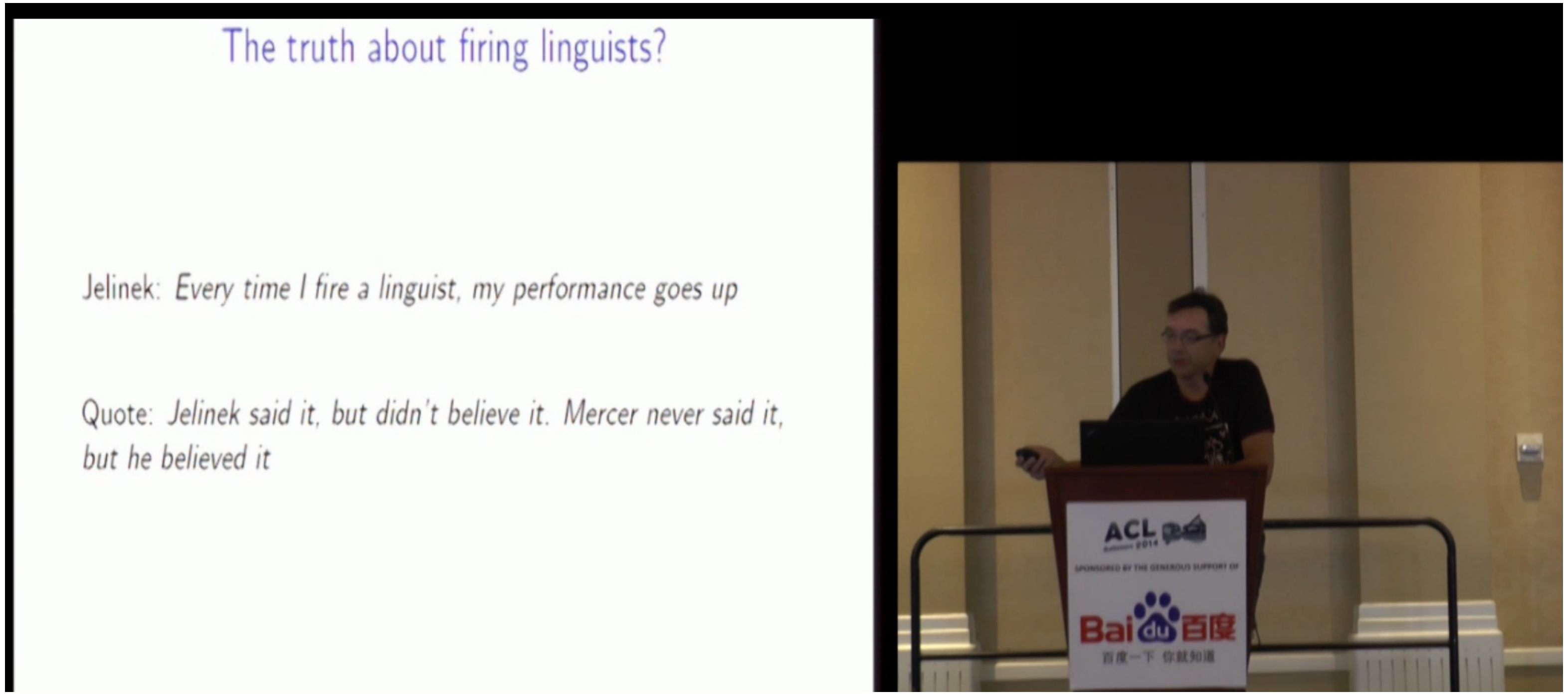
Robert Mercer
ACL Lifetime Achievement

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



Computational Linguistics:

Interdisciplinary Combination of Engineering and Humanities



The Case for Empiricism (With and Without Statistics)

Kenneth Church
1101 Kitchawan Road
Yorktown Heights, NY 10589
USA
Kenneth.Ward.Church@gmail.com

Abstract

These days we tend to use terms like *empirical* and *statistical* as if they are interchangeable, but it wasn't always this way, and probably for good reason. In *A Pendulum Swung Too Far* (Church, 2011), I argued that graduate programs should make room for both Empiricism and Rationalism. We don't know which trends will dominate the field tomorrow, but it is a good bet that it won't be what's hot today. We should prepare the next generation of students for all possible futures, or at least all probable futures. This paper argues for a diverse interpretation of Empiricism, one that makes room for everything from Humanities to Engineering (and then some).

Figure 1: Lily Wong Fillmore (standing) and Charles (Chuck) Fillmore

1 Lifetime Achievement Award (LTA)

Since the purpose of this workshop is to celebrate Charles (Chuck) Fillmore, I would like to take this opportunity to summarize some of the points that I made in my introduction to Chuck's LTA talk at ACL-2012.

I had the rather unusual opportunity to see his talk (a few times) before writing my introduction because Chuck video-taped his talk in advance.[1] I knew that he was unable to make the trip, but I had not appreciated just how serious the situation was. I found out well after the fact that the LTA meant a lot to him, so much so that he postponed an operation that he probably shouldn't have postponed (over his doctor's objection), so that he would be able to answer live questions via Skype after the showing of his video tape.

I started my introduction by crediting Lily Wong Fillmore, who understood just how much Chuck wanted to be with us in Korea, but also, just how impossible that was. Let me take this opportunity to thank her once again for her contributions to the video (technical lighting, editing, encouragement and so much more).

For many of us in my generation, C4C, Chuck's "The Case for Case" (Fillmore, 1968) was the introduction to a world beyond Rationalism and Chomsky. This was especially the case for me, since I was studying at MIT, where we learned many things (but not Empiricism).

After watching Chuck's video remarks, I was struck by just how nice he was. He had nice things to say about everyone from Noam Chomsky to Roger Schank. But I was also struck by just how difficult it was for Chuck to explain how important C4C was (or even what it said and why it mattered). To make sure that the international audience wasn't misled by his upbringing and his self-deprecating humor, I showed a page of "Minnesota Nice" stereotypes, while reminding the audience that stereotypes aren't nice, but as stereotypes go, these stereotypes are about as nice as they get.

[1] The video is available online at https://framenet.icsi.berkeley.edu/fndrupal/node/5489.

Priming & Word Associations

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

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;

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*

TABLE 1

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

Type of stimulus pair: Top string	Type of stimulus pair: Bottom string	Correct response	Proportion of trials	Mean RT (msec.)	Mean % errors
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*.

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

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] ,

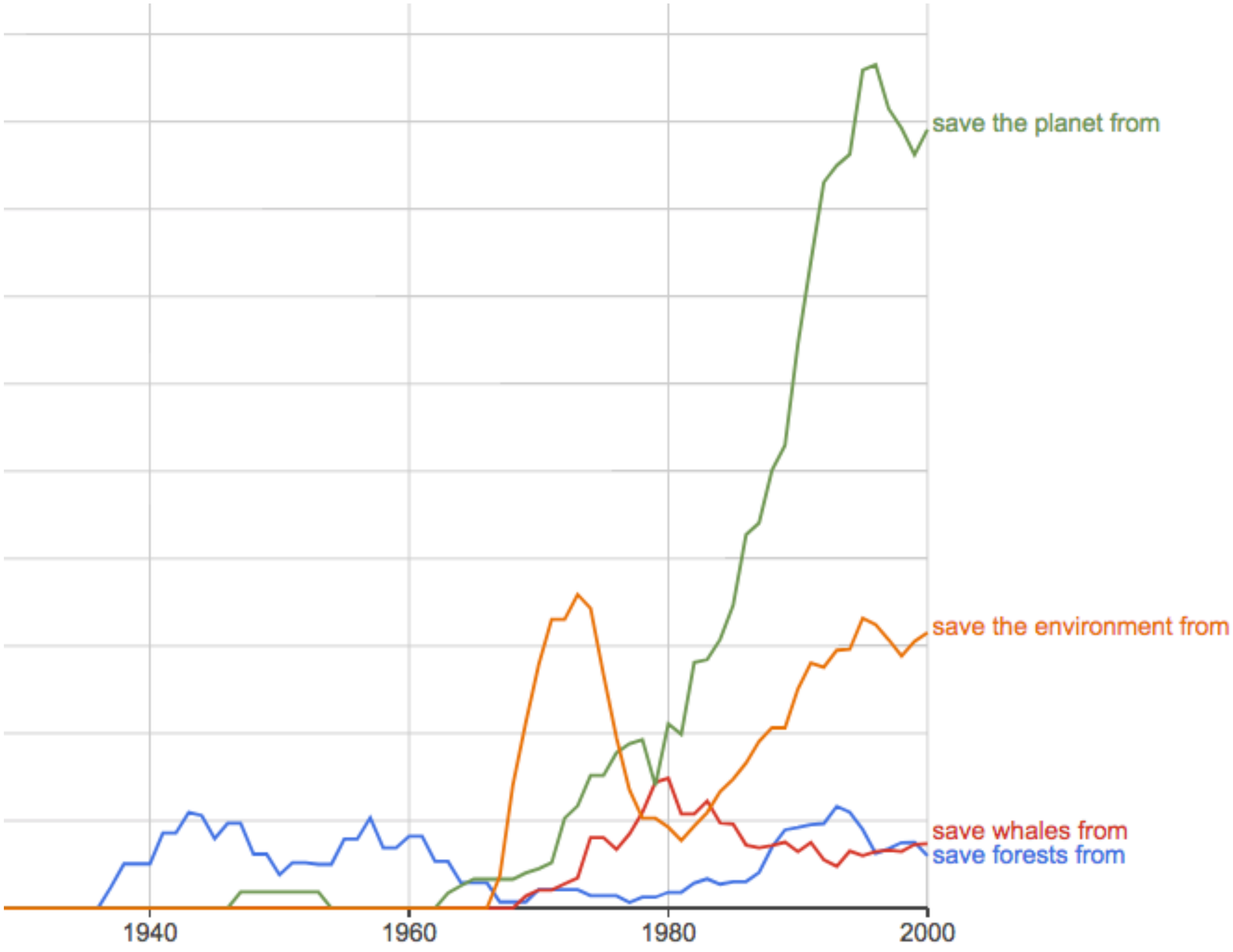
Save good shoppers from their evil $$

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

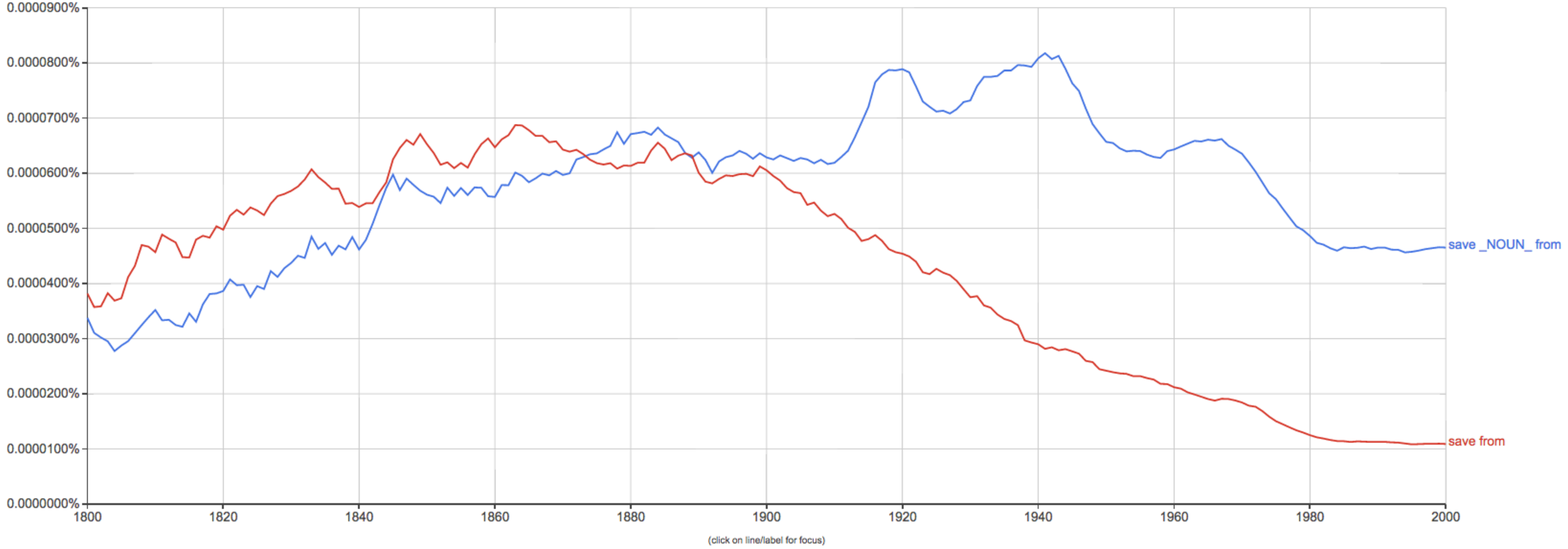
Patrick wanted me to "fix" my bug

Google Ngrams

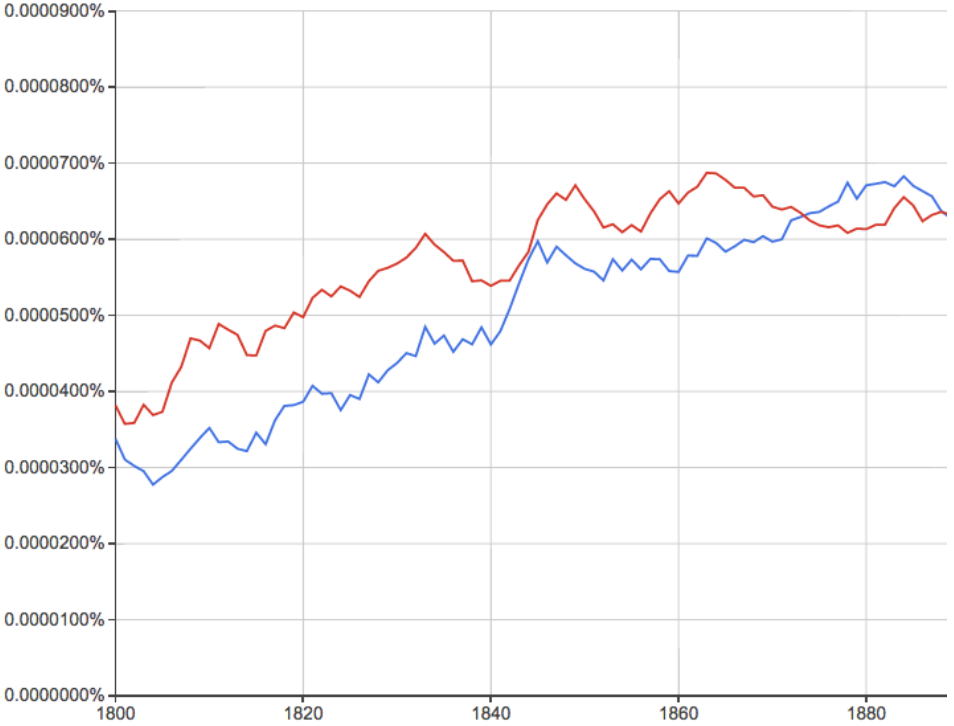


https://books.google.com/ngrams/graph?content=save+forests+from%2Csave+whales+from%2Csave+the+planet+from%2Csave+the+environment+from&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Csave%20forests%20from%3B%2Cc0%3B.t1%3B%2Csave%20whales%20from%3B%2Cc0%3B.t1%3B%2Csave%20the%20planet%20from%3B%2Cc0%3B.t1%3B%2Csave%20the%20environment%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

save from ≠ save X from Y



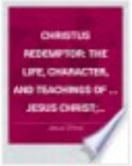
save from ≠ save X from Y



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.

Agenda

- Homework
  - Assignment 1: [Better Together]
  - Assignment 2: [HuggingFace Pipelines]
- Background Material
  - Python
    - numpy, matplotlib, requests, json
    - sklearn, scipy
    - requests: APIs (Semantic Scholar)
  - ➢ **Linear Algebra**
  - Graph Algorithms
  - Probability
  - Machine Learning
- Old Business
  - (Nearly) everything → Vector
    - Word2vec
    - Doc2vec
  - Similarity → Cosine
  - Approximate Nearest Neighbors
- New Business
  - [Colab]
  - Deep Nets: Inference
    - Classification & Regression
    - Anything → Vector
    - Machine Translation
    - Fill Mask

Linear Algebra

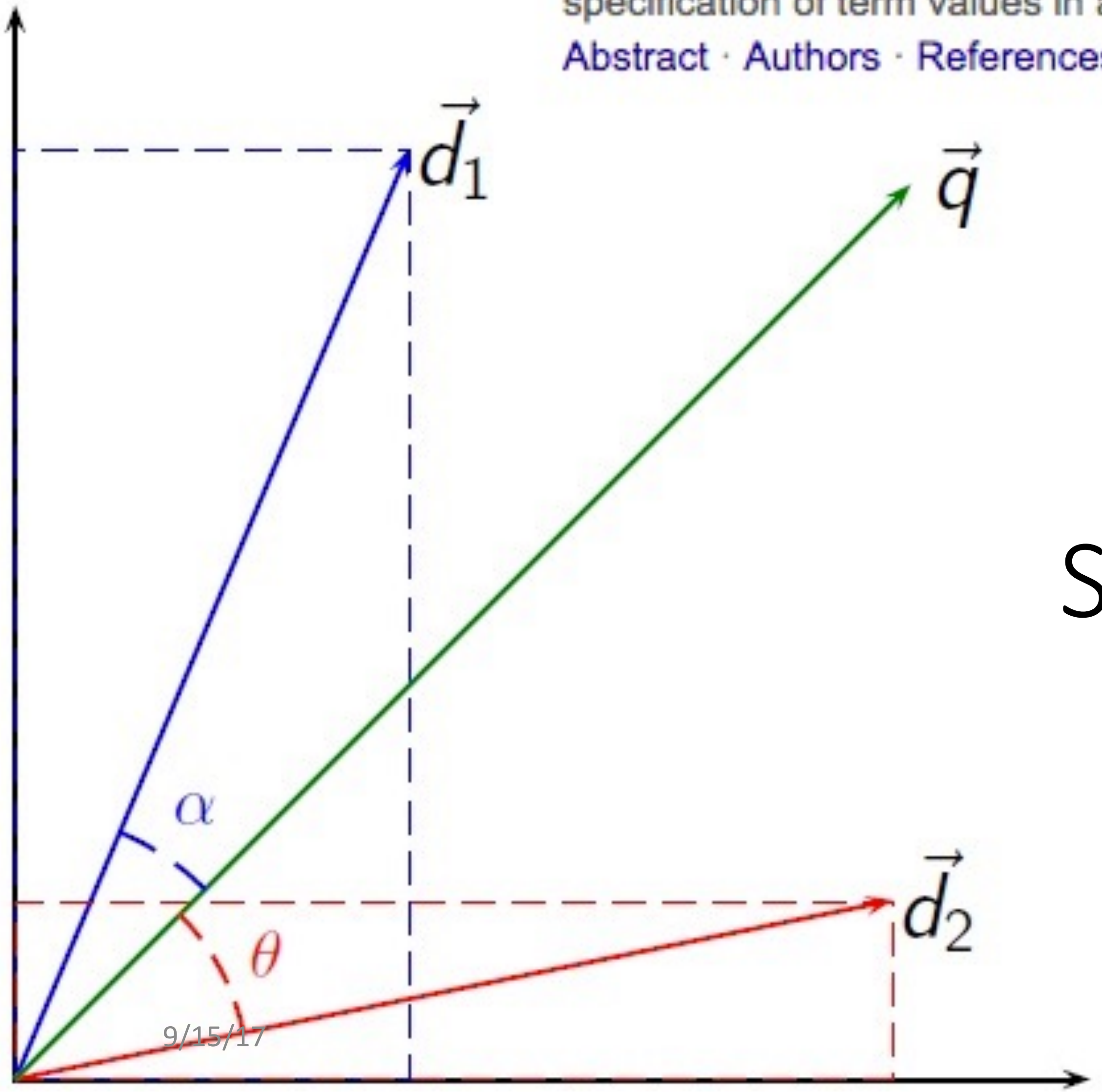
- Singular Value Decomposition (SVD)
- Principal Component Analysis (PCA)
- Dimension Reduction
- Rotations
- Approximate Nearest Neighbors (ANN)

Nearly Everything To Vectors (Embeddings)

- "Everything"
  - Words (Terms): word2vec
  - Documents (Text Strings):
    - doc2vec, BERT, Specter
  - Graphs (GNNs)
    - Example: citation graph
  - Semantics ("Meaning")
  - All the world's languages
  - Audio (Speech, Music)
  - Pictures and Videos

- Embeddings
  - Similarity $\approx$ Cosine
    - Similar documents
      - Word Overlap
      - Nearby in citation graph
      - Similar topics, venues, authors
  - Latent (Hidden) Dimensions
- Computational Convenience
  - Dimension Reduction
  - Rotations
  - Approximate Nearest Neighbors

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

Word2vec (Embeddings)

Slide from JM3

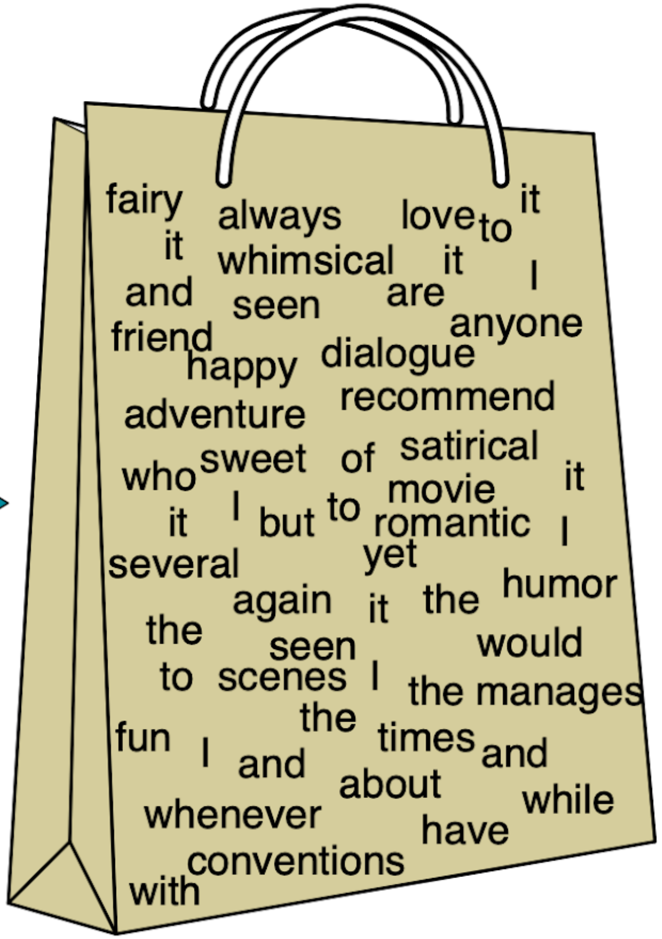
- $M \in \mathbb{R}^{V \times K}$ (tall-skinny matrix)
  - V: vocabulary size (≈ 500k)
  - K: hidden dimensions (≈ 300)
- $MM^T = cos(w_i, w_j) \propto PMI(w_i, w_j)$
  - Similarity of all pairs of words in V
  - It might be infeasible to materialize MM^T
    - But there are approximations (ANNs)
    - that find many/most of the large values
- Better for capturing collocations
  - Collocations: w_i & w_j appear near one another (more than chance)
- Less appropriate for other notions of similarity
  - Both synonyms and antonyms appear near one another
  - (But they don't mean the same thing)

For plotting purposes,

- use dimension reduction
- to reduce K down to 2D

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

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

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

- Interpretation:
  - Entropy: $H = -\log(P)$
  - where $P = \Pr(t \in d)^{count(t,d)}$

- 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

Bellcore Example

- Example of term by document matrix
  - A document ≈ a bag of words
  - A word ≈ a bag of documents
    - *You shall know a word by the company it keeps*
- Example of SVD for dimension reduction
  - Suggestion: reducing dimensions → better separation of classes of interest
- Motivate latent dimensions
  - as a method to embed both terms and documents
  - into a common (unified) vector space

Bellcore’s Example: Bag of Words + SVD

http://wordvec.colorado.edu/papers/Deerwester_1990.pdf

c1	Human machine *interface* for Lab ABC *computer* applications
c2	A *survey* of *user* opinion of computer *system response time*
c3	The EPS *user interface* management *system*
c4	*System* and *human system* engineering testing of EPS
c5	Relation of *user*-perceived *response time* to error measurement
m1	The generation of random, binary, unordered *trees*
m2	The intersection *graph* of paths in *trees*
m3	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4	*Graph minors*: A *survey*

Term by Documents Matrix

c1	Human machine interface for Lab ABC computer applications
c2	A survey of user opinion of computer system response time
c3	The EPS user interface management system
c4	System and human system engineering testing of EPS
c5	Relation of user-perceived response time to error measurement
m1	The generation of random, binary, unordered trees
m2	The intersection graph of paths in trees
m3	Graph minors IV: Widths of trees and well-quasi-ordering
m4	Graph minors: A survey

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

Term by Document Matrix

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

Singular Value Decomposition (SVD)

- $M \approx U\, D\, V^T$
- D is diagonal
  - Eigenvalues
  - Sorted from largest to smallest
- U and V are Eigenvectors
  - Orthogonal and unit length
    - $U^T U = I$
    - $V^T V = I$

- $\cos(M, M) = MM^T$
  - $U\, D\, V^T (U\, D\, V^T)^T$
  - $U\, D\, V^T (V\, D\, U^T)$
  - $UD^2\, U^T$
- $M \rightarrow UD$
  - Plus dimension reduction
  - Replace smaller Eigenvalues with 0

Dimension Reduction

- Standard Recipe
  - Set smaller Eigenvalues to 0
- Interpretation
  - L2 optimality (least squares)
- Recall that Eigenvalues are sorted from largest to smallest

- Motivation for dimension reduction
  - Computational resources:
    - Space
      - Specter: $M \in \mathbb{R}^{N \times K}$
      - N is 200M documents
      - K is 768 (BERT hidden layer)
      - $MM^T \in \mathbb{R}^{N \times N}$ (***very*** large)
    - Time
  - Statistical convenience:
    - Smoothing (soft thesaurus)
    - Replace zeros with small values
  - Computational convenience:
    - Approximate nearest neighbors
    - https://pypi.org/project/annoy/

SVD and PCA

SVD (Singular Value Decomposition)

- $M \approx U\, D\, V^T$
- D: Eigenvalues
- U: Eigenvectors
- M need not be square
  - (just non-singular)

PCA (Principal Componenet Analysis)

- $Q \propto X^T X = W \Lambda W^T$
- Q is square by construction
  - Λ: Eigenvalues
  - W: Covariances
    - Diagonal of W are variances

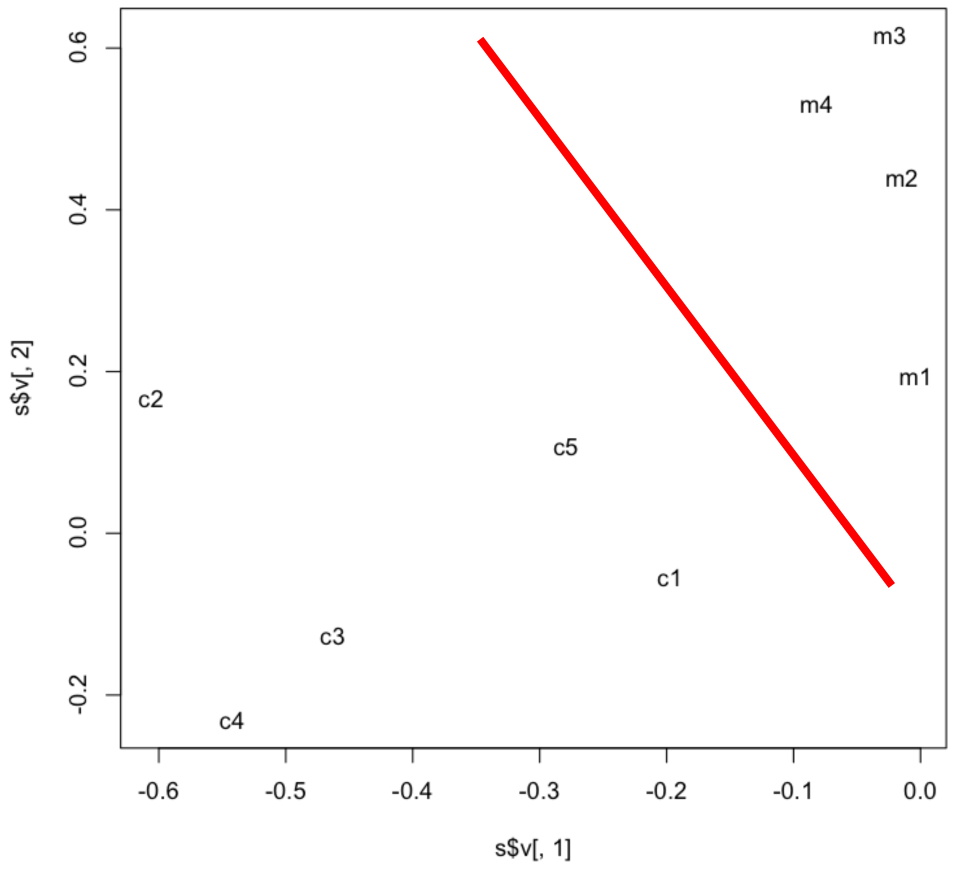
Dimension Reduction in R

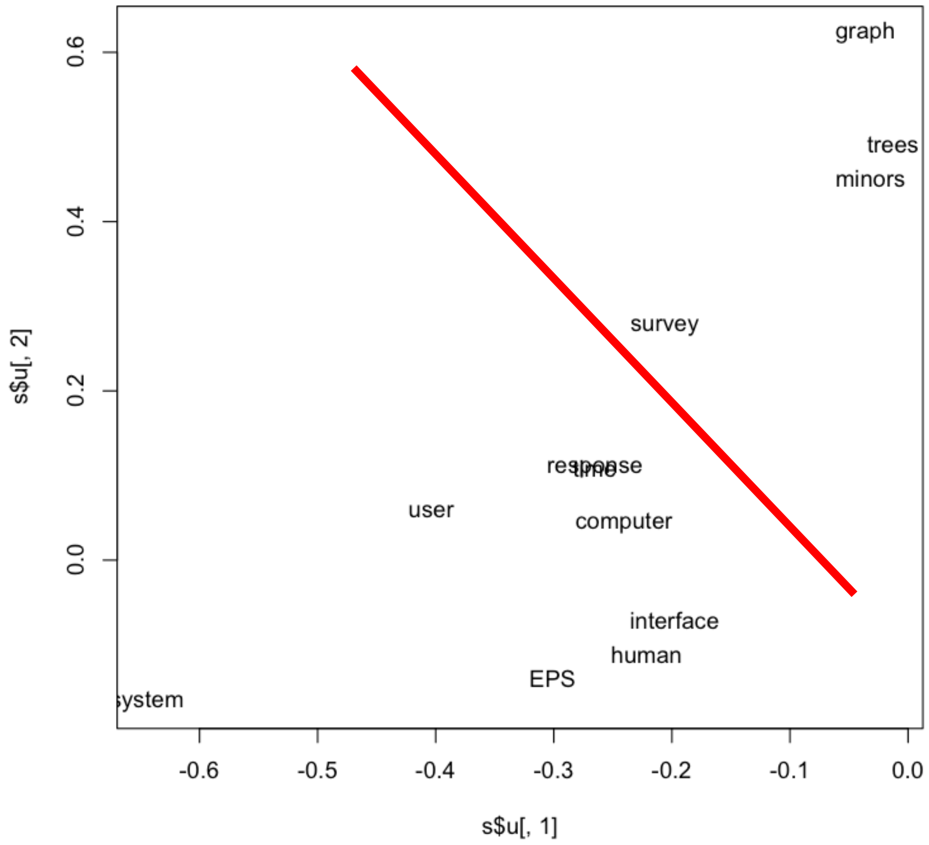
$$bellcore \approx U\ D\ V^T$$

```
bellcore =
structure(.Data = c(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0,
    0, 1, 0, 0, 0, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,
    0, 0, 0, 0, 0, 0, 1, 0, 1, 1),
.Dim = c(12, 9),
.Dimnames = list(c("human", "interface", "computer",
    "user", "system", "response", "time", "EPS",
    "survey", "trees", "graph", "minors"),
c("c1", "c2", "c3", "c4", "c5", "m1", "m2", "m3",
    "m4")))
```

```
b = svd(bellcore)
b2 = b$u[,1:2] %*% diag(b$d[1:2]) %*%
   t(b$v[,1:2])
dimnames(b2) = dimnames(bellcore)
par(mfrow=c(2,2))
plot(hclust(as.dist(-cor(bellcore))))
plot(hclust(as.dist(-cor(t(bellcore)))))
plot(hclust(as.dist(-cor(b2))))
plot(hclust(as.dist(-cor(t(b2)))))
```

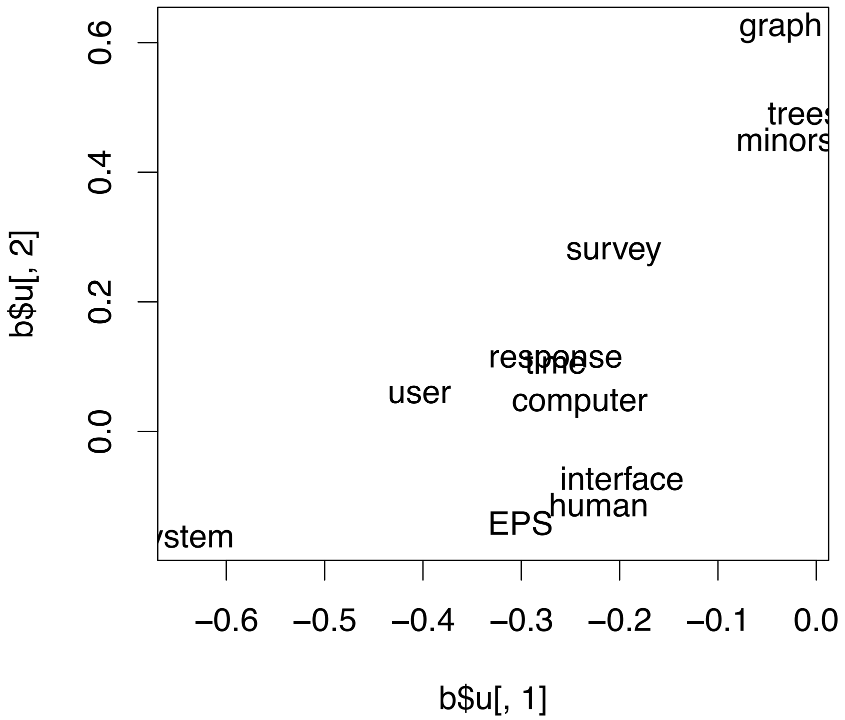
SVD maps terms & docs into internal dimensions



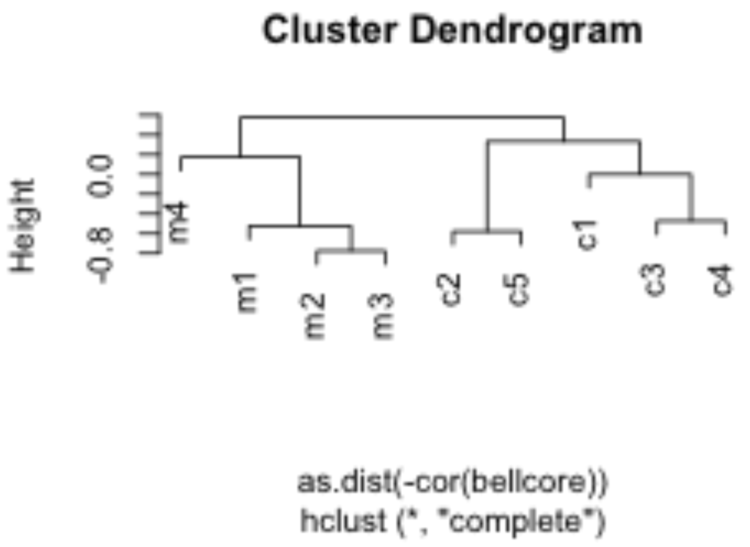


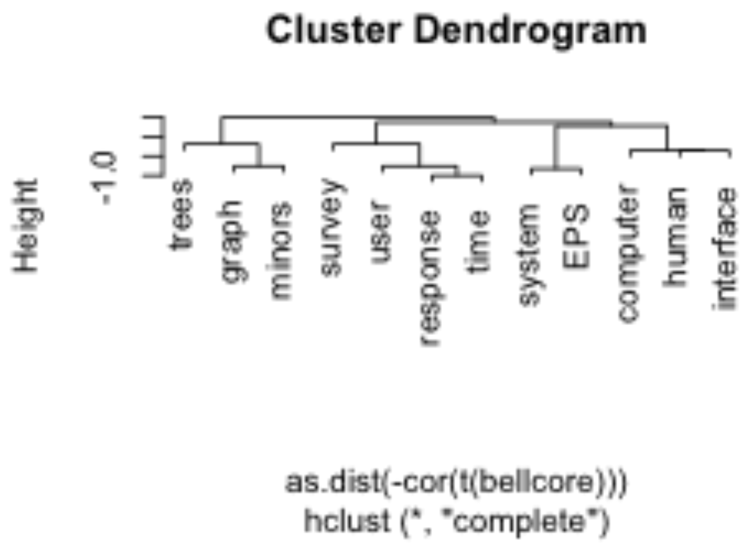
$$bellcore \approx U\ D\ V^T$$

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



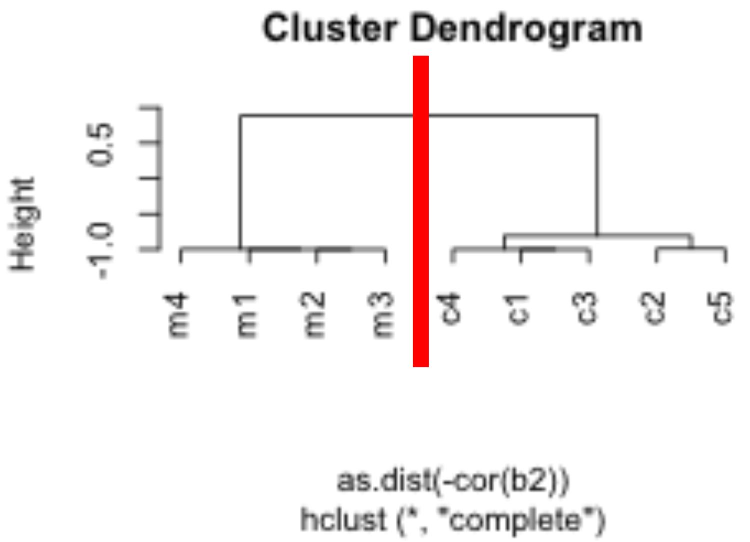
```
b = svd(bellcore)
b2 = b$u[,1:2] %*% diag(b$d[1:2]) %*% t(b$v[,1:2])
dimnames(b2) = dimnames(bellcore)
par(mfrow=c(2,2))
plot(hclust(as.dist(-cor(bellcore))))
plot(hclust(as.dist(-cor(t(bellcore)))))
plot(hclust(as.dist(-cor(b2))))
plot(hclust(as.dist(-cor(t(b2)))))
```

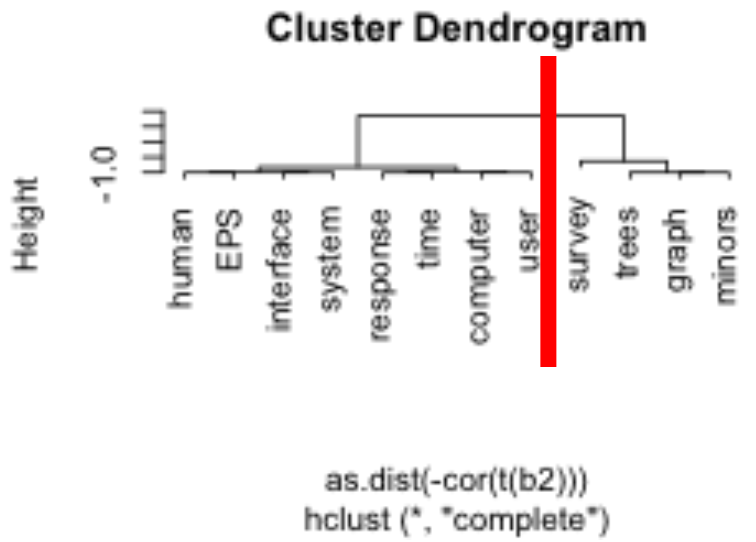




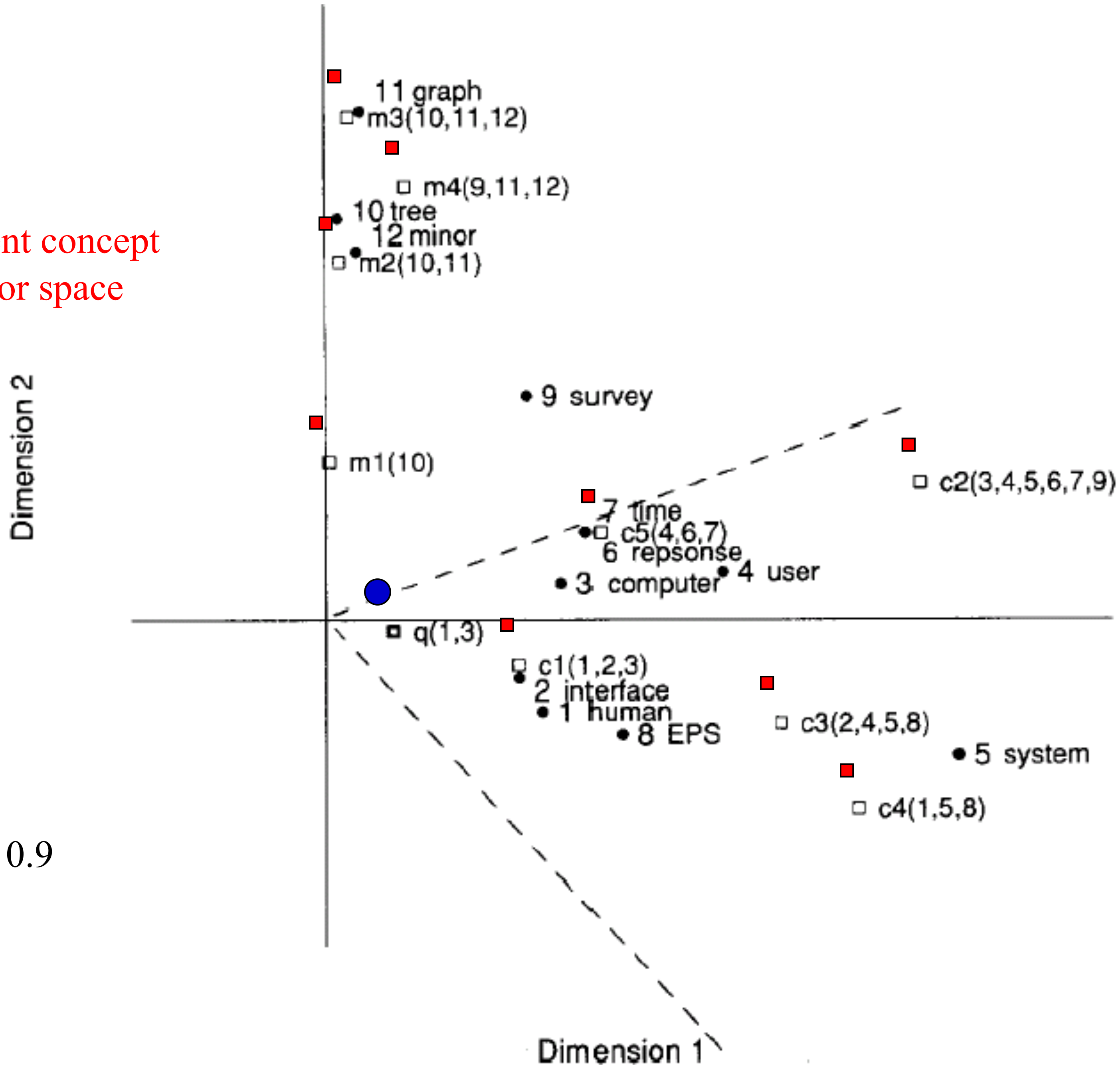
Term by Document Matrix

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





Latent concept vector space



- term
- document
- query
- cosine > 0.9

Approximate Nearest Neighbors (ANN)

- Indexing time:
  - Input: Embedding $M \in \mathbb{R}^{N \times K}$
  - Output: Indexes
- Query time:
  - Input:
    - Embedding, Indexes, query
  - Query: $q \in \mathbb{R}^{K}$
  - Output: candidates, $c \in \mathbb{R}^{K}$
    - where c is near q
    - sorted by $sim(q, c)$

```python
from gensim.similarities.annoy import AnnoyIndexer

# 100 trees are being used in this example
annoy_index = AnnoyIndexer(model, 100)
# Derive the vector for the word "science" in our model
vector = wv["science"]
# The instance of AnnoyIndexer we just created is passed
approximate_neighbors = wv.most_similar([vector], topn=11, indexer=annoy_index)
# Neatly print the approximate_neighbors and their corresponding cosine similarity values
print("Approximate Neighbors")
for neighbor in approximate_neighbors:
    print(neighbor)
```

https://radimrehurek.com/gensim/auto_examples/tutorials/run_annoy.html

https://pypi.org/project/annoy/

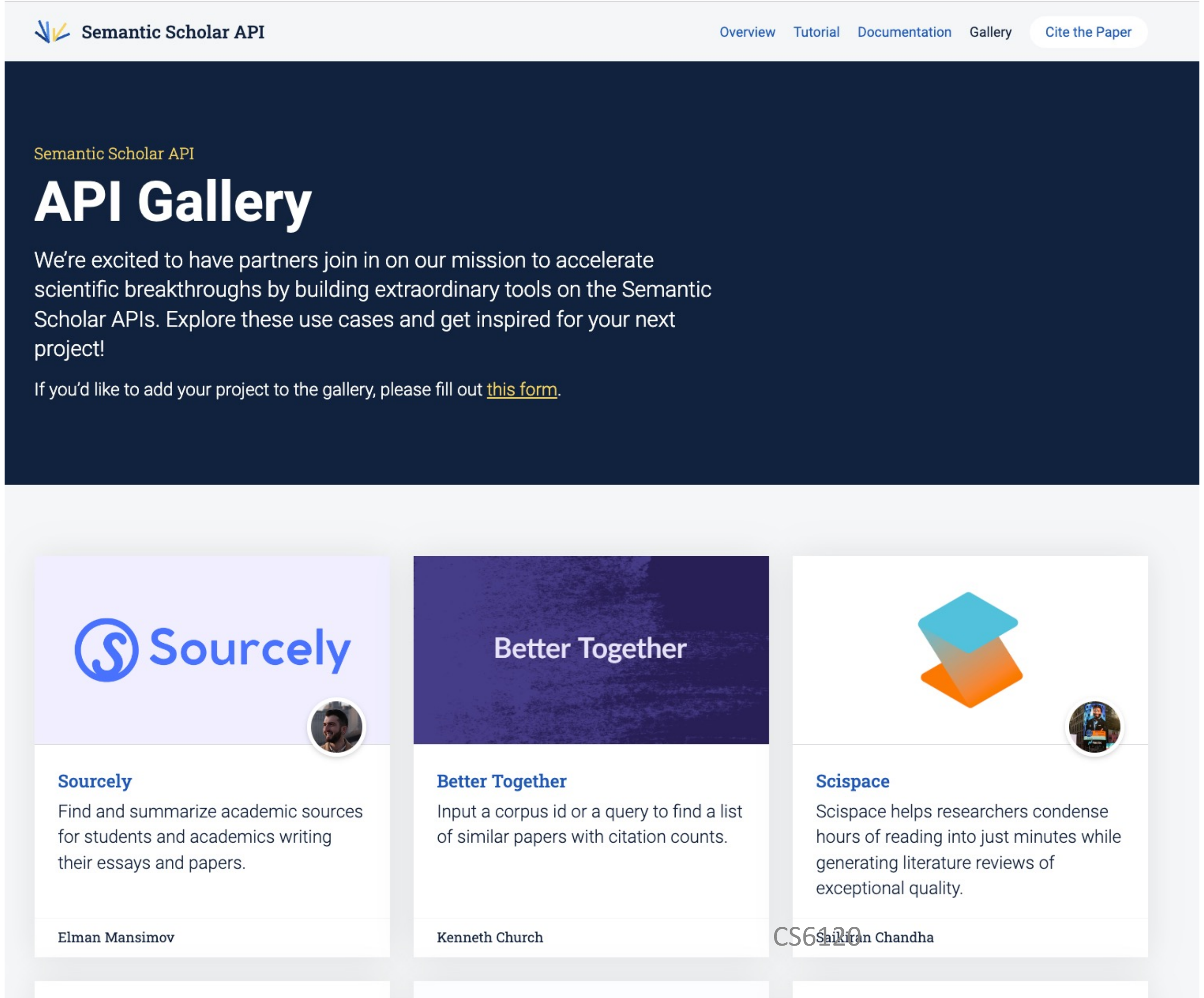
Formula for Survey Papers
(Start thinking about your final project)

✓Summarize main points of paper

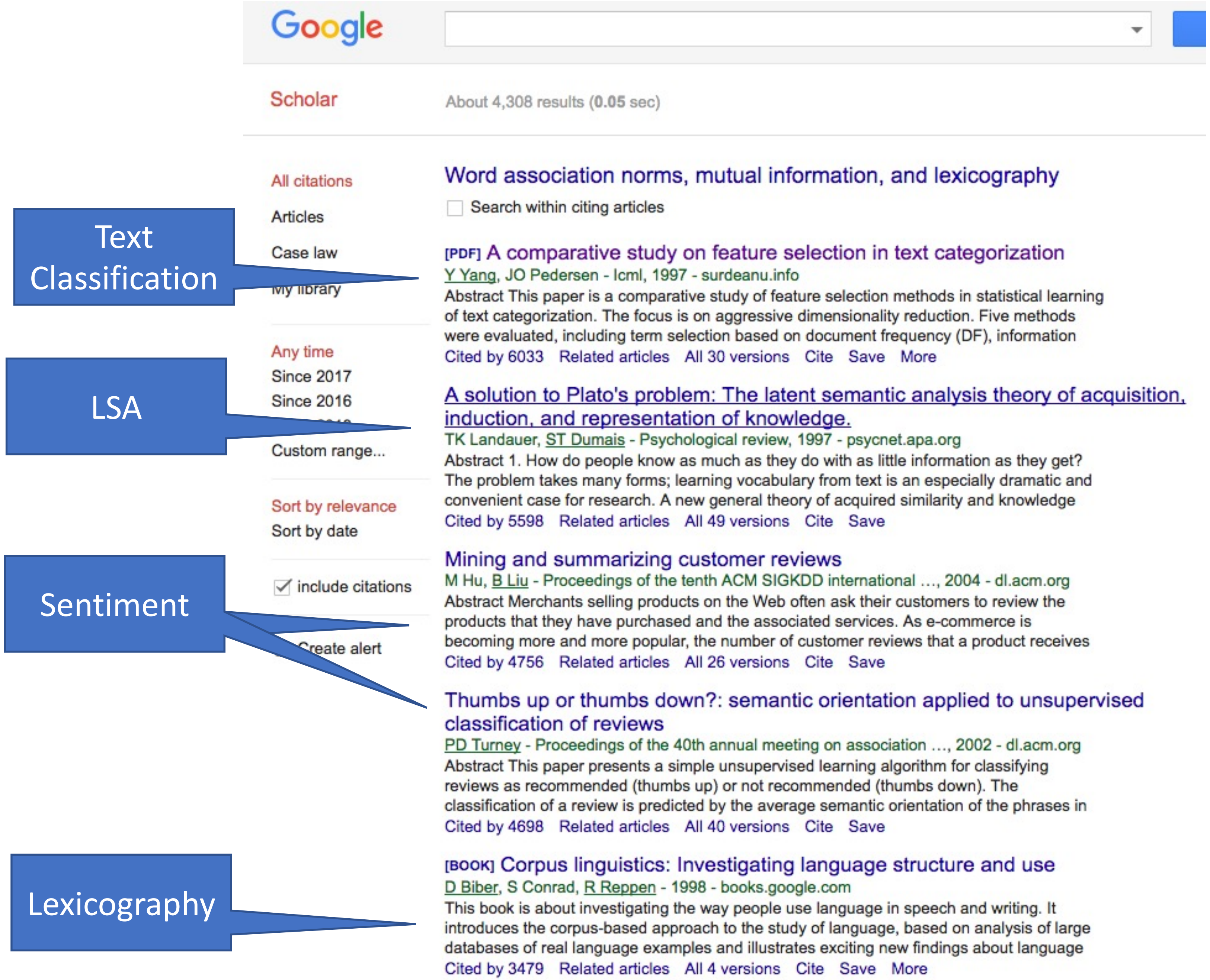
- Call out
  - ➢some highlights of subsequent literature
  - suggestions for future work

Shameless Plug

https://www.semanticscholar.org/product/api/gallery



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

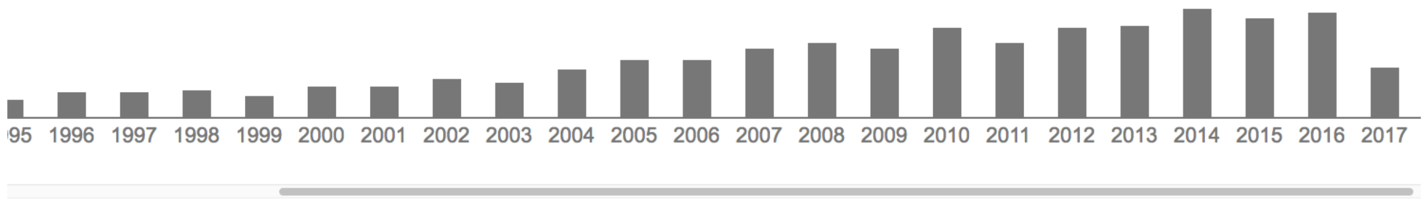
Issue 1

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)

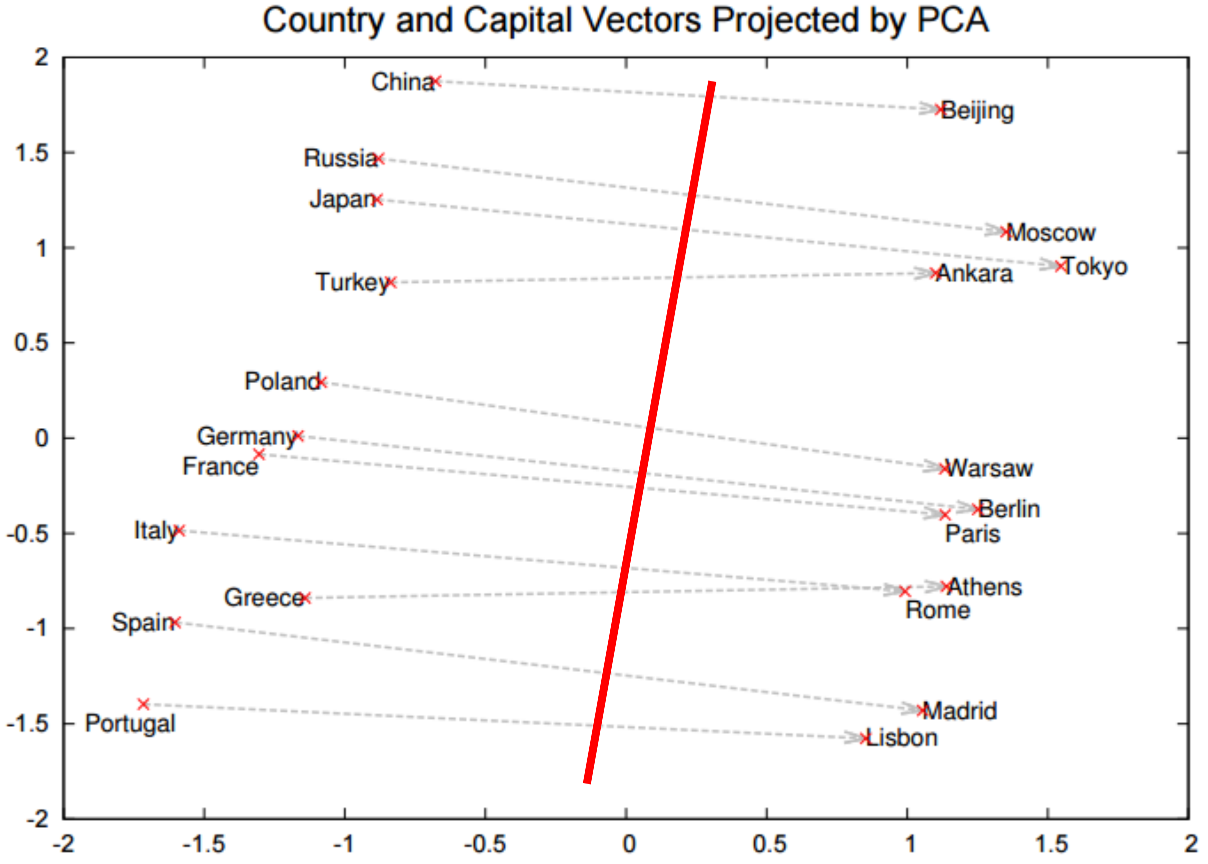
- Word2vec is not first, last or best to discuss
  - Vector spaces, embeddings, analogies, similarity metrics, etc.
- But word2vec is simple and accessible
  - Anyone can download the code and use it in their next paper.
  - Any many do (for better and for worse)
- Available downloads
  - Pre-computed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

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

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
  - $\hat{x} = \underset{x \in V}{\mathrm{ARGMAX}}\, sim(x', king + woman - man)$

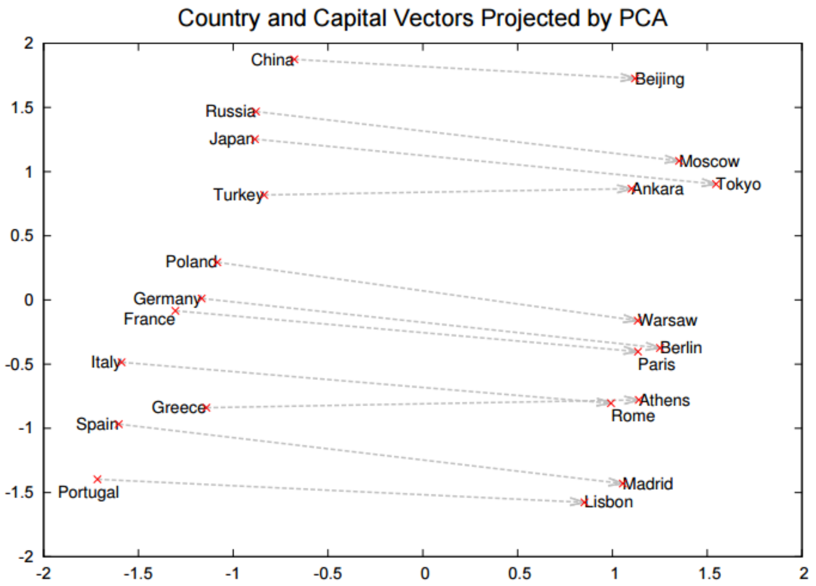
x	Gender	Number
Queen	f	sg
Monarch	m	sg
Princess	f	sg
Crown˙prince	m	sg
Prince	m	sg
Kings	m	pl
Queen˙Consort	m	sg
Queens	f	pl
Sultan	m	sg
Monarchy	m	sg

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

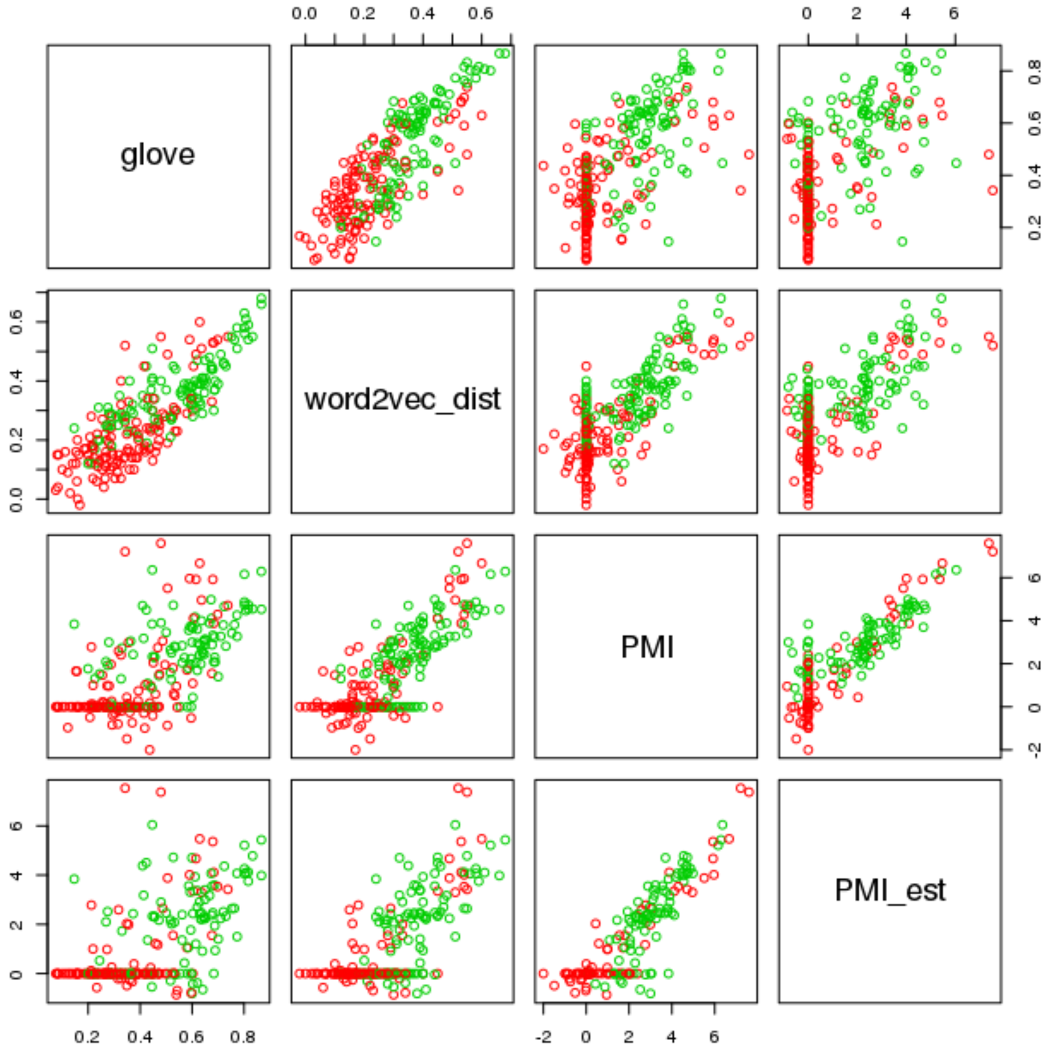
Table 2. *Some types of analogies are easier than others, as indicated by accuracies for top choice (A_1), as well as top 2 (A_2), top 10 (A_{10}) and top 20 (A_{20}). The rows are sorted by A_1. These analogies and the type classification come from the questions-words test set, except for the last row, 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	$\frac{young}{younger} = \frac{wide}{wider}$
0.90	0.94	0.97	0.98	1,599	Nationality-adjective	$\frac{Ukraine}{Ukrainian} = \frac{Switzerland}{Swiss}$
0.90	0.93	0.97	0.98	1,332	Plural	$\frac{woman}{women} = \frac{snake}{snakes}$
0.87	0.94	1.00	1.00	1,122	Superlative	$\frac{young}{youngest} = \frac{wide}{widest}$
0.85	0.90	0.97	1.00	506	Family	$\frac{uncle}{aunt} = \frac{stepson}{stepdaughter}$
0.83	0.89	0.97	0.98	335	Capital-countries	$\frac{Tokyo}{Japan} = \frac{Tehran}{Iran}$
0.79	0.86	0.94	0.96	4,695	Capital-world	$\frac{Zagreb}{Croatia} = \frac{Dublin}{Ireland}$
0.78	0.84	0.98	0.99	1,056	Present-participle	$\frac{write}{writing} = \frac{walk}{walking}$
0.71	0.79	0.90	0.92	2,467	City-in-state	$\frac{Worcester}{Massachusetts} = \frac{Cincinnati}{Ohio}$
0.68	0.78	0.93	0.95	870	Plural-verbs	$\frac{write}{writes} = \frac{work}{works}$
0.66	0.82	0.97	0.98	1,560	Past-tense	$\frac{writing}{wrote} = \frac{walking}{walked}$
0.43	0.48	0.64	0.69	812	Opposite	$\frac{tasteful}{distasteful} = \frac{sure}{unsure}$
0.35	0.42	0.57	0.62	866	Currency	$\frac{Vietnam}{dong} = \frac{USA}{dollar}$
0.29	0.37	0.63	0.73	992	Adjective-to-adverb	$\frac{usual}{usually} = \frac{unfortunate}{unfortunately}$
0.01	0.02	0.08	0.10	190	SAT questions	$\frac{audacious}{boldness} = \frac{sanctimonious}{hypocrisy}$



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

Agenda

- Homework
  - Assignment 1: [Better Together]
  - Assignment 2: [HuggingFace Pipelines]
- ✓ Background Material
  - ✓ Python
    - ✓ numpy, matplotlib, requests, json
    - ✓ sklearn, scipy
    - ✓ requests: APIs (Semantic Scholar)
  - ✓ Linear Algebra
  - ✓ Graph Algorithms
  - ✓ Probability
  - ✓ Machine Learning
- ✓ Old Business
  - ✓ (Nearly) everything → Vector
    - ✓ Word2vec
    - ✓ Doc2vec
  - ✓ Similarity → Cosine
  - ✓ Approximate Nearest Neighbors
- New Business
  - Colab
  - Deep Nets: Inference
    - Classification & Regression
    - Anything → Vector
    - Machine Translation
    - Fill Mask

HuggingFace Pipelines
[Colab]

- See https://huggingface.co/docs/transformers/main_classes/pipelines#transformers.pipeline.task for a list of currently supported tasks.
  - machine learning:
    - classification, regression, token classification, classify spans, fill mask
  - speech:
    - speech-to-text (automatic speech recognition (ASR), text-to-speech (speech synthesis), audio classification
  - vision:
    - image classification, video classification, image segmentation, image to text, visual question answering
  - natural language:
    - text classification, question answering, fill mask, translation

Back Translation and Conjunction

Synonyms (not equivalent)

- celestial and divine
  - 天天和天天
  - Every day and every day
- wisdom and erudition
  - 智慧和智慧
  - Wisdom and wisdom
- mournful and tearful
  - 悲伤和悲伤
  - Sadness and sadness

Antonyms

- coolness and eagerness
  - 寒凉和殷勤
  - The cold and the warmth
- fractious and blithesome
  - 讨人厌 讨人厌 讨人厌
  - I'm sick of it. I'm sick of it. I'm sick of it.

backup

tf-idf: Term Weighting

- Words have different importance, overlooked by simple count
- tf: term frequency: $\Pr(t|d)$, where t (terms) are IID events

$$tf_{t,d} = \frac{count(t,d)}{\sum_t count(t,d)} = \Pr(t|d)$$

- idf: inverse document frequency

$$idf_t = \log\left(\frac{\#\, total\ docs}{\#docs\ that\ have\ term\ t}\right) = -\log(\Pr(t \in d))$$

- tf-idf for word t in document d: $tf_{t,d} \times idf_t$
  - Interpretation:
    - Entropy: $H = -\log(P)$
    - where $P = \Pr(t \in d)^{count(t,d)}$