SAND2016-11367PE

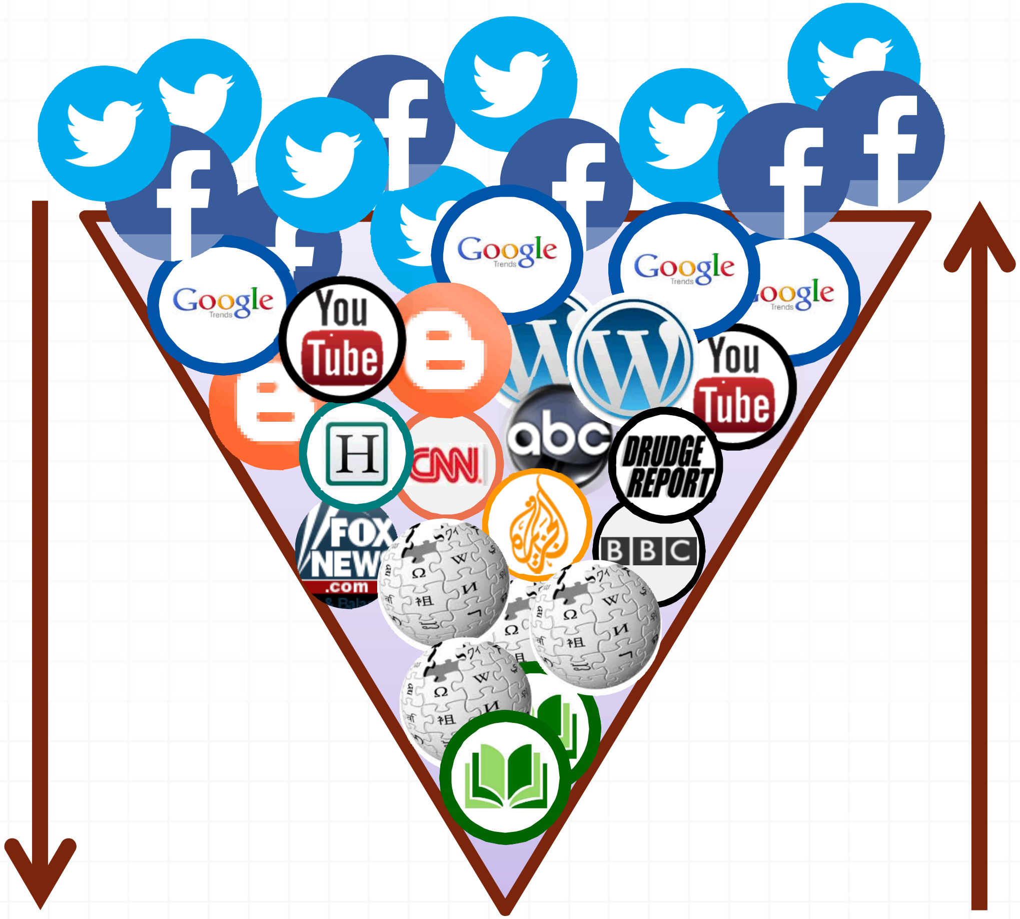
An Intro to Text Analytics

Sandia National Laboratories

November 2016

Content from Jeremy Wendt and Travis Bauer

Written communication varies a lot



Info density

Cost/effort

Impact

Durability

Latency

Intent/fortitude

Throughput

Immediacy

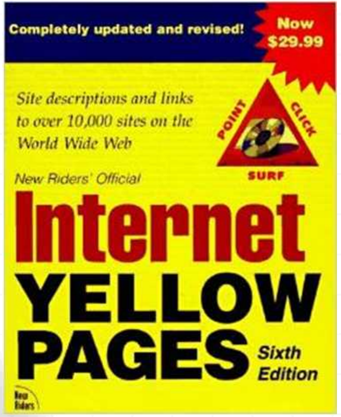
Every Search Problem Is Unique

Google-scale

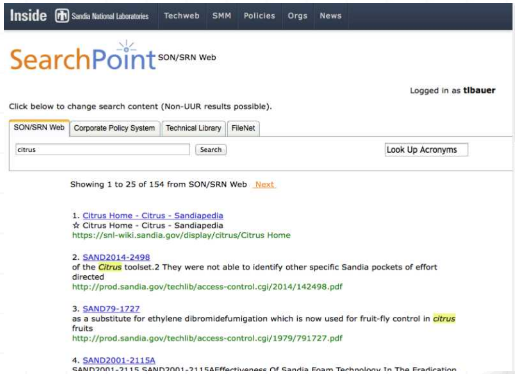
- 100,000,000 GB index
- Few canonical answers
- Rapidly changing
- Hidden content
  - Dark web
  - Paywalls/usernames
- Dynamically generated content

Sandia-scale

- 813,000 documents
- Canonical answers exist
- Slower to change
- Content issues
  - All content must be approved
  - Some requires authorities
- Some dynamically generated content


Completely updated and revised!
Now $29.99
Site descriptions and links to over 10,000 sites on the World Wide Web
POINT
CLICK
SURF
New Riders' Official
Internet
YELLOW
PAGES
Sixth Edition
New Riders

Very different problems…
Very different algorithms required



Logged in as **tlbauer**

Click below to change search content (Non-UUR results possible).

Showing 1 to 25 of 154 from SON/SRN Web Next

1. Citrus Home - Citrus - Sandiapedia
☆ Citrus Home - Citrus - Sandiapedia
https://snl-wiki.sandia.gov/display/citrus/Citrus Home

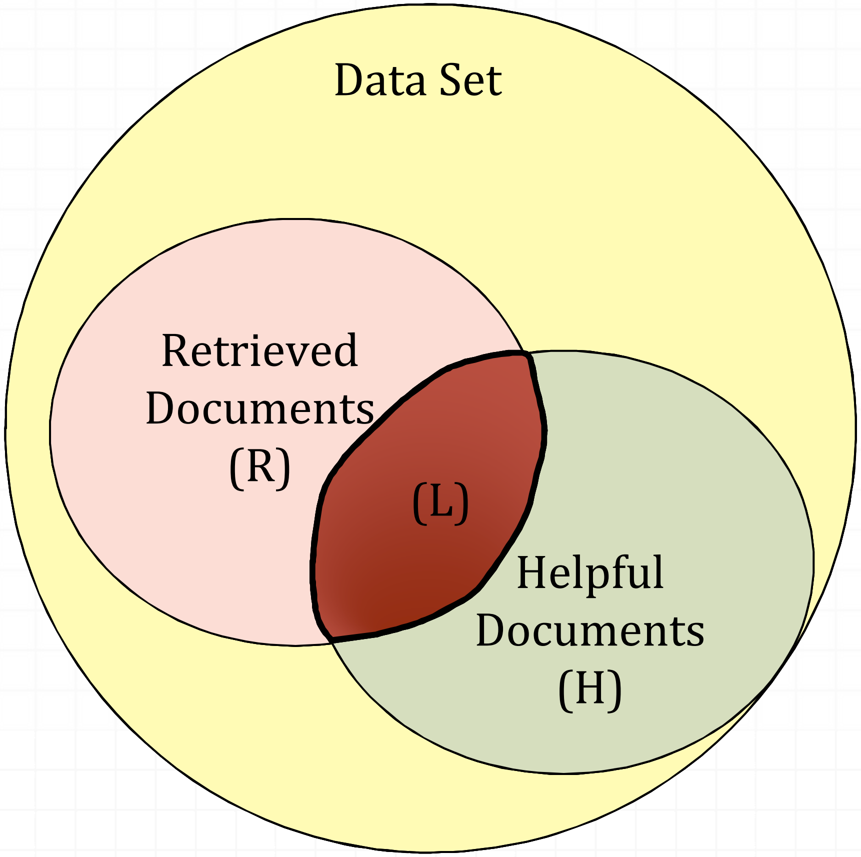
2. SAND2014-2498
of the *Citrus* toolset.2 They were not able to identify other specific Sandia pockets of effort directed
http://prod.sandia.gov/techlib/access-control.cgi/2014/142498.pdf

3. SAND79-1727
as a substitute for ethylene dibromidefumigation which is now used for fruit-fly control in *citrus* fruits
http://prod.sandia.gov/techlib/access-control.cgi/1979/791727.pdf

4. SAND2001-2115A
SAND2001-2115 SAND2001-2115AEffectiveness Of Sandia Foam Technology In The Eradication

Define Quality Measure

- Precision:
  - L / R – The fraction of retrieved documents that are useful
- Recall:
  - L / H – The fraction of the useful documents retrieved
- Perfection?



Zipf’s Law

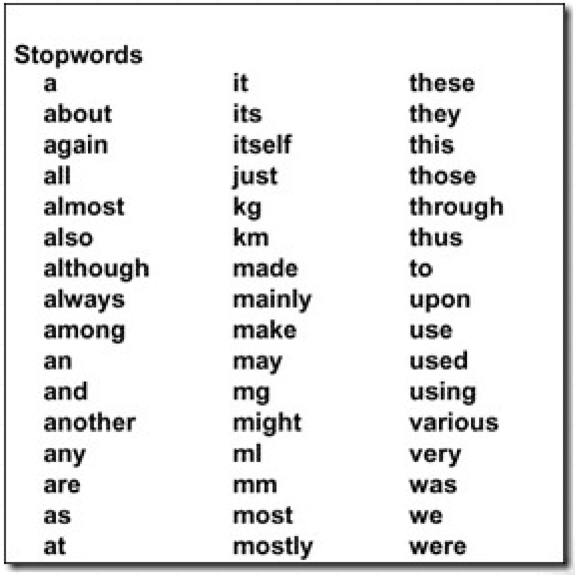
“Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.”

George Kinsley Zipf, 1935

- Which of those questions does this affect?

Stopwords

- The most common words show up in nearly all documents
  - Not very useful for search
- They also show up the most in many documents
  - Visualizations can be dominated
- Don’t include them in the matrix

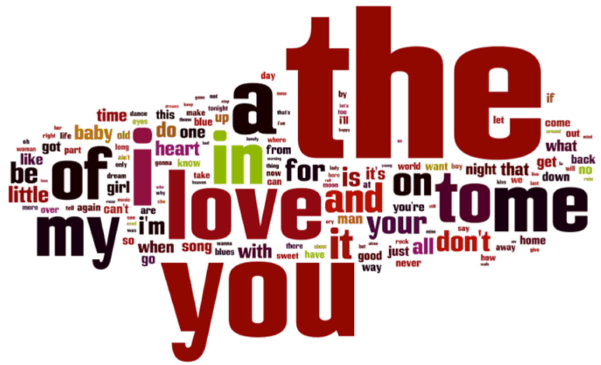


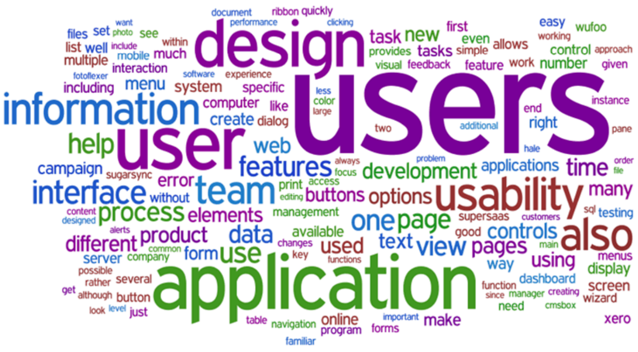
Stopwords

a
about
again
all
almost
also
although
always
among
an
and
another
any
are
as
at

it
its
itself
just
kg
km
made
mainly
make
may
mg
might
ml
mm
most
mostly

these
they
this
those
through
thus
to
upon
use
used
using
various
very
was
we
were





Sample Document Term Matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	1	1	0	0	0	0	0	0	0	0
boat	0	0	1	1	1	0	0	0	0	0
cargo	1	1	1	1	1	0	0	0	0	0
board	0	0	0	0	1	1	0	0	0	0
player	0	0	0	0	0	1	1	1	1	1
turn	0	0	0	0	0	0	0	0	1	1
win	0	0	0	0	0	1	1	1	0	0

- Linear Independence
  - Synonymy
  - Polysemy

Synonymy

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	1	1	0	0	0	0	0	0	0	0
boat	0	0	1	1	1	0	0	0	0	0
cargo	1	1	1	1	1	0	0	0	0	0
board	0	0	0	0	1	1	0	0	0	0
player	0	0	0	0	0	1	1	1	1	1
turn	0	0	0	0	0	0	0	0	1	1
win	0	0	0	0	0	1	1	1	0	0

- Linear Independence
  - **Synonymy**
  - Polysemy
- Compression

Polysemy

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	1	1	0	0	0	0	0	0	0	0
boat	0	0	1	1	1	0	0	0	0	0
cargo	1	1	1	1	1	0	0	0	0	0
board	0	0	0	0	1	1	0	0	0	0
player	0	0	0	0	0	1	1	1	1	1
turn	0	0	0	0	0	0	0	0	1	1
win	0	0	0	0	0	1	1	1	0	0

- Linear Independence
  - Synonymy
  - **Polysemy**
- Compression

After SVD

This matrix is the result of running SVD

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	.28	.28	.35	.35	.40	.02	-.03	-.03	-.03	-.03
boat	.55	.55	.68	.68	.78	.10	-.01	-.01	-.02	-.02
cargo	.83	.83	1.02	1.02	1.18	.12	-.04	-.04	-.05	-.05
board	.21	.21	.27	.27	.35	.35	.27	.27	.21	.21
player	-.05	-.05	-.04	-.04	.12	1.18	1.02	1.02	.83	.83
turn	-.03	-.03	-.03	-.03	.02	.40	.35	.35	.28	.28
win	-.02	-.02	-.01	-.01	.10	.78	.68	.68	.55	.55

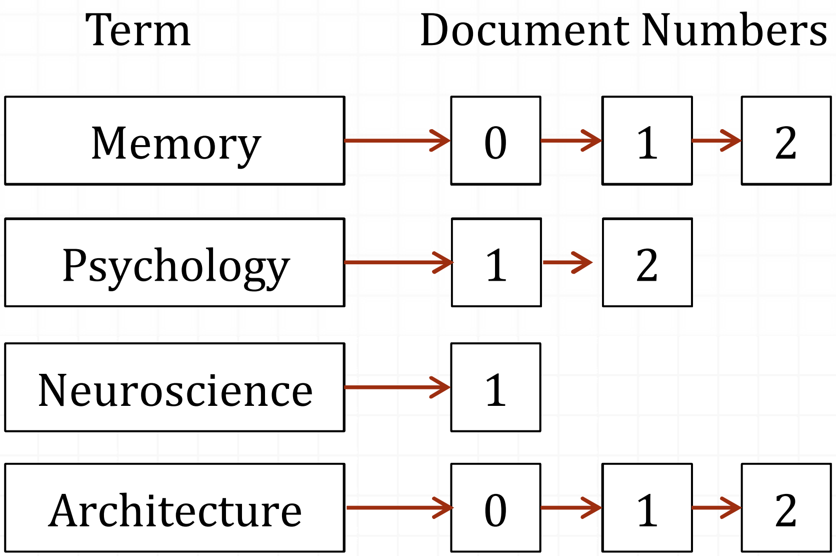
The Two Matrices

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	1	1	0	0	0	0	0	0	0	0
boat	0	0	1	1	1	0	0	0	0	0
cargo	1	1	1	1	1	0	0	0	0	0
board	0	0	0	0	1	1	0	0	0	0
player	0	0	0	0	0	1	1	1	1	1
turn	0	0	0	0	0	0	0	0	1	1
win	0	0	0	0	0	1	1	1	0	0

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
ship	.28	.28	.35	.35	.40	.02	-.03	-.03	-.03	-.03
boat	.55	.55	.68	.68	.78	.10	-.01	-.01	-.02	-.02
cargo	.83	.83	1.02	1.02	1.18	.12	-.04	-.04	-.05	-.05
board	.21	.21	.27	.27	.35	.35	.27	.27	.21	.21
player	-.05	-.05	-.04	-.04	.12	1.18	1.02	1.02	.83	.83
turn	-.03	-.03	-.03	-.03	.02	.40	.35	.35	.28	.28
win	-.02	-.02	-.01	-.01	.10	.78	.68	.68	.55	.55

Inverted Index

- Only record the non-zero values



Term | Document Numbers

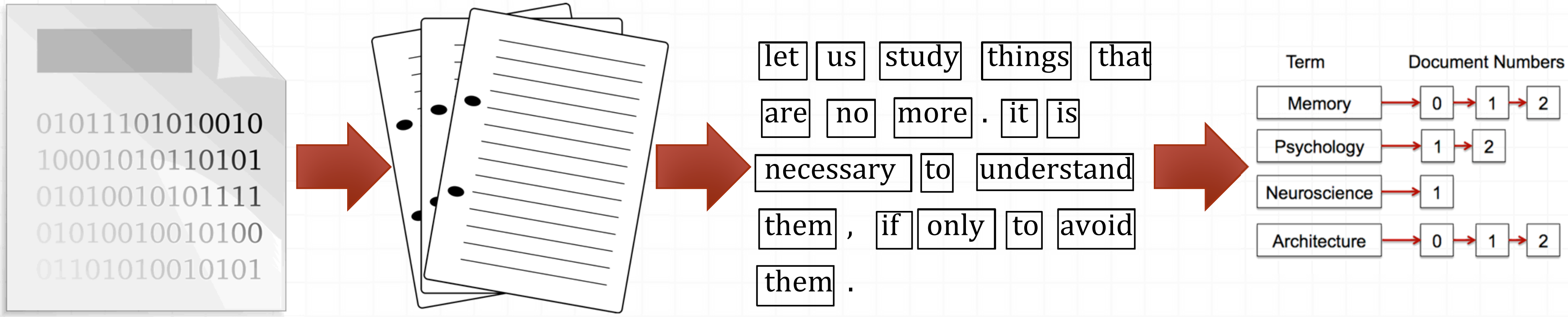
Memory → 0 → 1 → 2

Psychology → 1 → 2

Neuroscience → 1

Architecture → 0 → 1 → 2

Creating the Index



let us study things that are no more . it is necessary to understand them , if only to avoid them .

Term	Document Numbers
Memory	0 → 1 → 2
Psychology	1 → 2
Neuroscience	1
Architecture	0 → 1 → 2

File of Bytes

Extract Fields of Text

Extract Terms

Building the Inverted Index

Extract Terms

- Tokenization
  - Cut character sequences into word tokens
    - Split on spaces
    - Make decisions regarding hyphens, colons, etc.
    - Phrases
- Normalization
  - Change case
  - Remove internal punctuation (for example, change U.S.A. to USA)
- Stemming
  - Run, running, runner get turned into “run”
  - Often not useful
    - A document that is sufficiently about some term is likely to include the relevant stems
    - Collapse terms in unwanted ways (e.g. runny)
    - There are other ways to deal with this issue

Consistency is critical

Term Frequency

- We can store the count of the term occurrence in the index for each document (zeros still implied)
- More formally, $tf_{d,t}$ is the frequency of term *t* in document *d*
- Example:
  - “John is quicker than Mary” = [john:1, mary:1, quicker:1]
  - “Mary is quicker than John” = [john:1, mary:1, quicker:1]

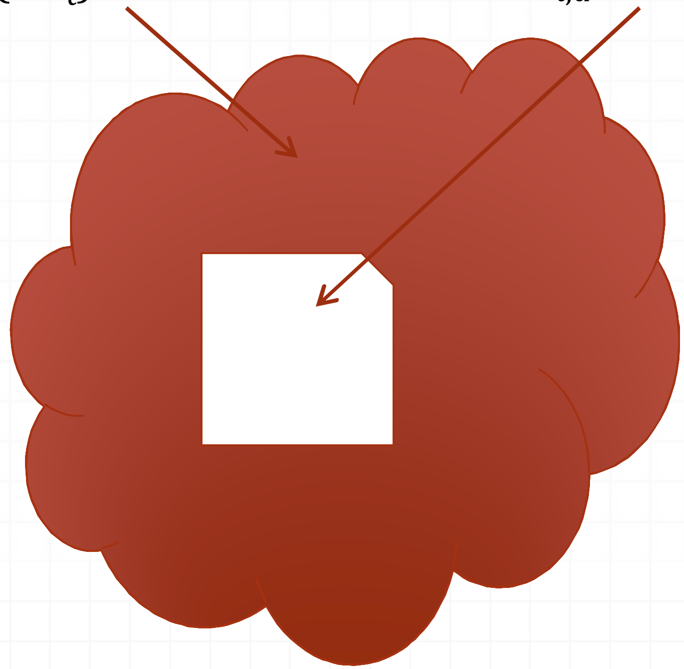
Inverse Document Frequency

- Document frequency: The number of documents that contain some term

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

  - N = number of documents in the corpus
  - Denominator = document frequency (often corrected for divide by zero by adding 1 to denom)
- Thus, tf * idf gets you the words that are most common in the document and least common in the corpus

TF*IDF



Global Weight
(idf_t)

Local Weight
($tf_{t,d}$)

A document as a query

- We can use a whole document as a query and search for the most similar document in the corpus
  - Sort results by similarity
- How do we compute similarity?
  - Absolute value of vector difference
    - If the vectors have similar distribution, but different scale…
  - Dot product
    - Longer documents have higher probability of containing the queried terms
  - Cosine similarity
    - Normalizes the dot product by document length
- Another scoring metric that uses the entropy of the term across the corpus and the count within a document
  - Well normalized

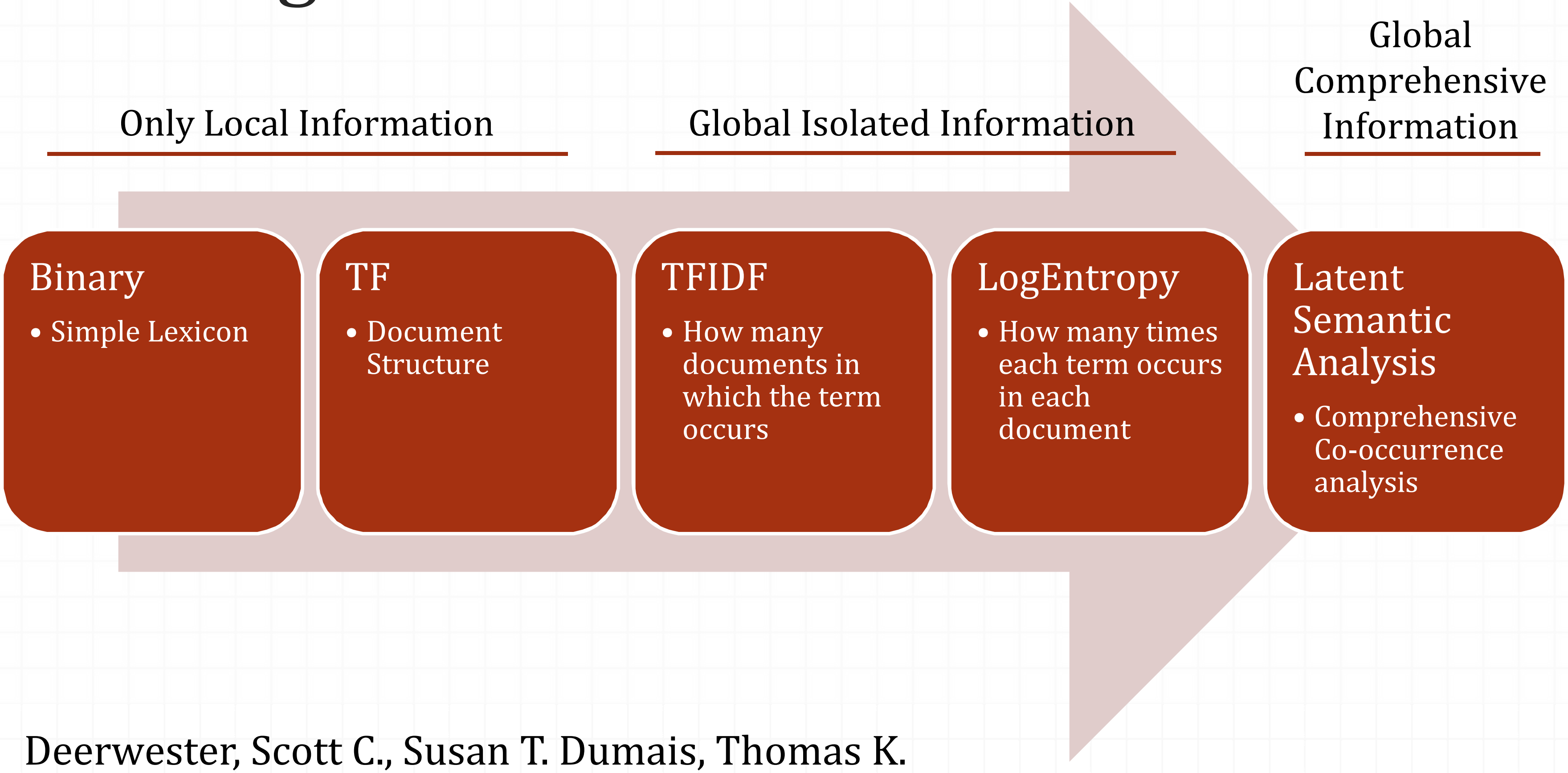
Topic Modeling

- Leaving search
- Given a large document corpus
  - What topics are covered in the documents?
  - Which documents fit into which topics?
  - Which documents don’t fit in any of these topics?
    - Both as a quality measure for the topics and to look at the documents

We’ve Built a Matrix…

- … so, let’s do matrix operations
- Latent Semantic Analysis (LSA) is based on a singular value decomposition
- Latent Dirichlet Allocation (LDA)
  - LDA uses Bayesian stats to assign probabilities to document-topic pairs, and uses a sampling method (usually Gibbs) to iterate over the documents and terms.
  - LDA outputs both:
    - Per-document topic distribution
    - Per-topic term distribution
  - These outputs are often converted (via setting a probability threshold) into an overlapping distribution of documents in topics (i.e., some documents are assigned to multiple topics and some to no topic at all).
- LDA topics tend to be a little more human-understandable
- Overlapping documents is preferable for many applications (effectively it increases precision)

Taking More information Into Account



Only Local Information

Global Isolated Information

Global Comprehensive Information

Binary
- Simple Lexicon

TF
- Document Structure

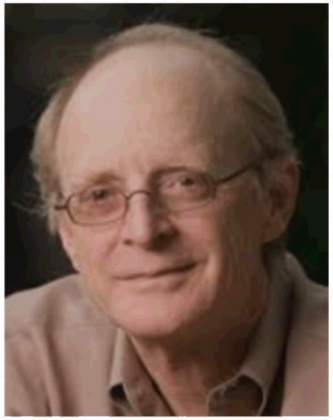
TFIDF
- How many documents in which the term occurs

LogEntropy
- How many times each term occurs in each document

Latent Semantic Analysis
- Comprehensive Co-occurrence analysis

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. "Indexing by latent semantic analysis." *JAsIs* 41, no. 6 (1990): 391-407.

James Pennebaker



- 1991: Psychology professor studying how/why people recover from traumatic events
- Ran a series of studies
  - People wrote about traumatic experiences
    - Many subjects improved considerably
      - Immune function boosts
      - Blood pressure drops
      - Depression reduces
      - Mood improves

- Why did writing work? Why did it not for some?

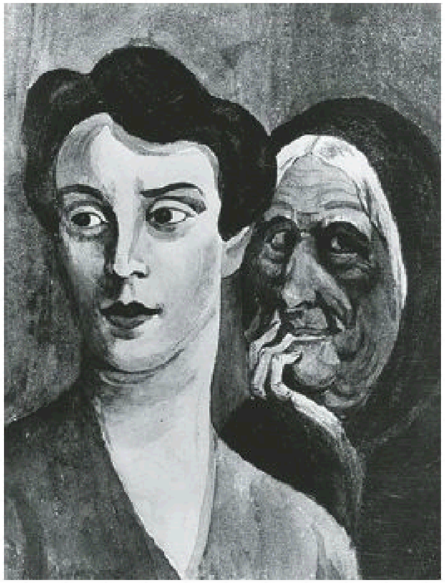
Evaluate the Essays

- LSA was brand new and shiny
  - Tried looking at topics the subjects wrote about
  - No evidence that topics led to different outcomes
- Flip from looking at topic to writing style
  - Stopwords critical to writing style
    - ~~One~~ ~~might~~ ~~want~~ ~~to~~ try working smarter
    - ~~I~~ ~~want~~ ~~to~~ try working smarter

Results

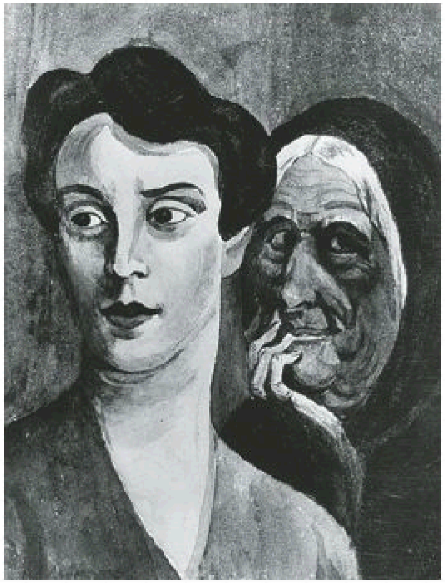
- "The results were breathtaking. (Ok, if you are not a computational linguist, 'breathtaking' may be a bit of an overstatement. You had to be there.) The more people changed in the ways they used function words from writing to writing, the more their health later improved … More specifically, the more people changed their use of first-person singular pronouns (e.g., *I, me, my*) compared with other pronouns (e.g. *we, you, she, they*), the better their health later became.

What’s a function word?



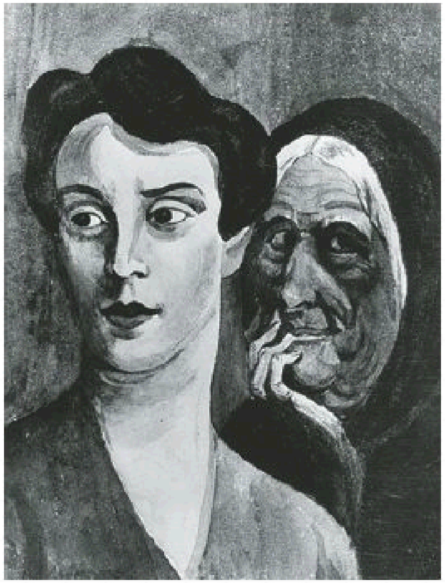
- PERSON 1: In the aforementioned picture an elderly woman is about to speak to a middle aged woman who looks condescending and calculating
- PERSON 2: I see an old woman looking back on her years remembering how it was to be beautiful and young.
- PERSON 3: The old woman is a witch or something. She looks kinda like she is coaxing the young one to do something.

What's a function word? Not these



- PERSON 1: In the aforementioned picture an elderly woman is about to speak to a middle aged woman who looks condescending and calculating
- PERSON 2: I see an old woman looking back on her years remembering how it was to be beautiful and young.
- PERSON 3: The old woman is a witch or something. She looks kinda like she is coaxing the young one to do something.

What's a function word? These



- PERSON 1: In the aforementioned picture an elderly woman is about to speak to a middle aged woman who looks condescending and calculating
- PERSON 2: I see an old woman looking back on her years remembering how it was to be beautiful and young.
- PERSON 3: The old woman is a witch or something. She looks kinda like she is coaxing the young one to do something.

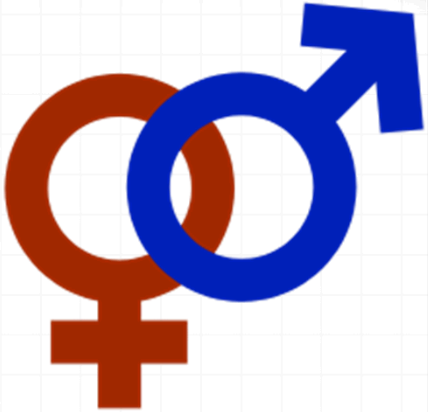
Function Words

- Average vocabulary size = ~100,000 words
- Total number of function words = 450 words
    - ~55% of word occurrences
- 99.96% of our vocabulary responsible for less than half of word occurrences!

Word	Percent of all words
I	3.64
the	3.48
and	2.92
to	2.91
a	1.94
of	1.83
that	1.48
in	1.29
it	1.19
my	1.08
is	1.06
you	1.05
was	1.01
for	0.8
have	0.7
with	0.67
he	0.66
me	0.64
on	0.63
but	0.62
TOTAL	**29.6**

Extension to Other Areas

- When you hit gold, keep digging!
- They found function word usage varies over many areas
  - Personality
  - Age
  - Gender
  - Social class
  - Stress levels
  - Biological activity
  - Social relationships



Finding	Discussion
Women use first-person singular pronouns (I-words) more than men	Research suggests that women, on average, are more self-aware and self-focused than are men.
Men and women use first person plural words (we-words) at the same rate.	“We” is actually two very different words. • “Warm and fuzzy” we (me and my dog) (women) • Impersonal We – “We really need to analyze that data.” (men)
Men use articles (a, an, the) more than women.	More concrete, highly specific nouns.
No difference in positive emotional words	
Women use more cognitive words than men.	Because of the next line
Women use social words at far higher rates than men.	“Social words” are about relating to other human beings. Women talk more about other people.
Men also use more “big words” and swear words.	
Women use more negative emotion words, negations, certainty words (always, absolutely), and hedge phrases (“I think”)	

How big is the difference?

- "Although men and women use words differently, the differences can often be subtle. In one large study of over fourteen thousand language samples, we found that 14.2 percent of women's words were personal pronouns compared with 12.7 percent for men. From that statistical perspective, this is a *huge* difference. The kind of whopping statistical effect that brings tears of joy to a scientist's eyes (or at least mine). But …" at a speaking rate of 100 words per minute a woman would only mention about one and a half pronouns more than a man.
- 100,000 blog posts (19,320 authors)
  - Computer 72% correct
  - Humans 55%-65% correct

Deception

Letters of Recommendation

- 200 letters of recommendation written by Pennebaker
- Rated how he truly felt about the student
- For students he rated highly
  - Used Longer Sentences
  - Bigger words
  - *Fewer* positive emotion words (really)
  - Provided more detailed information
    - Talked more about what the students did than about the students themselves
  - Paid little attention to the reader
    - "As you can see…"
    - "I'm sure you agree that …"

Deception

Stephen Glass

- New Republic journalist from the late 1990’s
  - 6 articles completely invented
  - 21 articles partially fraudulent
  - 14 articles likely trustworthy
- Real or likely Real Stories
  - Used more words, more numbers, more details
  - Fewer emotion (especially positive emotion) and cognitive words
  - Fewer verbs
  - Fewer self-references (I-words)

Verbal Mimicking

“[When engaged in conversation] people also converge in the ways they talk – they tend to adopt the same levels of formality, emotionality, and cognitive complexity. In other words, people tend to use the same groups of function words at similar rates.

“The matching of function words is called language style matching, or LSM. Analyses of conversations find that LSM occurs within the first fifteen to thirty seconds of any interaction and is generally beyond conscious awareness.”

LSM and Love

- 80 daters recorded during speed dating
  - Above-average LSM – almost twice as likely to want future contact as those with below-average LSM
  - LSM was a better indicator than the people themselves (because both have to agree)
- 80 young dating couples
  - Read their IM's (with permission)
  - Among the 43 couples with the highest LSM scores, 77% were still dating three months later (52% of others)
  - LSM was a better indicator than self-reports

Text Analysis

Conclusions

- Some really cool results
  - Large-scale search is largely “solved”
    - Corporate-level search … less so
  - Topic modeling early cool results
    - How do we describe the topics? Are they “right”?
  - Early results in understanding word “meaning”
    - Polysemy problems
- However, still can’t solve a lot of big problems
- Almost all tools require training