From Frequency to Meaning

Raf Alvarado
UVA DS 5001

Meaning and significance, statistical-semantic hypothesis, Zipf's Law, bags of words, vector space representations

Business

TBA

Review

See revised HMM notebook (if interested)

- Includes perplexity measures of performance

Also see NLTK notebook on generating NGrams

- Faster, but uses lazy iterators
- And data hard to get at

Meaning and Significance

What do we mean by meaning?

Reference — LOCUTION

Things **pointed** to (real or imagined)
Ostensive reference
Deixis

Motivation — ILLOCUTION

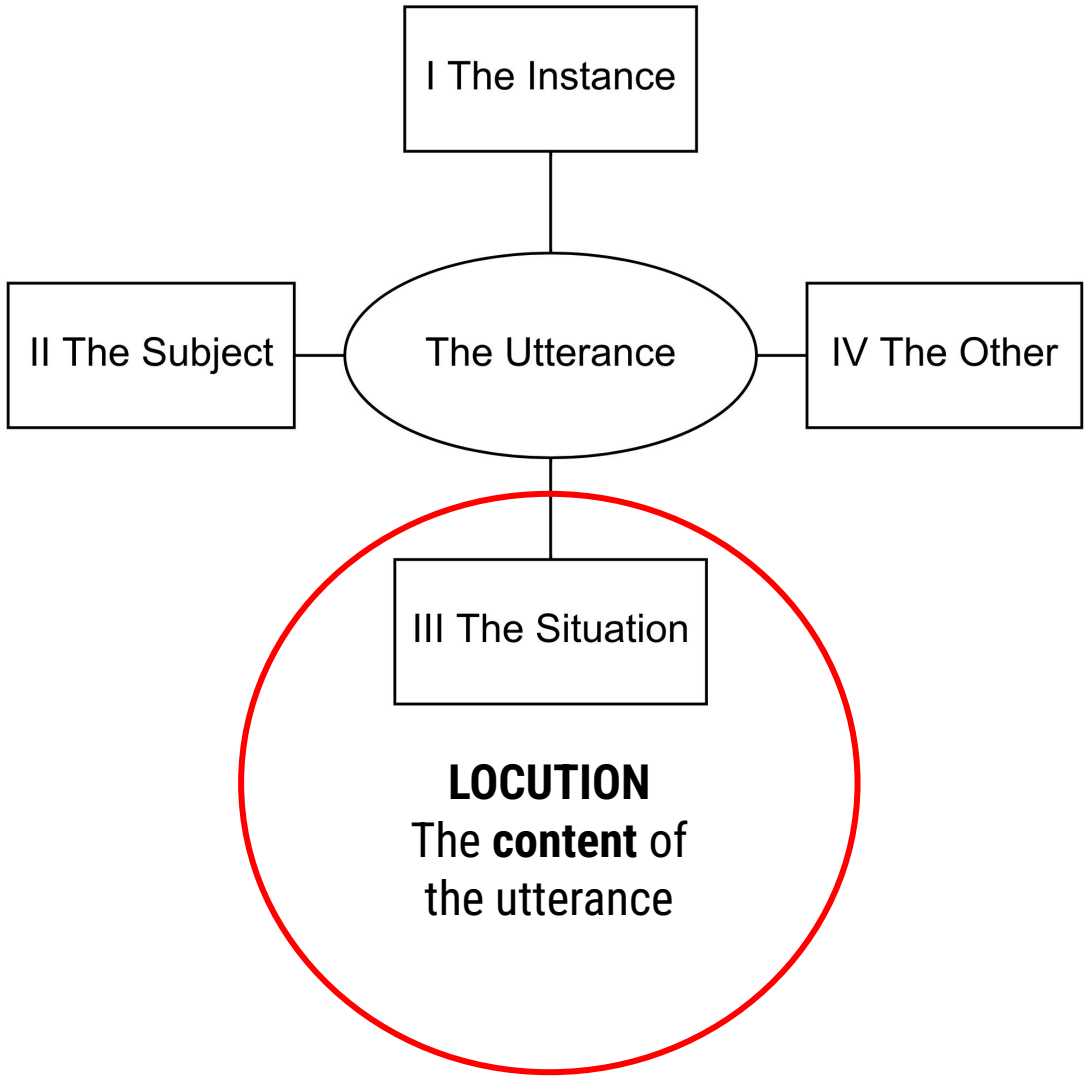
Why something is said

Effect — PERLOCUTION

The **impact** of words on others' **behavior**

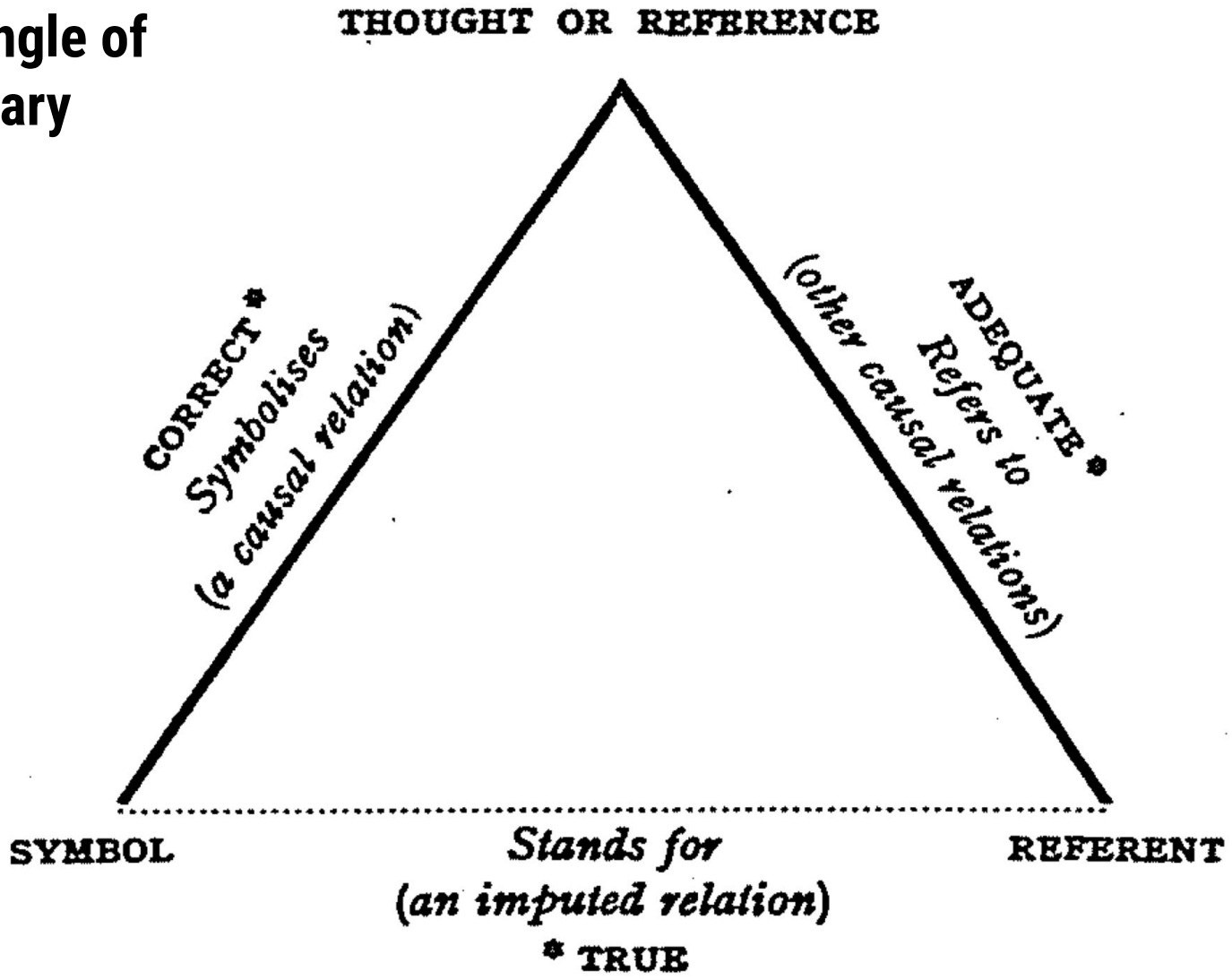
Model of Discourse as Model of Meaning

ILLOCUTION
The **force** or **sentiment** behind the utterance



PERLOCUTION
The **effect** of the utterance on other people

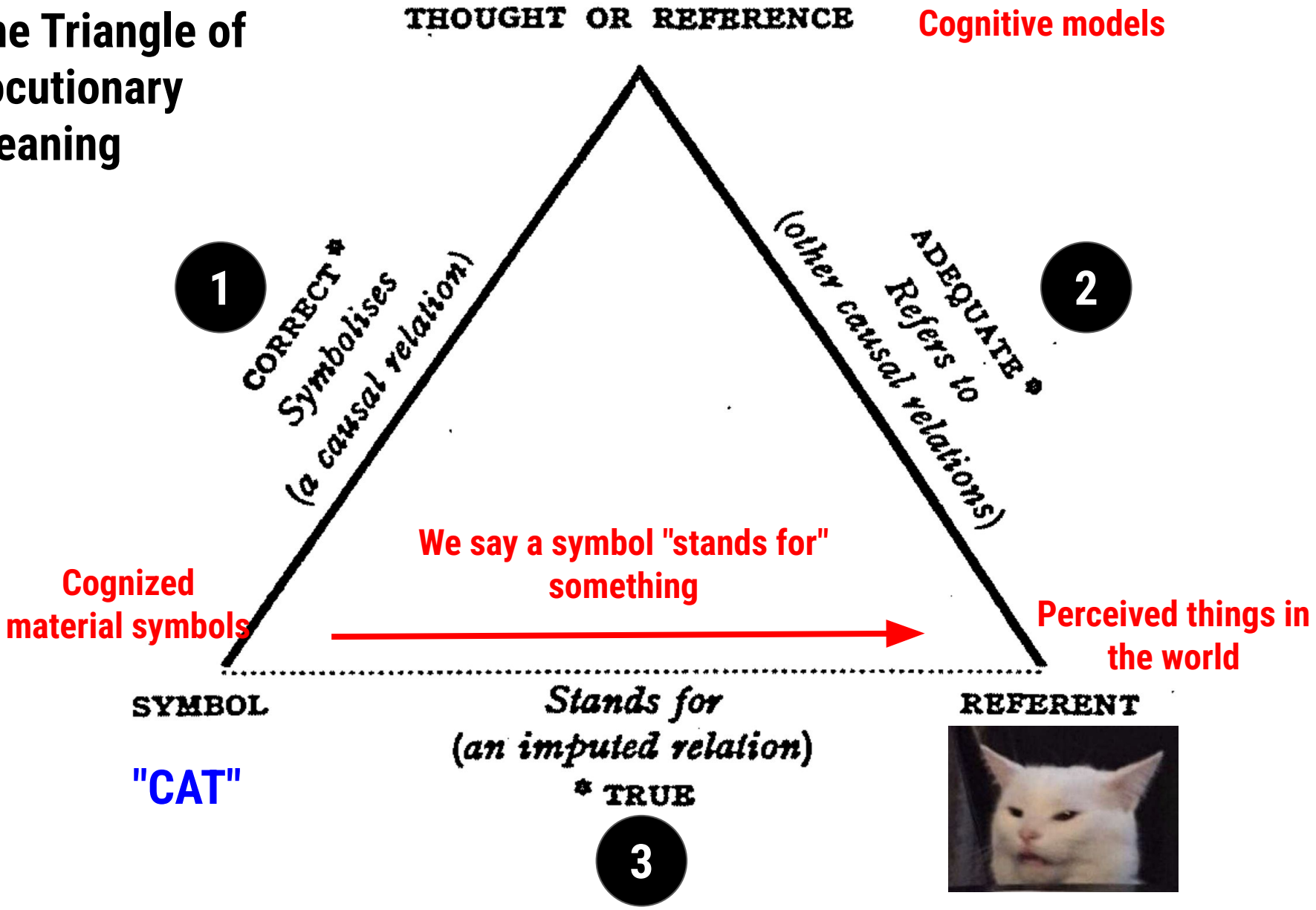
The Triangle of Locutionary meaning



From **Ogden and Richards,** 1923, *The Meaning of Meaning*

The Triangle of Locutionary meaning

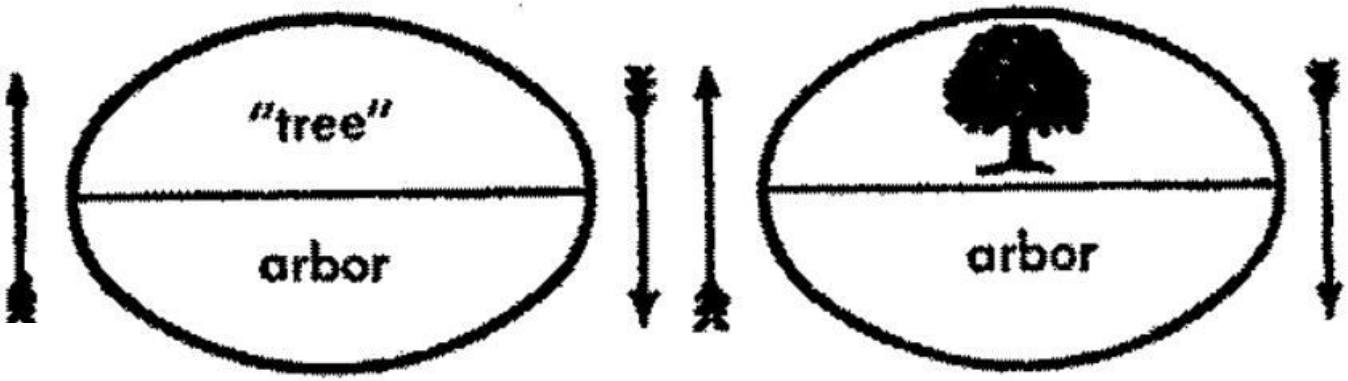
Cognitive models



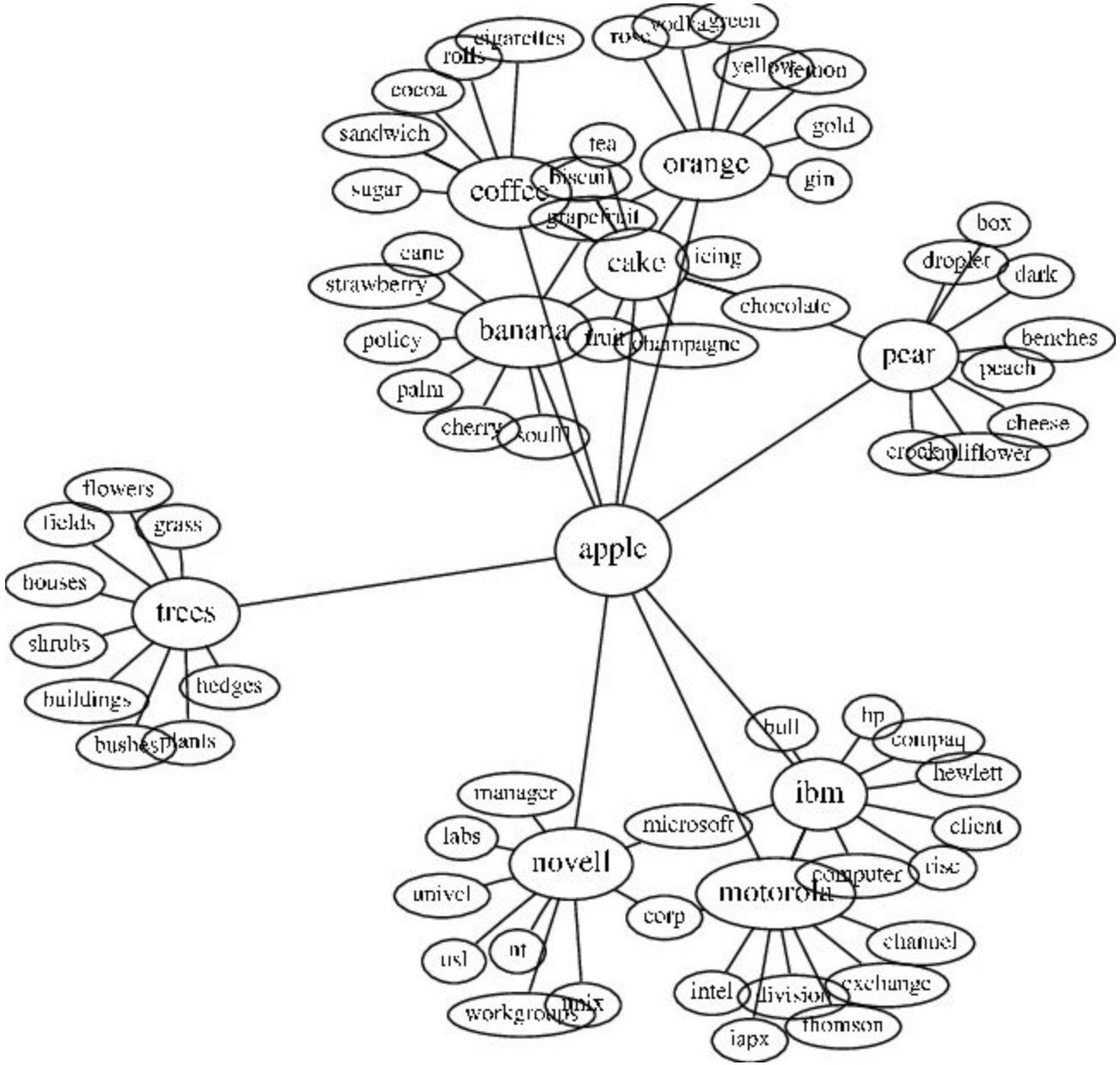
Meaning is primarily **psychological**. Words first stand for **thoughts**.

This claim is more **strongly** made by Saussure

From de Saussure's *Course in General Linguistics* (1916), on the nature of the Linguistic Sign



- "The linguistic sign unites, not a thing and a name, but a concept and a sound-image. The latter is not a material sound, a purely physical thing, but the psychological imprint of the sound, the impression it makes on our senses." (p.66)



In this view, meaning is a function of **location** in a network of symbols and meanings

An important variant of this view is called **structuralism**

Widdows, Dominic, and Beate Dorow. 2002. “A Graph Model for Unsupervised Lexical Acquisition.” In *COLING 2002: The 19th International Conference on Computational Linguistics*. https://aclanthology.org/C02-1114.pdf.

What do we mean by meaning?

Reference — LOCUTION

Things **pointed** to (real or imagined)
Ostensive reference
Deixis

Motivation — ILLOCUTION

Why something is said

Effect — PERLOCUTION

The **impact** of words on others' **behavior**

Structure — NON-OSTENSIVE REFERENCE (Ricoeur)

Connection to **other words** or ideas; symbols and meanings
Location in a **conceptual** space
Geometry of meaning

A Paradox

Ricoeur

Texts have been removed from the situation that gives discourse meaning (fixation and distantiation)

Makes evident **non-ostensive** meanings

Texts have a "**surplus of meaning**"

Shannon

Meaning is **irrelevant** to the engineering problem

All we need are **probabilities** and state machines (language models)

A Paradox

Luhn

> “The intellectual aspects of writing and of meaning cannot serve as elements of such machine systems. To a machine, words can be only so many physical things.”

A Paradox

But

Are **information** and **meaning** really *unrelated*?

Statistical significance and semantic significance?

Hypothesis

Statistical patterns of how words are used can be used to infer their **meanings**

In other words, **significance may be a proxy for meaning**

Statistical Semantic Hypothesis (revised)

$$\text{If } F(Sr_1) = F(Sr_2) \text{ then } Sd_1 \cong Sd_2$$

Where **F** refers to some **frequency** (probability) distribution and Sr and Sd are **sets**

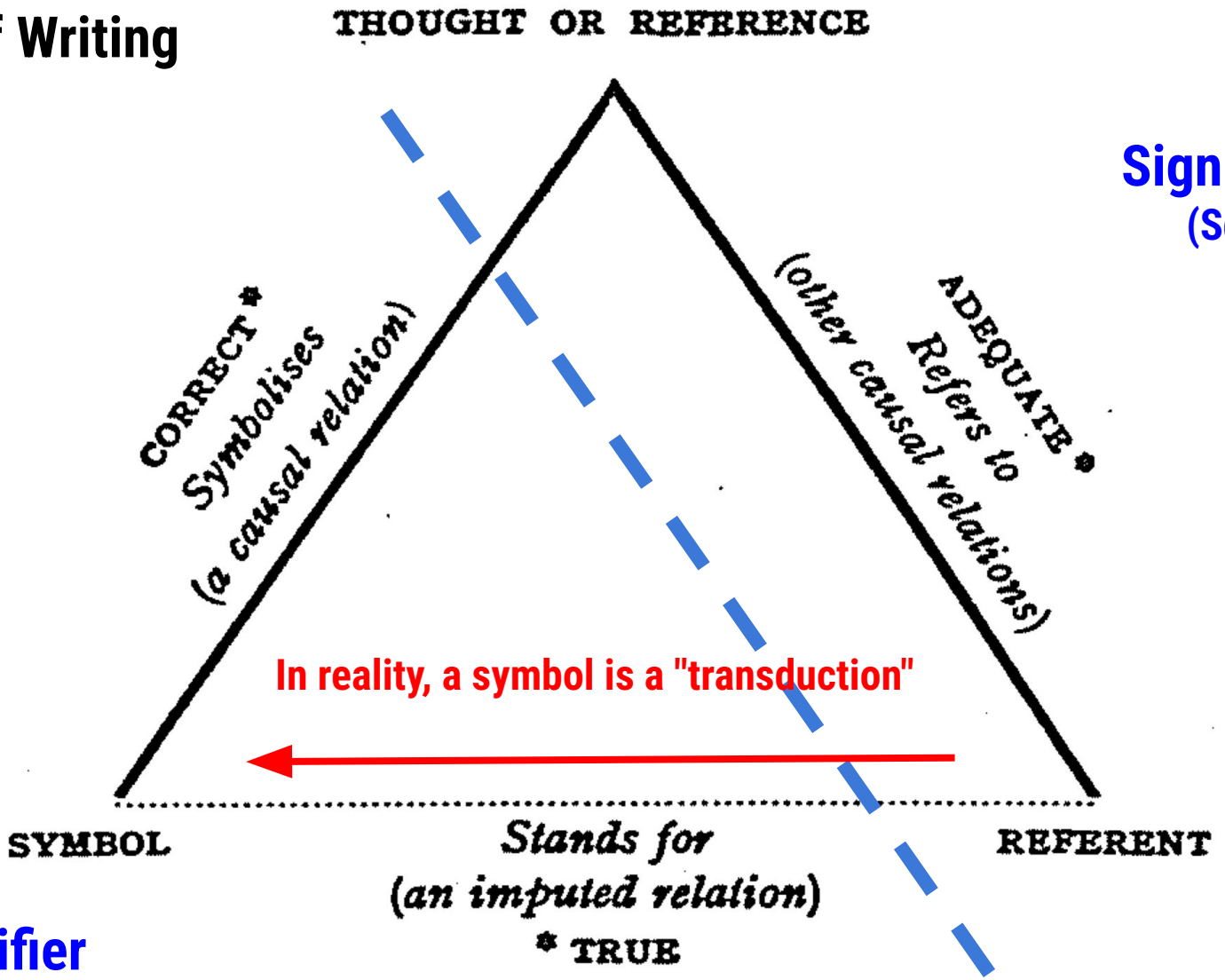
Sr = **SIGNIFIER (symbol)** ← **THE DATA**

Sd = **SIGNIFIED (meaning)** ← **LATENT**

Turney and Pantel ascribe this theory
to **Wittgenstein** and others

The **meaning** of a word is its **use**
(not what it refers to in the world)

Effect of Writing



Signified
(Sd)

Signifier
(Sr)

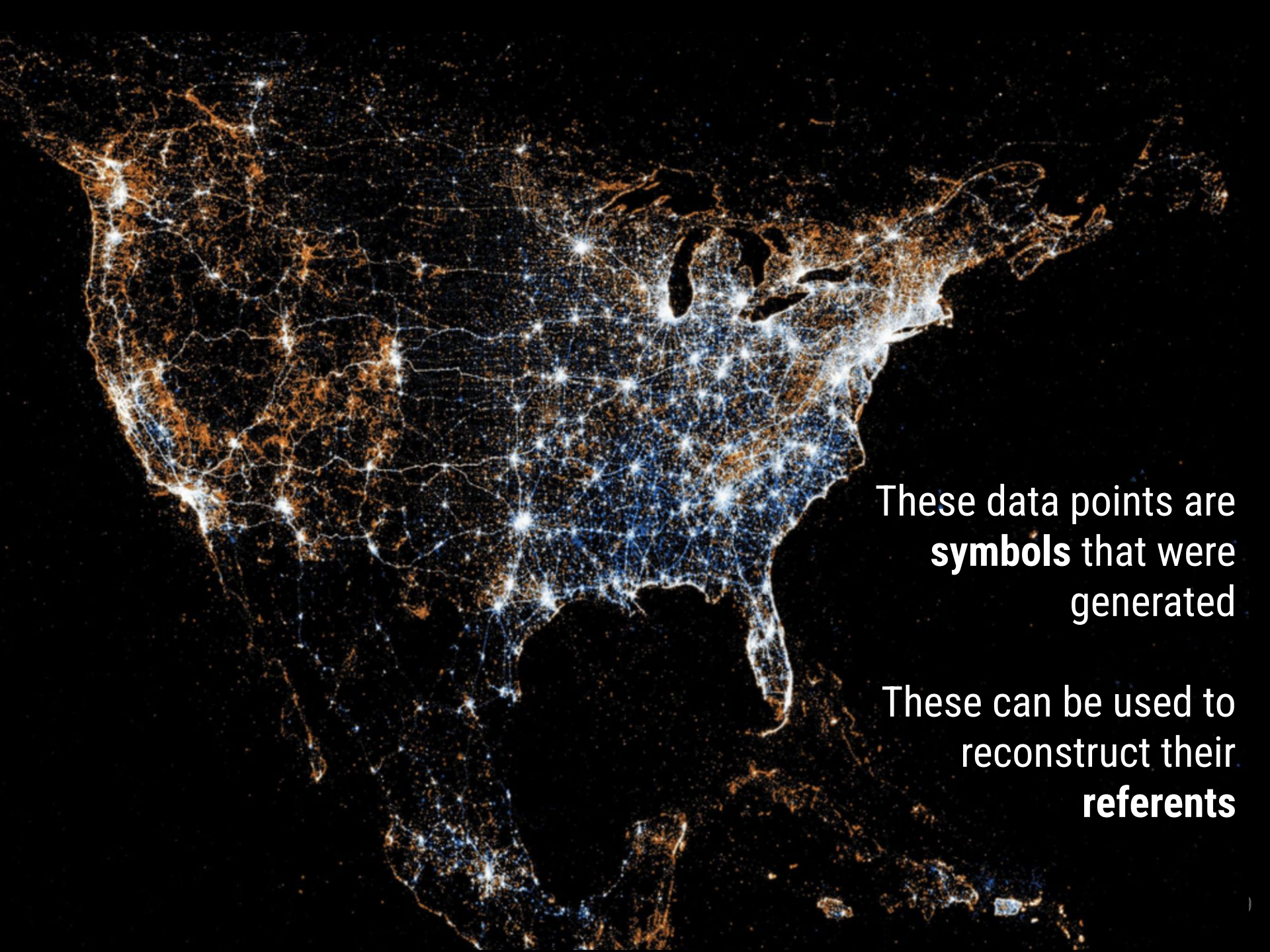
$$F(Sr_1) = F(Sr_2)$$

$$Sd_1 \cong Sd_2$$

Why would this be true?

The **statistical patterns** of the **traces** (signs) left by the act of **signification** (the utterance) reflect the **statistical patterns** in the **references**

Think of **references** -- ideas in the mind, things in the world -- as **causes** and **signs** as **effects**


These data points are **symbols** that were generated
These can be used to reconstruct their **referents**

Approaches

Approaches and Hypotheses

Global Term Frequencies (Zipf's Law)

Global term frequencies in a corpus follow a **universal pattern** that provides some insights into language

Bag-of-Words (BOW) Hypothesis

Within-document word frequencies **characterize the content of documents** and so can be used to group documents by similarity (and other things); Origin unknown — *Tristan Tzara's Dada poetry*?

Distributional Hypothesis

Term that occur in **similar term contexts** (e.g. sentences) tend to have similar meanings; Zellig Harris and J R Firth

Approaches and Hypotheses

TF-IDF

Combination of **local** (document) and **global** (corpus) term frequencies in a **bag-of-words** representation

Luhn

Anticipates TF-IDF

But combines **local term frequencies** with local term **clustering**

Ahead of his time ...

HANS PETER LUHN, MENTOR, 68, DIES; Data-Processing Specialist Served F.B.M. 20 Years

Give this article

Aug. 20, 1964

Prototypical **data scientist**

In 1958, Mr. Luhn, in a demonstration, took a 2,326 word article on hormones of the nervous system from The Scientific American, inserted it in the form of magnetic tape into an I.B.M. computer, and pushed a button. Three minutes later, the machine's automatic typewriter typed four sentences giving the gist of the article, of which the machine had made an abstract.

https://www.nytimes.com/1964/08/20/archives/hans-peter-luhn-mentor-68-dies-dataprocessing-specialist-served-fbm.html

Luhn

Argues that non-stopwords that are frequent in a document **and close to each other** such words are considered significant

> "ideas most **closely associated intellectually** are found to be implemented by words most **closely associated physically**"

Example of Statistical-Semantic Hypothesis

> Also, Interesting takes advantage of OHCO . . .
>
> "The divisions of written text into **sentences**, **paragraphs**, **chapters**, et cetera, is another physical manifestation of the graduating degree of **association of ideas**."

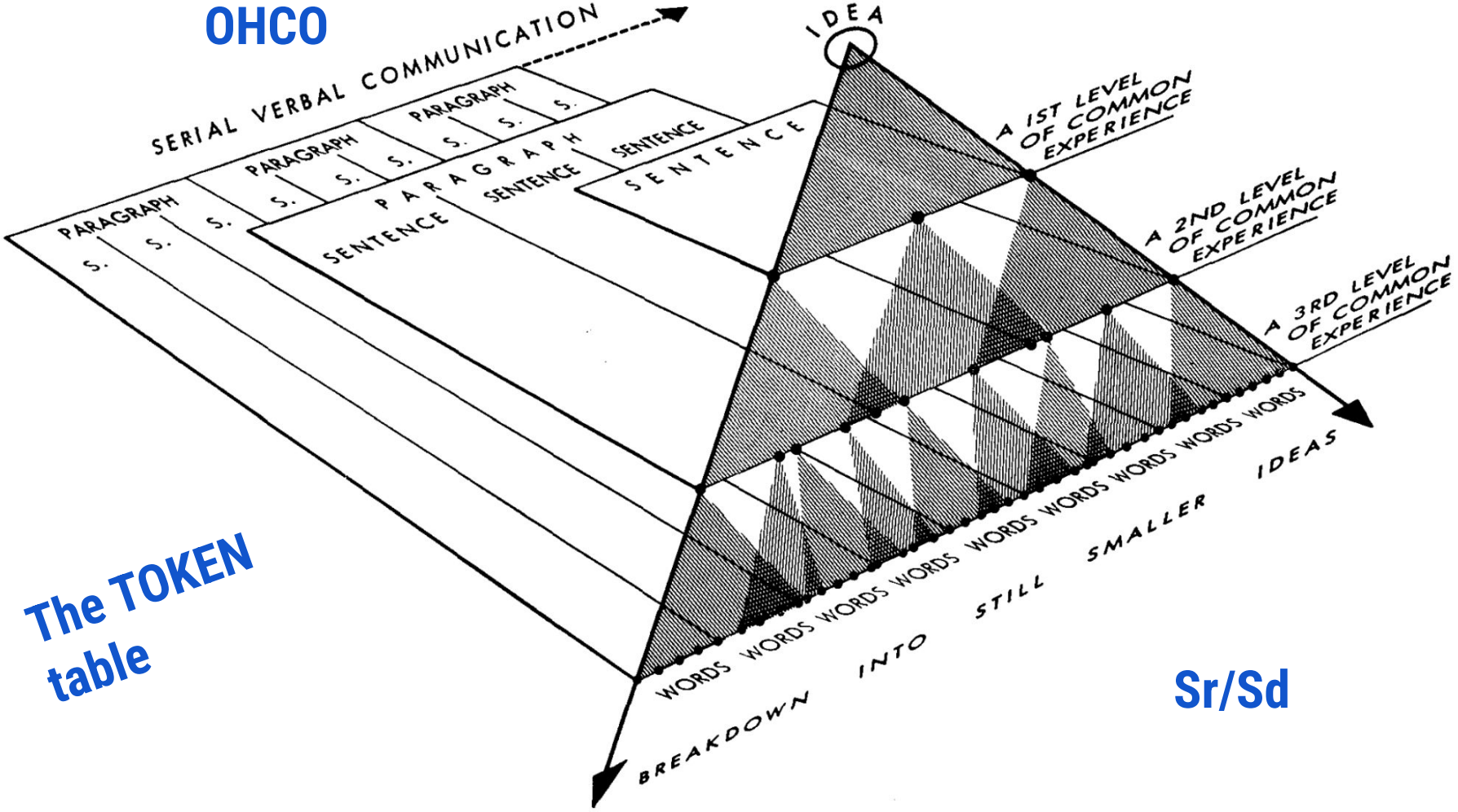
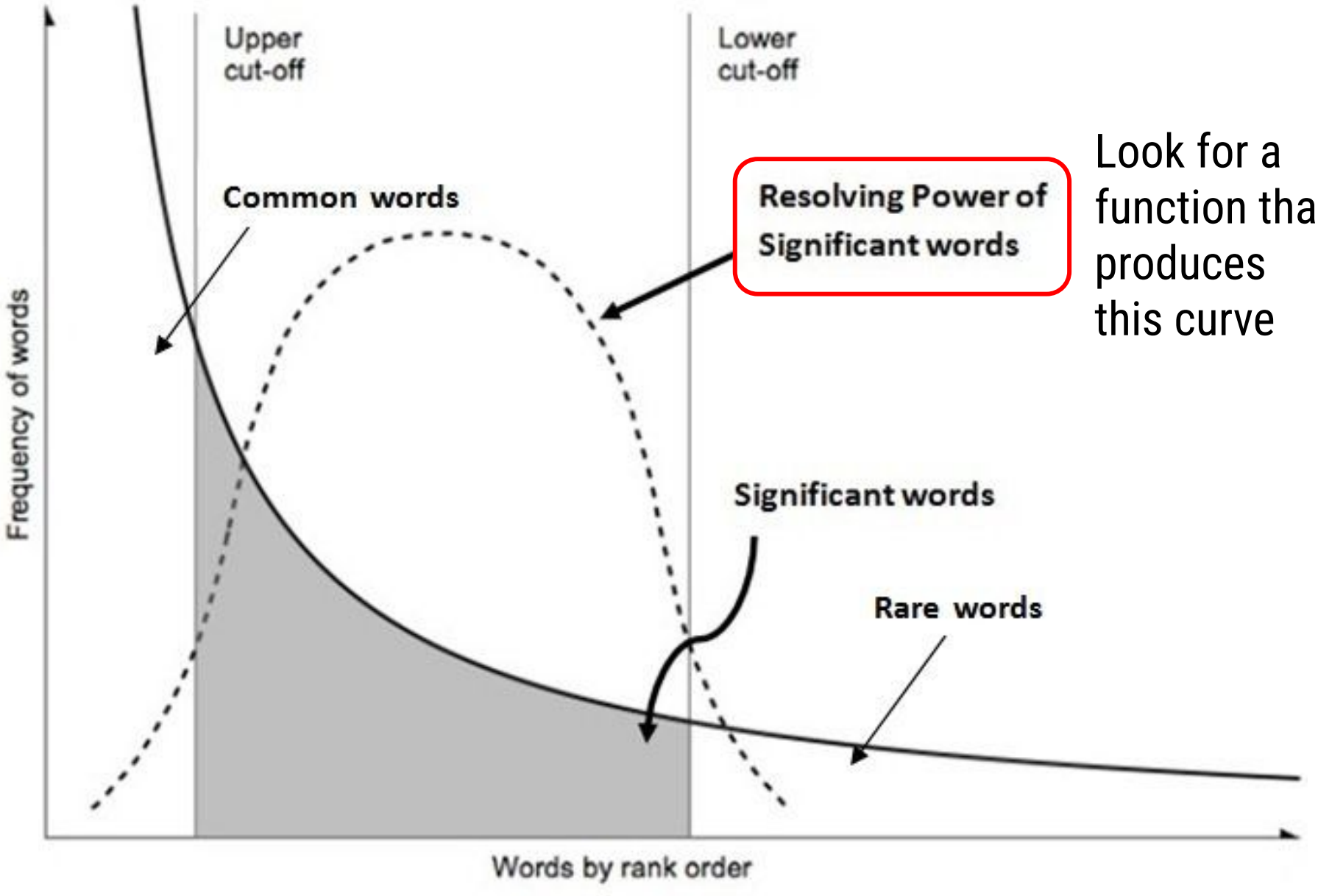


Figure 1 **Communication of ideas.**
Breakdown of basic idea into elementary concepts on experiential level common to reader and writer.

Importance of OHCO

In the process of communicating ideas, an author pursues a certain plan of organizing his ideas. The external evidence of such a plan is the grouping of his ideas into chapters, paragraphs, and sentences. Figure 1 illustrates how this organization may come about. Notions are most closely and specifically related to each other within a sentence. One sentence immediately following another might either be related in its entirety to previous notions or serve to relate these notions to new ones. The same might be said of succeeding sentences. However, a significant new argument is usually introduced in a new paragraph. A still more decisive change of aspects might be denoted by the start of a new chapter.

Back to frequency ...



Look for a function that produces this curve

Luhn's concept of **significant words** (Luhn 1958)

Resolving power means

the ability of words to **discriminate** content

What makes a document **stand out** from other documents

A means by which to **summarize** documents

Global Term Frequency

Global Frequency

Global term frequency means **corpus frequency CF**

Recall that the he Culturomics School is based on global **ngram frequencies** over time

- Texts are unbundled in the NGram Viewer
- Here we are looking a frequencies **within** documents as well

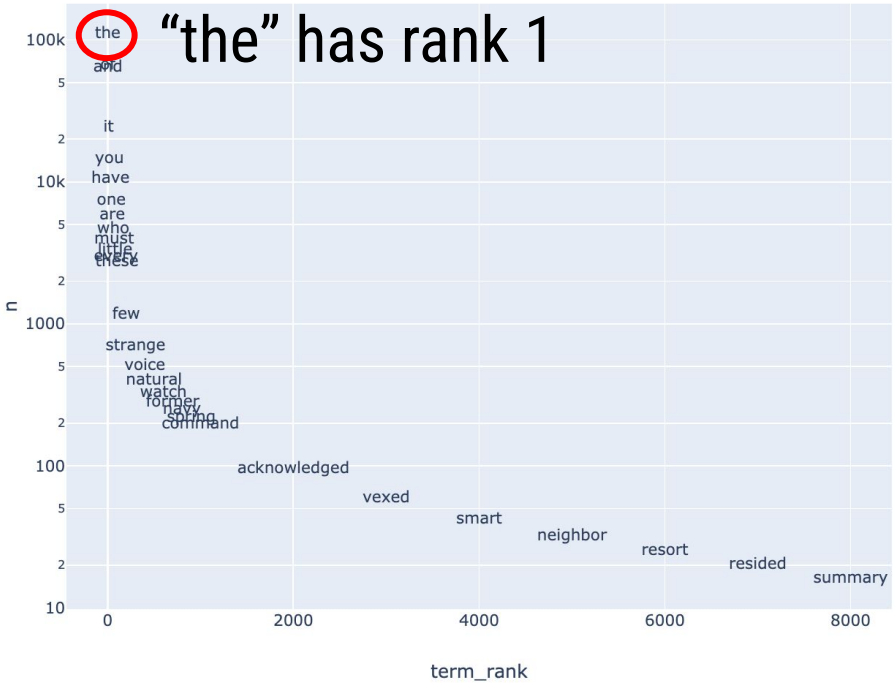
Since many in-document measures of significance use global frequencies, let's look at this first

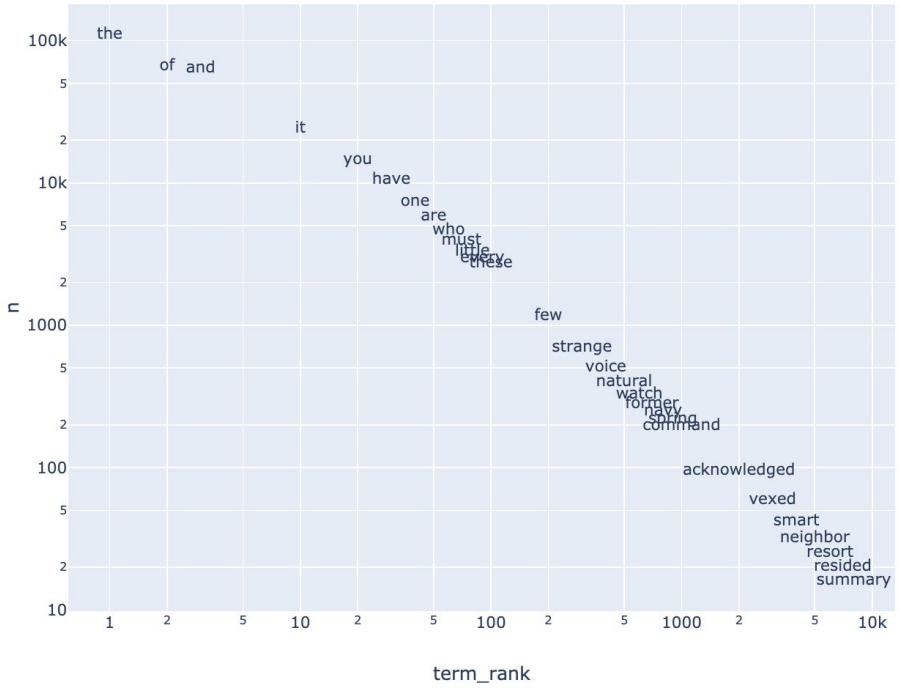
- The most basic fact of global frequency **is Zipf's Law ...**

Zipf's Law

The frequency (f) of a word is inversely proportional to its rank (r), i.e. there is theoretically a constant (k) that is the product of f and r

$$f \propto \frac{1}{r} \qquad k = fr$$





Zipf's Law

One of the **first statistical patterns** discovered in corpus linguistics

- A **power law** distribution of terms by frequency
  George Kingsley Zipf in 1935, "The Psychology of Language"

Universal but **not well understood**

- May have to do with **network theory** — power laws describe scale-free networks
- Related to **Pareto's 80/20** rule
  - 80% effect due to 20% of causes

term_str	term_rank	n	zipf_k	pos_max
the	1	110093	110093	DT
of	2	65993	131986	IN
and	3	63528	190584	CC
it	10	23697	236970	PRP
you	20	14466	289320	PRP
so	30	9843	295290	RB
their	40	7118	284720	PRP$
what	50	5739	286950	WP
upon	60	4504	270240	IN
into	70	3899	272930	IN
other	80	3198	255840	JJ
after	90	2935	264150	IN
might	100	2638	263800	MD
night	200	1128	225600	NN
passed	300	689	206700	VBN
means	400	498	199200	NNS
lost	500	394	197000	VBN
mere	600	322	193200	JJ
scene	700	279	195300	NN
women	800	243	194400	NNS
unless	900	218	196200	IN
assured	1000	195	195000	VBD
cosmopolitan	2000	95	190000	NN
inconvenience	3000	59	177000	NN
fifth	4000	43	172000	JJ
feeble	5000	32	160000	JJ
plumes	6000	25	150000	NN
mildness	7000	20	140000	NN
templars	8000	17	136000	NNP

Here is a list of words from our collection of novels by Austen and Melville

As `n` (f) decreases, `term_rank` (r) increases

Note that in practice k is not constant but appears like Luhn's function

Note that **verbs and nouns** (not PRPs) don't start appearing until r is **200** (in this list)

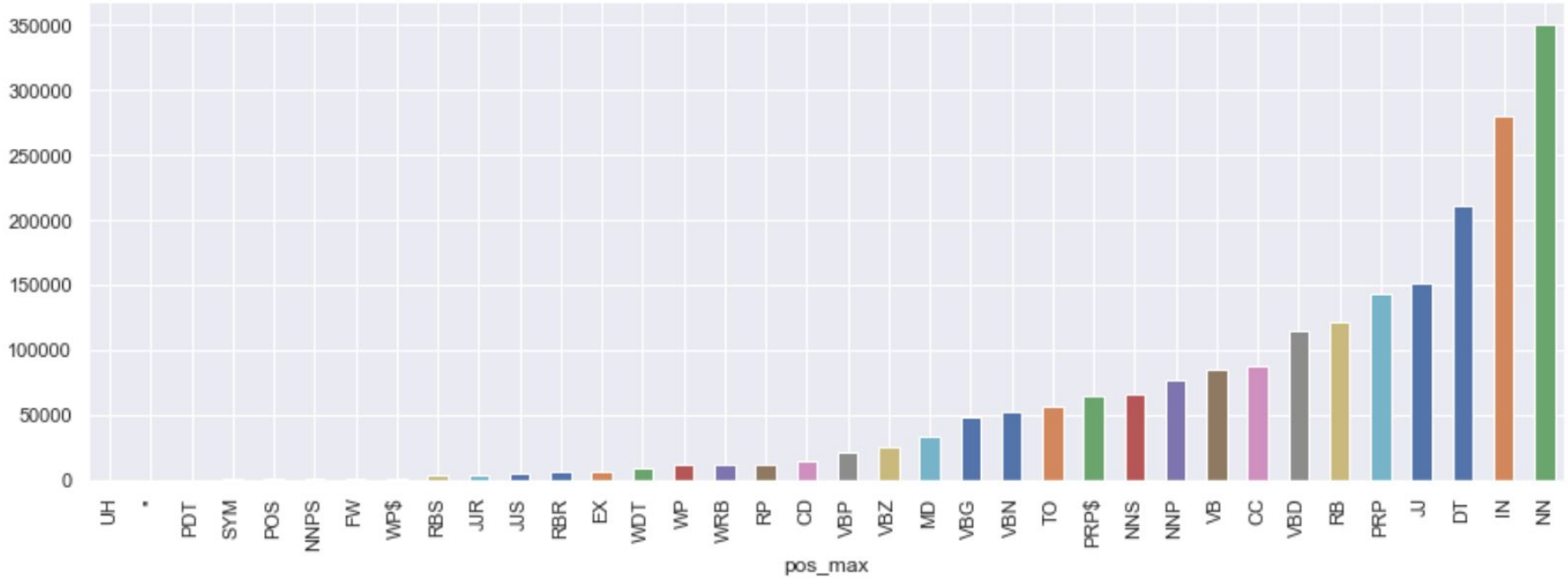
Zipf's law may help distinguish **closed** and **open** class words (function words vs nouns and verbs)

closed class
open class

Parts-of-speech can be divided into two broad supercategories: **closed class** types and **open class** types. Closed classes are those with relatively fixed membership, such as prepositions—new prepositions are rarely coined. By contrast, nouns and verbs are open classes—new nouns and verbs like *iPhone* or *to fax* are continually being created or borrowed. Any given speaker or corpus may have different open class words, but all speakers of a language, and sufficiently large corpora, likely share the set of closed class words. Closed class words are generally **function words** like *of*, *it*, *and*, or *you*, which tend to be very short, occur frequently, and often have structuring uses in grammar.

function word

Note that **discriminant** words are going to be **open** class

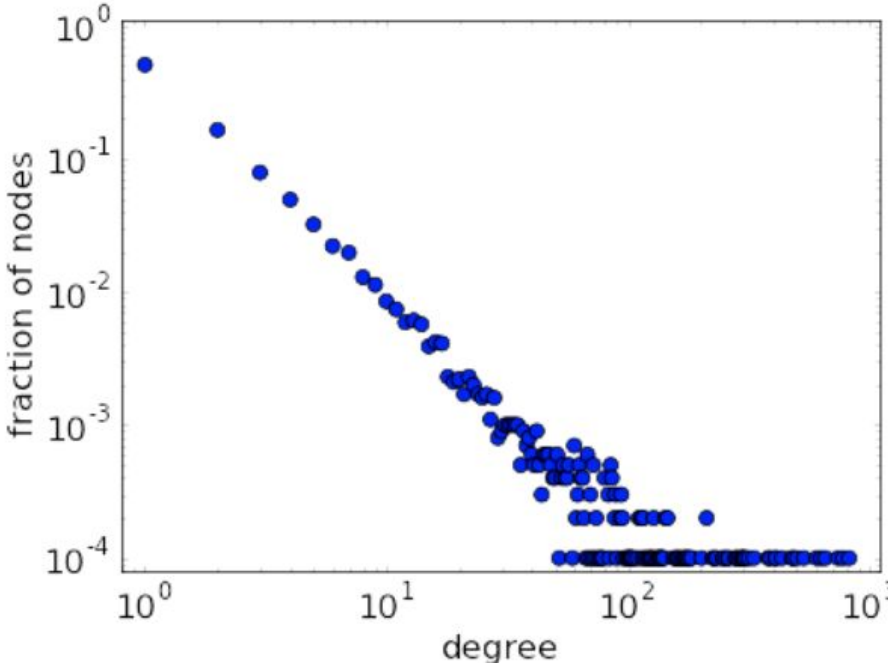


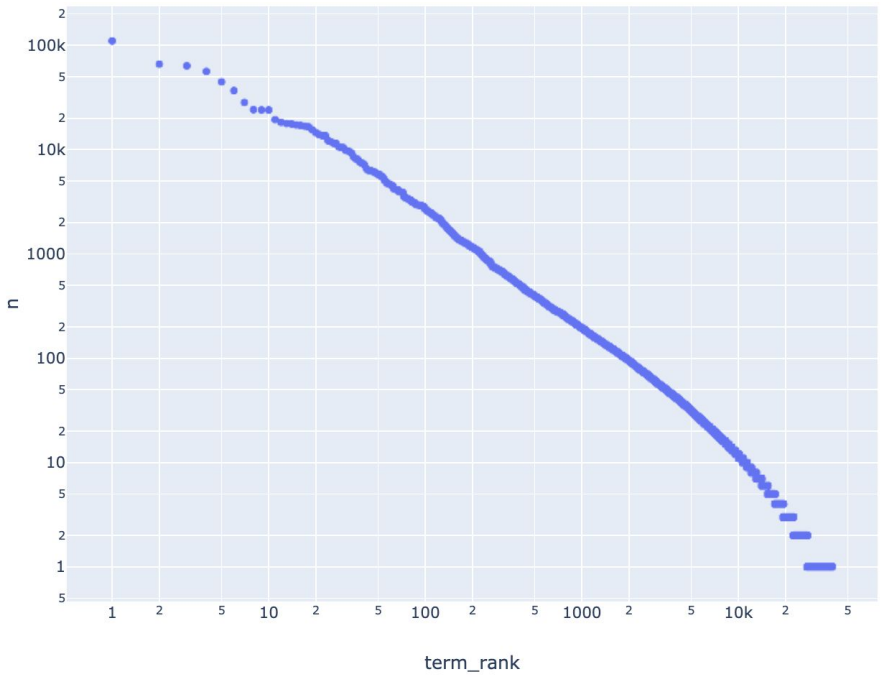
term_rank	term_str	n	num	stop	p_stem	pos_max	zipf_k
68	man	3923	0	0	man	NN	266764
74	time	3446	0	0	time	NN	255004
122	nothing	2164	0	0	noth	NN	264008
124	day	2106	0	0	day	NN	261144
126	way	2072	0	0	way	NN	261072
138	thing	1748	0	0	thing	NN	241224
141	sea	1695	0	0	sea	NN	238995
153	house	1471	0	0	hous	NN	225063
156	however	1440	0	0	howev	NN	224640
157	indeed	1414	0	0	inde	NN	221998

Nouns are the most frequent POS

Given that there are a large number of nouns, they may have resolving power

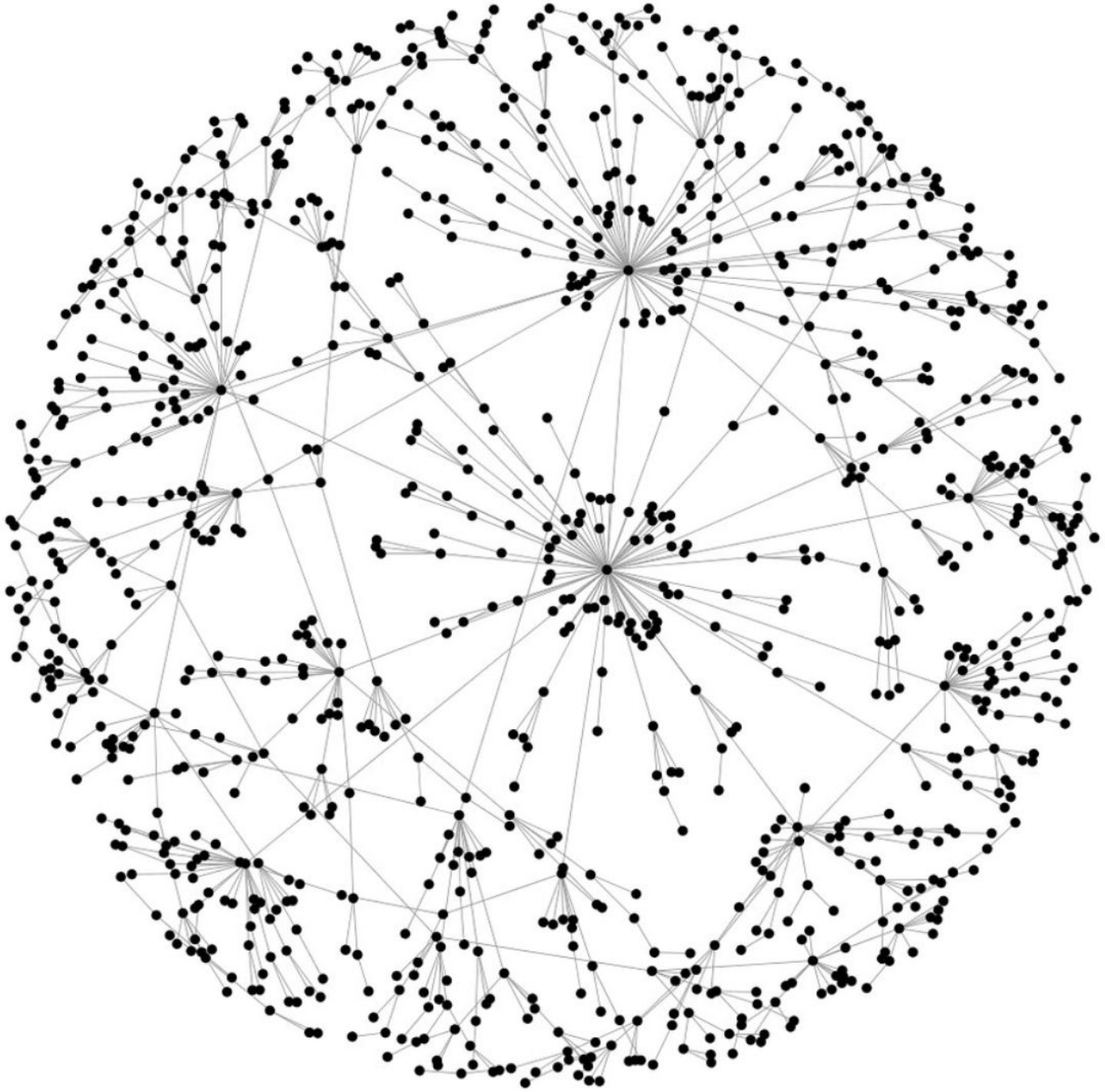
Comparing Zipf's Law with Scale-Free Networks





From Math Insight, "Scale-Free Networks."

This may be evidence for the theory that **language is organized as a network** in the brain — like the **semantic network** we saw earlier



A scale-free network has many **dense nodes** to which other nodes **preferentially attach**

As the network grows, these nodes grow at a slower rate

Maybe language is like this

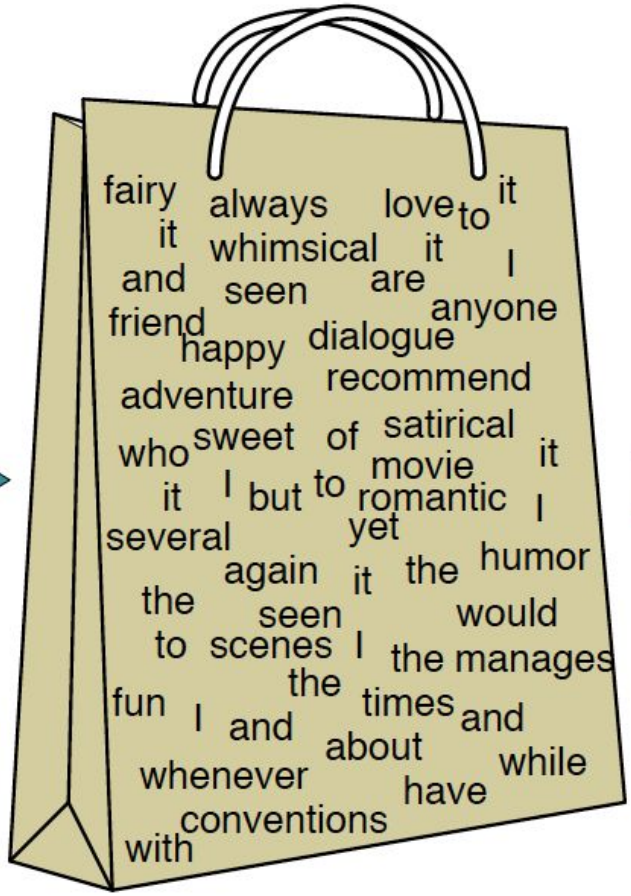
Zipf's law is interesting, but it does not take advantage of the **OHCO structure** of texts within a corpus

The Bag of Words representation
takes advantage of this information

A "**bag**" is a **document**, or **content object**

The Bag of Words Hypothesis

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
…	…

1

chap_num	para_num	sent_num	token_num	term_str
1	1	0	0	sir
			1	walter
			2	elliot
			4	of
			5	kellynch
			6	hall
			8	in
			9	somersetshire

2

chap_num	para_num	term_str	term_str
1	1	a	2
		admiration	1
		affairs	1
		almost	1
		always	1
		amusement	1
		an	2
		and	4
		any	2
		arising	1
		as	1
		at	1

3

chap_num	para_num	term_str	1	15	16	1760	1784	1785	1787	1789	1791	1800	...	your	yours	yourself	yo
1	1		0	0	0	0	0	0	0	0	0	0	...	0	0	0	
	2		0	0	0	0	0	0	0	0	0	0	...	0	0	0	
	3		2	1	0	1	1	1	1	1	1	1	...	0	0	0	
	4		0	0	1	0	0	0	0	0	0	0	...	0	0	0	
	5		0	0	0	0	0	0	0	0	0	0	...	0	0	0	

Bag of Words Hypothesis — Intuition

The hypothesis is based on the **intuition** that even though we **lose word order** (and thus syntax), words in document are **selected** on the basis of the **topic** or theme of the document

- "Guilt by association" principle
- There is some **cognitive mechanism** that is choosing words based on what is being talked about
- This mechanism tends to choose **similar words given similar topics**
- Follows from original hypothesis regarding Signifier and Signified

This intuition is developed more fully with **topic modeling**

Bag of Words Hypothesis — Generalization

More generally, this hypothesis applies to all situations in which items are **selected and collected in baskets** of some kind

- **Documents are like shopping baskets** and terms are items
- Each document represents a set of motivated **choices**

Therefore, **order does not matter**

- What is captured is **what** is being talked about,
  not so much **how** it is being talked about
- Although **sentiment analysis** may help with that

Bag of Words Hypothesis

Once words are grouped into bags, we can **observe statistics** at the bag level

The relative frequency in which a word appears in a bag (TF)

$$p(w \mid d)$$

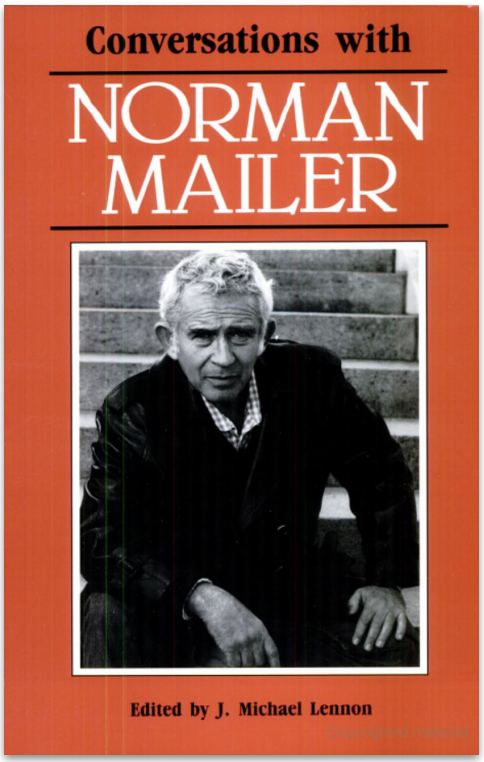
The relative frequency of bags in which a word appears (DF)

$$p(d \mid w)$$

The relative frequency of a word co-occurring with another word

$$p(w_1, w_2 \mid d)$$

A few words from the novelist on how to make movies: "In the novel, you start with a bag of words, and the way you put the words together has meaning. In movies, you get another kind of vocabulary. You have little bits of film strip, each the equivalent of words—five words or five thousand. But the strips are put together as individual words. Constructing a movie out of the field of experience that you have recorded gets to be wholly fascinating because you are working with a brand new vocabulary. You are putting things together in a way that nobody has ever put them together before.



1988 [1969], p. 149

The Latest Model Mailer

Joseph Roddy/1969

From *Look*, 27 May 1969, 23-28. Reprinted by permission of Joseph Roddy.

Dada Science :-)

How to Make a Dadaist Poem
(method of Tristan Tzara)

To make a Dadaist poem:

- Take a newspaper.
- Take a pair of scissors.
- Choose an article as long as you are planning to make your poem.
- Cut out the article.
- Then cut out each of the words that make up this article and put them in a bag.
- Shake it gently.
- Then take out the scraps one after the other in the order in which they left the bag.
- Copy conscientiously.
- The poem will be like you.
- And here are you a writer, infinitely original and endowed with a sensibility that is charming though beyond the understanding of the vulgar.

--Tristan Tzara

The BOW representation has several **uses**

One is to generate **co-occurrence networks**

Another is to compute statistics such as **TF-IDF** and other measures of **significance**

It is also the foundation for creating a common **vector space representation** of a text, which is in the foundation for creating **topic models**, etc.

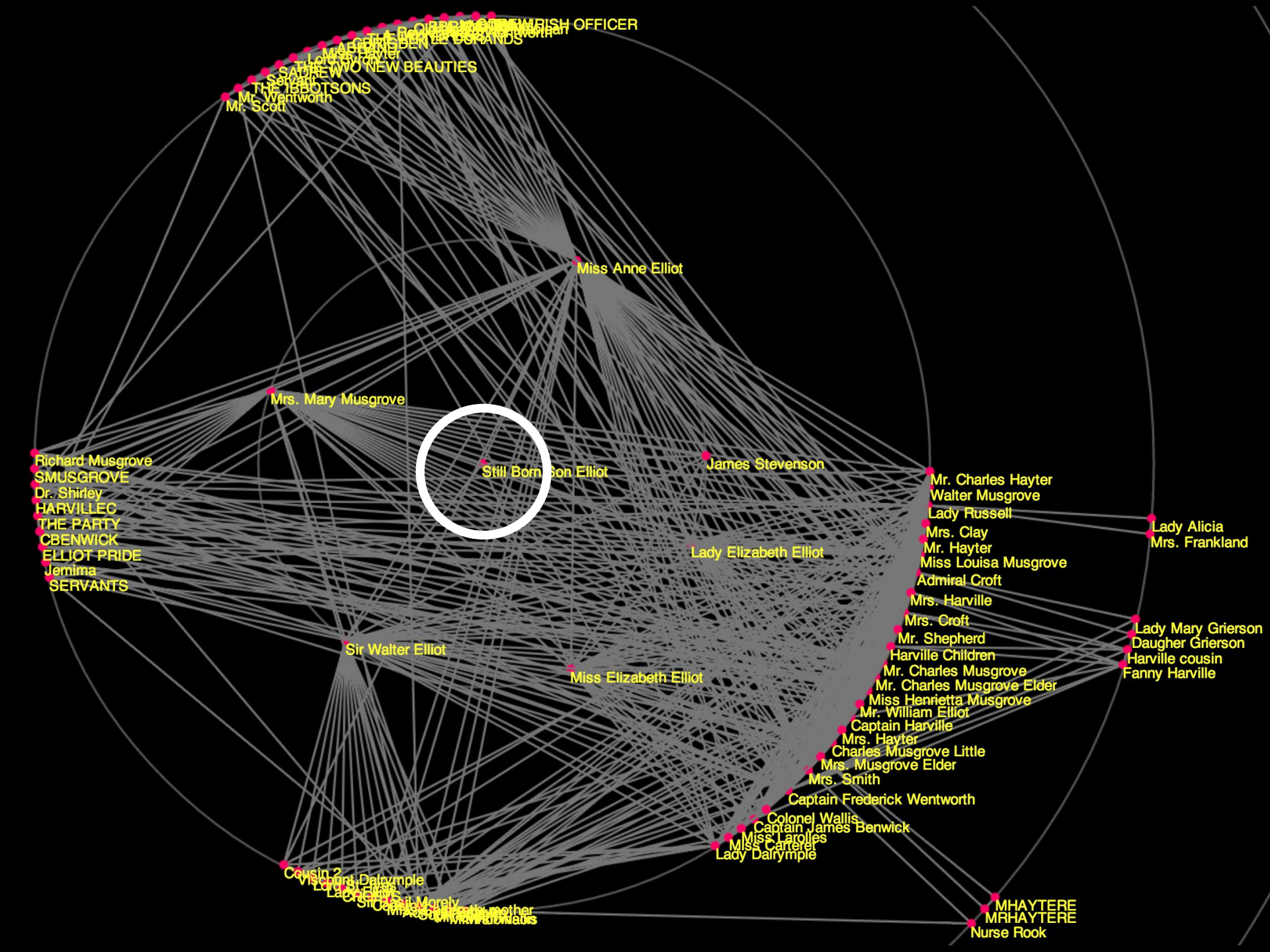
"Bags" are containers

Think OHCO

Containers are co-occurrence contexts

Co-occurrence is a property of the BOW representation that allows us to move from frequency to meaning in an intuitive way

(I call this "common container correlation")


Miss Anne Elliot
Mrs. Mary Musgrove
Still Born Son Elliot
James Stevenson
Lady Elizabeth Elliot
Sir Walter Elliot
Miss Elizabeth Elliot
Richard Musgrove
SMUSGROVE
Dr. Shirley
HARVILLEC
THE PARTY
CBENWICK
ELLIOT PRIDE
Jemima
SERVANTS
Mr. Charles Hayter
Walter Musgrove
Lady Russell
Mrs. Clay
Mr. Hayter
Miss Louisa Musgrove
Admiral Croft
Mrs. Harville
Mrs. Croft
Mr. Shepherd
Harville Children
Mr. Charles Musgrove
Mr. Charles Musgrove Elder
Miss Henrietta Musgrove
Mr. William Elliot
Captain Harville
Mrs. Hayter
Charles Musgrove Little
Mrs. Musgrove Elder
Mrs. Smith
Captain Frederick Wentworth
Colonel Wallis
Captain James Benwick
Miss Larolles
Miss Carteret
Lady Dalrymple
Lady Alicia
Mrs. Frankland
Lady Mary Grierson
Daugher Grierson
Harville cousin
Fanny Harville
Mr. Wentworth
Mr. Scott
Servant
THE ABBOTSONS
TWO NEW BEAUTIES
Cousin 2
Viscount Dalrymple
Sir Josiah Morely
MHAYTERE
MRHAYTERE
Nurse Rook

TF-IDF

Improving on Frequency

We have seen that term **frequencies are not the best indicator** of a term's significance in a document

> **Stop words** and common words are frequent,
> but not indicative of a specific theme or topic

One common way to solve this problem is to **weight** the terms in the document by multiplying their raw frequencies by the inverse of their frequency in the document corpus as a whole

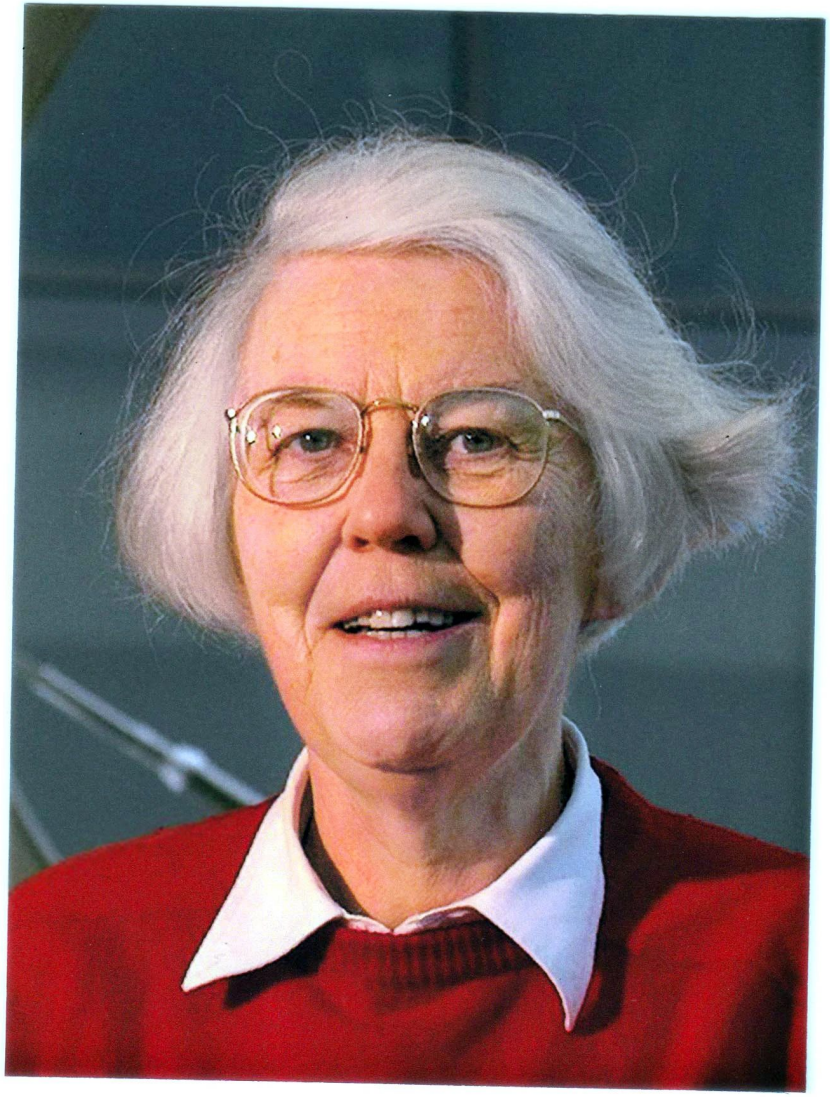
> The is measure is called **Term Frequency Inverse Document Frequency**, or **TF-IDF**

Origins of the Idea

Invented by **Karen Spärck-Jones** in 1972 at Cambridge University

See optional reading:

Spärck-Jones, Karen. 1972. "A statistical interpretation of **term specificity** and its application in retrieval." *Journal of Documentation,* 28(1): pp. 11-21.



The **exhaustivity** of document descriptions and the **specificity** of index terms are usually regarded as independent.

It is suggested that **specificity should be interpreted statistically, as a function of term use rather than of term meaning**.

...

It is argued that **terms should be weighted according to collection frequency**, so that **matches on less frequent, more specific, terms are of greater value than matches on frequent terms**.

...

Exhaustivity

The extent to which a document's content is **comprehensively** completely described by the terms within it.

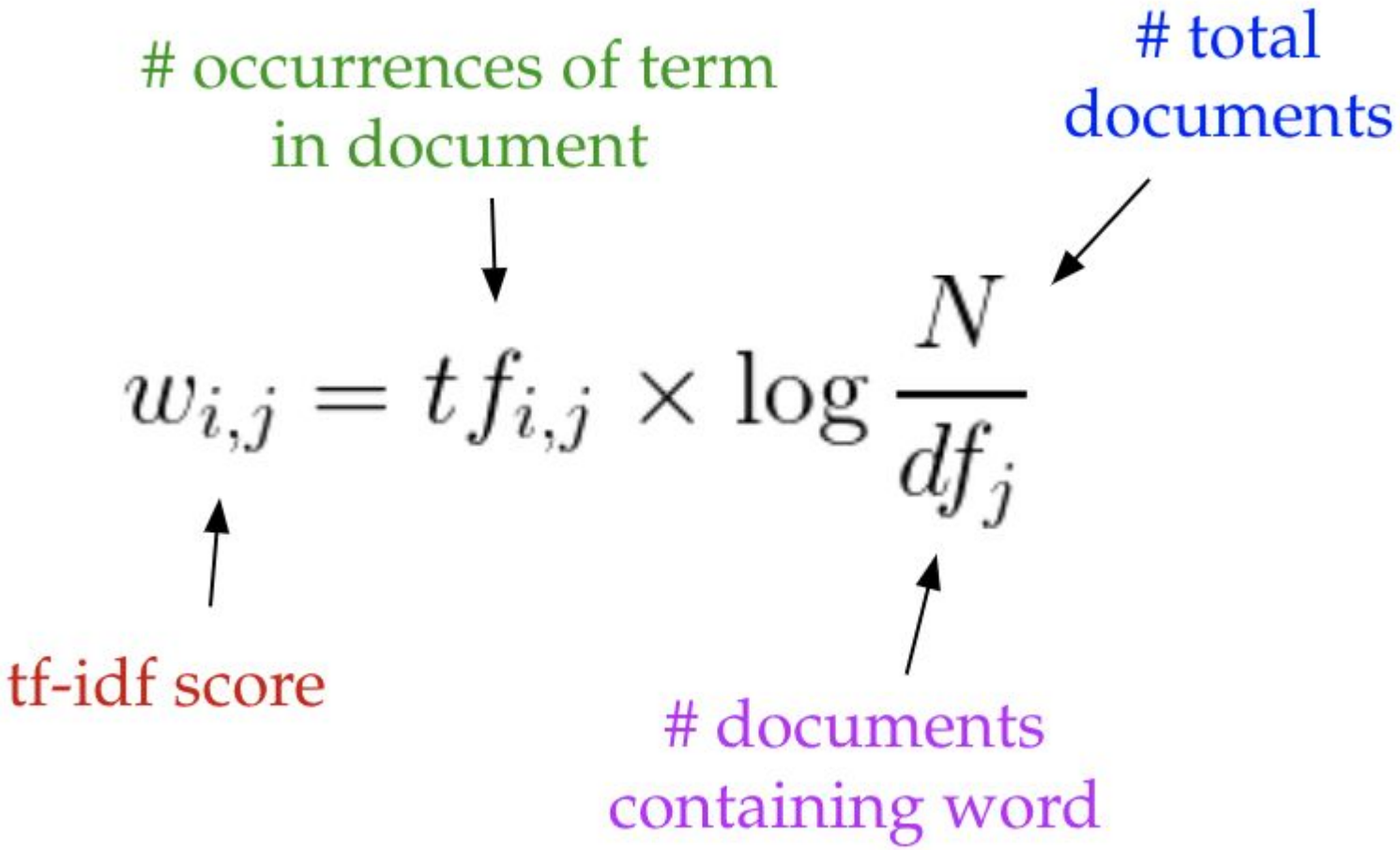
Specificity

How **uniquely or specifically** a term describes the document's content within the context of the corpus

→ Interpret **statistically**

Similar to Luhn's concern for resolving power

TF-IDF



$tf_{i,j}$ = term j count in document i

df_j = term j count in documents (binary)

IDF is just the Information of the word over documents

$$P(d \mid w_j) = df_j / N$$

$$1/P = N/df_j$$

$$\log \frac{N}{df_j}$$

$$I(P) = \log_2(1/P)$$

*So, TFIDF is just the frequency of a word in a document **weighted** by that word's information value over the corpus*

Is TF-IDF Cross Entropy?

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

$p(x) = tf_{i,j} / M_i$

relative frequency of a term in a document

$q(x) = df_j / N$

relative binary frequency of the term in the collection

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_j}$$

i = the doc

j = the term

N = total number of docs

M_i = total number of terms in i

$tf_{i,j}$ = term j count in document i

df_j = term j count in documents

Not really -- the two probabilities are from **different event spaces**

$DF = P(d|w)$ and $TF = P(w|d)$

DF is binary, TF usually not

TFIDF is generally regarded as effective but untheorized (heuristic)

(But clearly there is some relationship to entropy)

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \lvert\{d \in D : t \in d\}\rvert$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

There are **many ways** to compute TF and IDF

The simple methods work well

We will look at these in lab

From Wikipedia, "tf-idf"

Uses

Improved search

Treat query as sentence, TF-IDF as language model for document
Rank results

Feature extraction

TFIDF (e.g. sum) stats can be applied to documents

Dimensionality reduction

Reduce the vocabulary to small N (e.g. 40,000 $\longrightarrow$ 4000)

Content analysis

Explore semantics of significant nouns and verbs

TF-IDF and the Data Model

TF-IDF is **per-document term measure**

- It is not a feature of the VOCAB table *per se*
- Instead, it is a feature of the BOW table (i.e. collapsed TOKEN)

As a **measure of terms** in the VOCAB table, we may take **aggregated** values of TF-IDF

- SUM and MEAN

This helps with developing the **language model** for a given corpus

- TF-IDF can be compared and combined with term probability

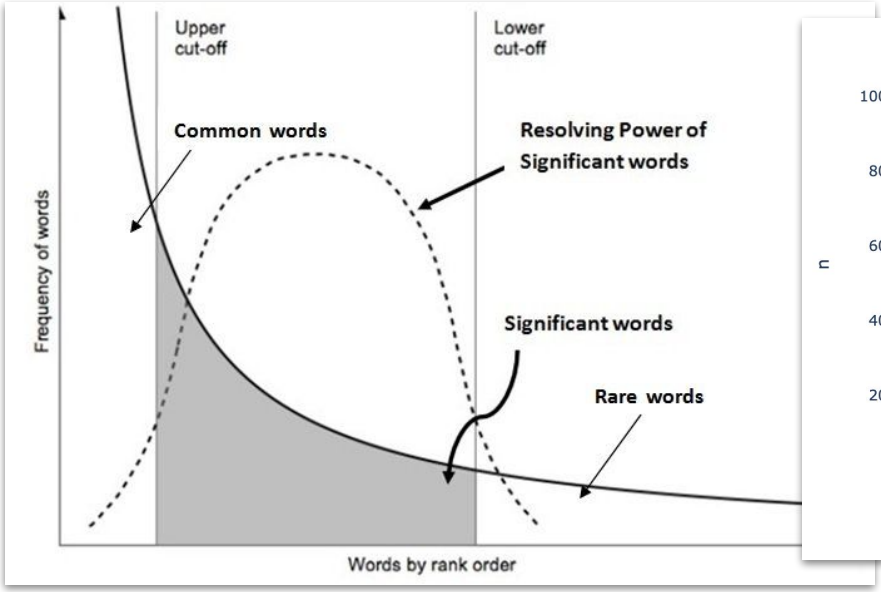
Examples

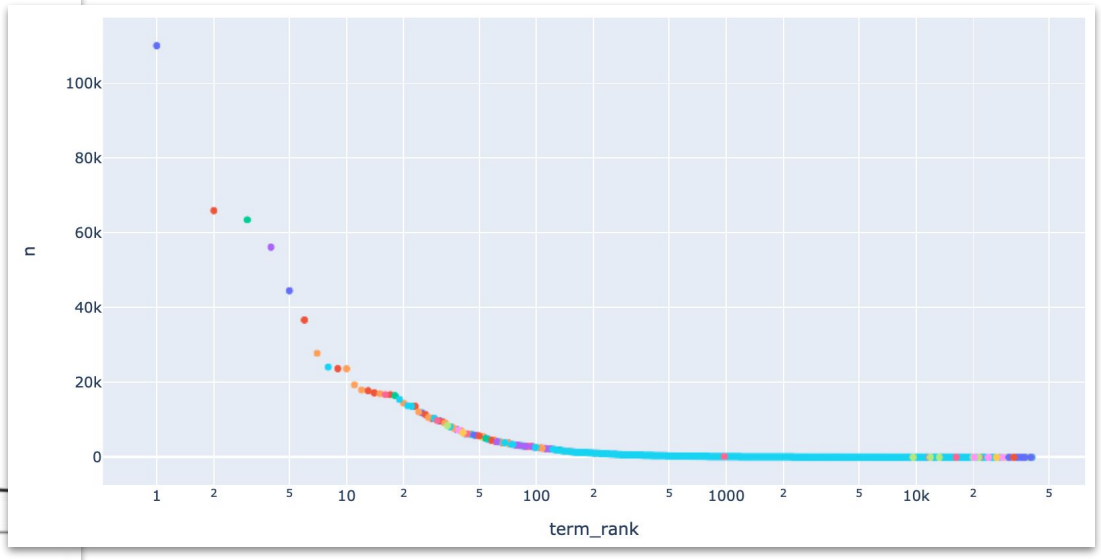
Most Significant Words by TF Method

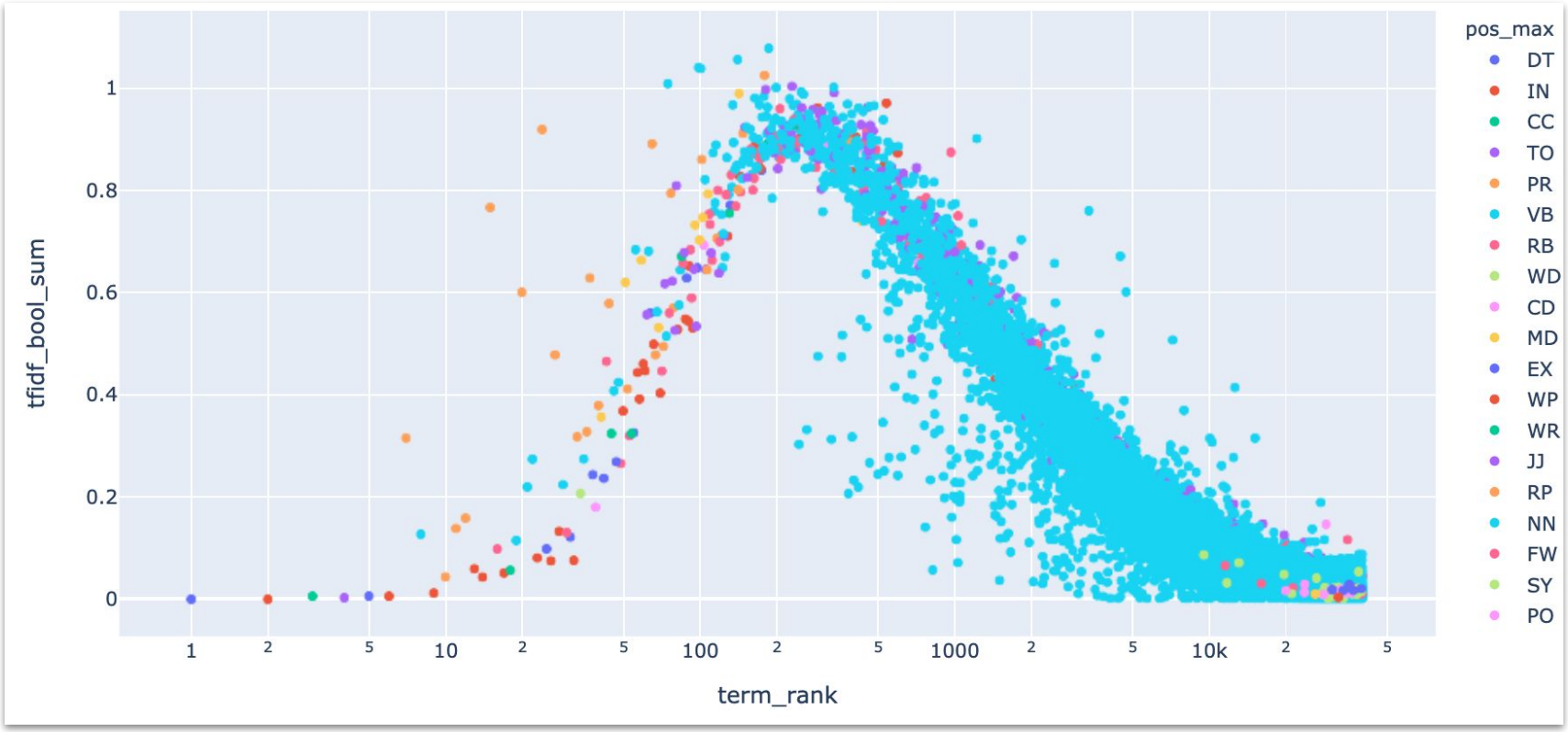
	0	1	2	3	4	5	6
tfidf_n_mean	gee/NN	lombardo/NN	bug/NN	oberlus/NN	hunilla/NN	steelkilt/NN	ugh/NN
tfidf_n_sum	mr/NN	pierre/NN	mrs/NN	miss/NN	fanny/NN	emma/NN	whale/NN
tfidf_cp_mean	um/NN	gee/NN	plujii/NN	ugh/NN	dunderfunk/NN	egbert/NN	southampton/NN
tfidf_cp_sum	pierre/NN	mr/NN	whale/NN	mrs/NN	thou/NN	um/NN	babbalanja/NN
tfidf_cpmax_mean	ugh/NN	gee/NN	plujii/NN	lombardo/NN	um/NN	hvalt/NN	grammars/NN
tfidf_cpmax_sum	pierre/NN	mr/NN	mrs/NN	thou/NN	miss/NN	babbalanja/NN	thee/NN
tfidf_logn_mean	gee/NN	lombardo/NN	bug/NN	oberlus/NN	hunilla/NN	steelkilt/NN	gees/NN
tfidf_logn_sum	mr/NN	mrs/NN	miss/NN	pierre/NN	sir/NN	captain/NN	ship/NN
tfidf_l2_mean	gee/NN	um/NN	ugh/NN	plujii/NN	dunderfunk/NN	lombardo/NN	gees/NN
tfidf_l2_sum	pierre/NN	mr/NN	mrs/NN	thou/NN	whale/NN	babbalanja/NN	media/NN
tfidf_bool_mean	55/CD	whys/NN	naïvely/RB	um/NN	altho/NN	gluepots/NN	hooroosh/NN
tfidf_bool_sum	lady/NN	sea/NN	mrs/NN	mr/NN	white/JJ	soul/NN	night/NN

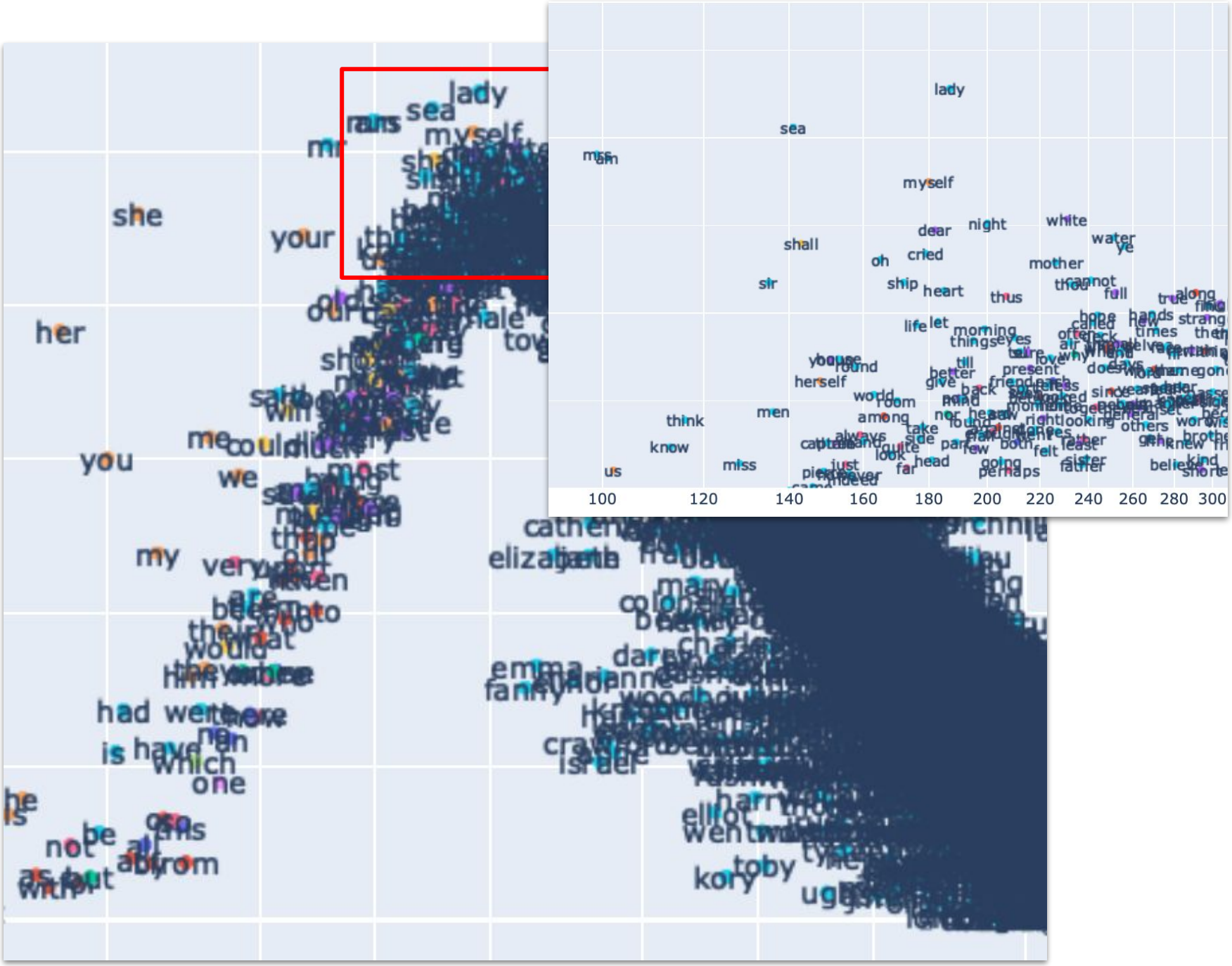
term_str	term_rank	pos_max	n	df	tfidf_bool_mean	tfidf_bool_sum	tfidf_n_mean	tfidf_n_sum
the	1	DT	110093	1122	0.000000	0.000000	0.000000	0.000000
and	3	CC	63528	1120	0.000005	0.005919	0.145998	163.517493
i	7	PR	27810	1007	0.000313	0.315549	4.308451	4338.610084
it	10	PR	23697	1107	0.000039	0.043503	0.415660	460.135403
you	20	PR	14466	878	0.000684	0.600749	5.828906	5117.779036
so	30	RB	9843	1074	0.000122	0.130525	0.578104	620.883473
their	40	PR	7118	963	0.000394	0.379156	1.629564	1569.269676
what	50	WP	5739	972	0.000379	0.368514	1.222457	1188.228139
upon	60	IN	4504	928	0.000496	0.460615	1.329243	1233.537349
into	70	IN	3899	950	0.000425	0.403429	0.985311	936.045631
other	80	JJ	3198	879	0.000599	0.526342	1.281156	1126.136062
after	90	IN	2935	863	0.000631	0.544599	1.287728	1111.309021
might	100	MD	2638	751	0.000936	0.702929	2.034484	1527.897583
night	200	NN	1128	490	0.002043	1.001279	2.751443	1348.207056
passed	300	VB	689	416	0.002176	0.905146	2.370785	986.246480
means	400	NN	498	344	0.002559	0.880452	2.469142	849.384919
lost	500	VB	394	284	0.003161	0.897780	2.749828	780.951277
mere	600	JJ	322	237	0.003366	0.797749	3.047606	722.282615
scene	700	NN	279	191	0.003848	0.734909	3.731337	712.685449
women	800	NN	243	160	0.004002	0.640359	4.267579	682.812714
unless	900	IN	218	178	0.003926	0.698758	3.253005	579.034929
assured	1000	VB	195	157	0.004250	0.667204	3.523956	553.261061
cosmopolitan	2000	NN	95	20	0.014561	0.291218	27.597162	551.943242
inconvenience	3000	NN	59	51	0.005540	0.282548	5.158950	263.106465
fifth	4000	NN	43	36	0.007600	0.273601	5.926752	213.363074
feeble	5000	JJ	32	28	0.007431	0.208054	6.085145	170.384065
plumes	6000	NN	25	21	0.010397	0.218337	6.832785	143.488488
mildness	7000	NN	20	19	0.007936	0.150777	6.193610	117.678589
templars	8000	NN	17	4	0.007540	0.030162	34.560392	138.241568

Note how sum of boolean TF-IDF is concentrated in the middle, around **nouns**, **verbs**, and **adjectives**, consistent with our earlier observation


Upper
cut-off
Lower
cut-off
Common words
Resolving Power of
Significant words
Significant words
Rare words
Frequency of words
Words by rank order


n
100k
80k
60k
40k
20k
0
1
2
5
10
2
5
100
2
5
1000
2
5
10k
2
5
term_rank


tfidf_bool_sum
1
0.8
0.6
0.4
0.2
0
1
2
5
10
2
5
100
2
5
1000
2
5
10k
2
5
term_rank
pos_max
DT
IN
CC
TO
PR
VB
RB
WD
CD
MD
EX
WP
WR
JJ
RP
NN
FW
SY
PO


lady
sea
myself
she
your
her
you
me
we
my
had
is
which
one
not
toby
kory
100
120
140
160
180
200
220
240
260
280
300

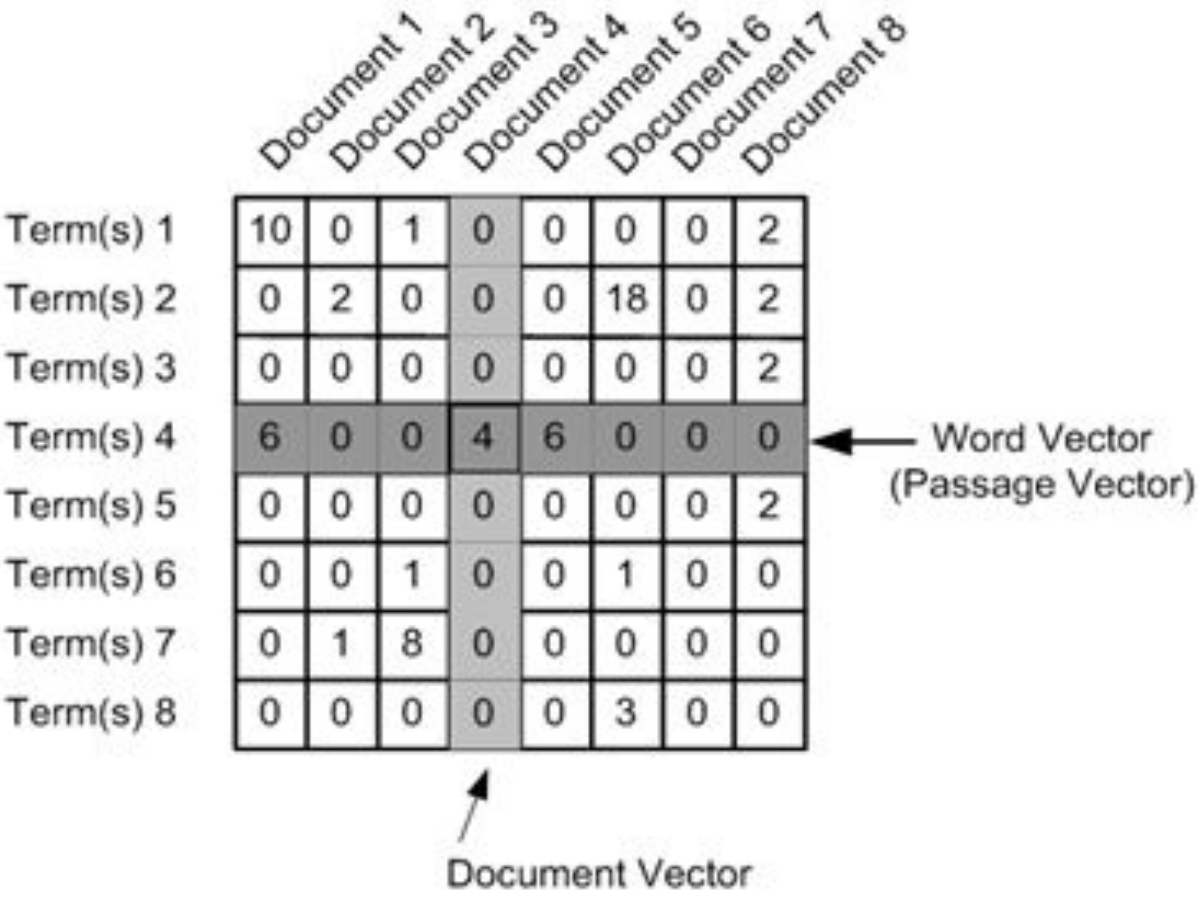
The Vector Space Model

Vector Spaces

Intuitively, a Vector Space is just a **data table** (aka data frame) with certain properties:

> *Data table = columns $\times$ rows = features $\times$ observations*
>
> **Features** consist of a finite set of **abstract attributes** that may appear in combination -- e.g. terms in vocabulary, psychological traits, etc. In VS, they are the **dimensions** of the coordinate system
>
> **Events** consist of **observed frequencies and/or weights** associated with each feature for a specific, concrete instance of a random variable, e.g. a message or a person. In VS, these are **coordinates** — or **vectors**

One way to **represent textual** data in vector space is as a **document-term matrix**



D4 = (0, 0, 0, 4, 0, 0, 0, 0)

In this example, the **observations** (term counts) are represented as **columns** and the **features** as **rows**

Often we will **flip** this and represent the features as columns and observations as rows

Note that each row or column is a **bag of words**

Also, note the diagram also shows a what Turney and Pantel call a **word-context matrix**

The big idea is to regard
terms as random variables and
documents as coordinates in term space

Salton, et al (1975) originally referred to **index** terms and **keywords**, but this has been generalized to entire vocabulary of a corpus

Basically, a document-term matrix
is an **unstacked** bag-of-words table

Representing Vector Spaces

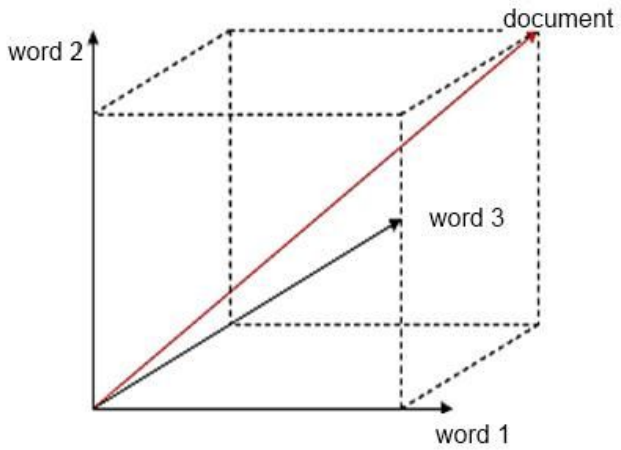
A vector space may be **visually represented** in at least two ways:

As a **table** in which **columns** are **events** and **rows** are **features** (or *vice versa*)

Aa a **coordinate system** in which features correspond to **dimensions** and observations correspond to **points**

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

The **rows** of this table correspond to the **axes** on the 3-D coordinate system



Note that it is **impossible** to visually represent the entire space in 3D geometric space

```
BOW.to_frame().head()
```

chap_num	term_id	n
0	33	1
	70	1
	71	1
	101	2
	131	1

Basically, a document-term matrix is an **unstacked** bag-of-words table

Notice how the DTM has lots of **zeros** — this is a **sparse** matrix

Convert BOW to DTM

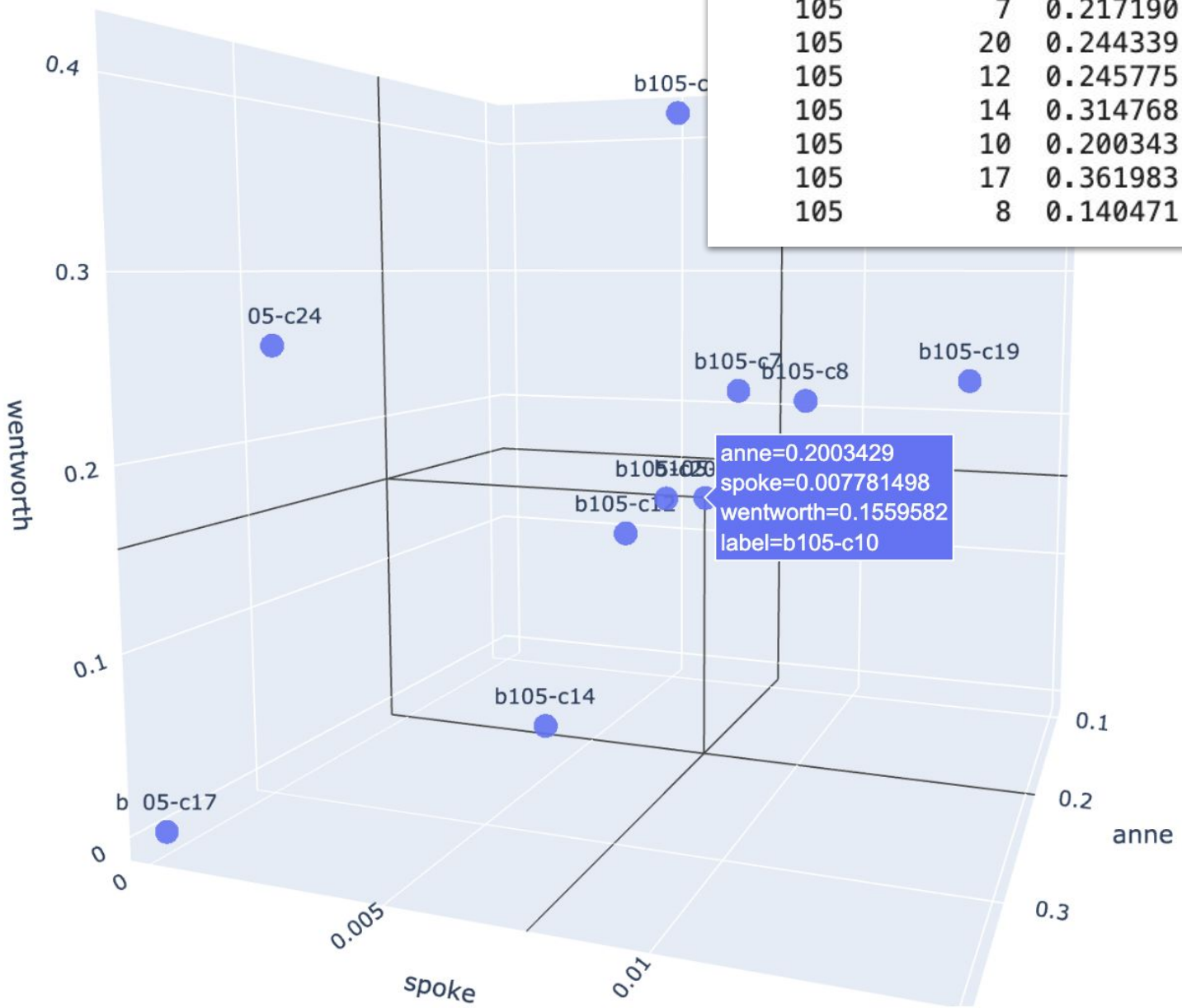
```
DTM = BOW.unstack().fillna(0)
```

```
DTM.head()
```

term_id	1	2	3	4	5	6	7	8	9	10	...	169
chap_num												
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
1	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	

Words become **dimensions**
Documents become **points**

book_id	chap_num	anne	spoke	wentworth
105	19	0.344746	0.014729	0.253034
105	24	0.293783	0.000000	0.256701
105	9	0.094174	0.005030	0.417609
105	7	0.217190	0.008838	0.227731
105	20	0.244339	0.007733	0.166054
105	12	0.245775	0.006848	0.143791
105	14	0.314768	0.006724	0.057758
105	10	0.200343	0.007781	0.155958
105	17	0.361983	0.000000	0.000000
105	8	0.140471	0.009233	0.211491

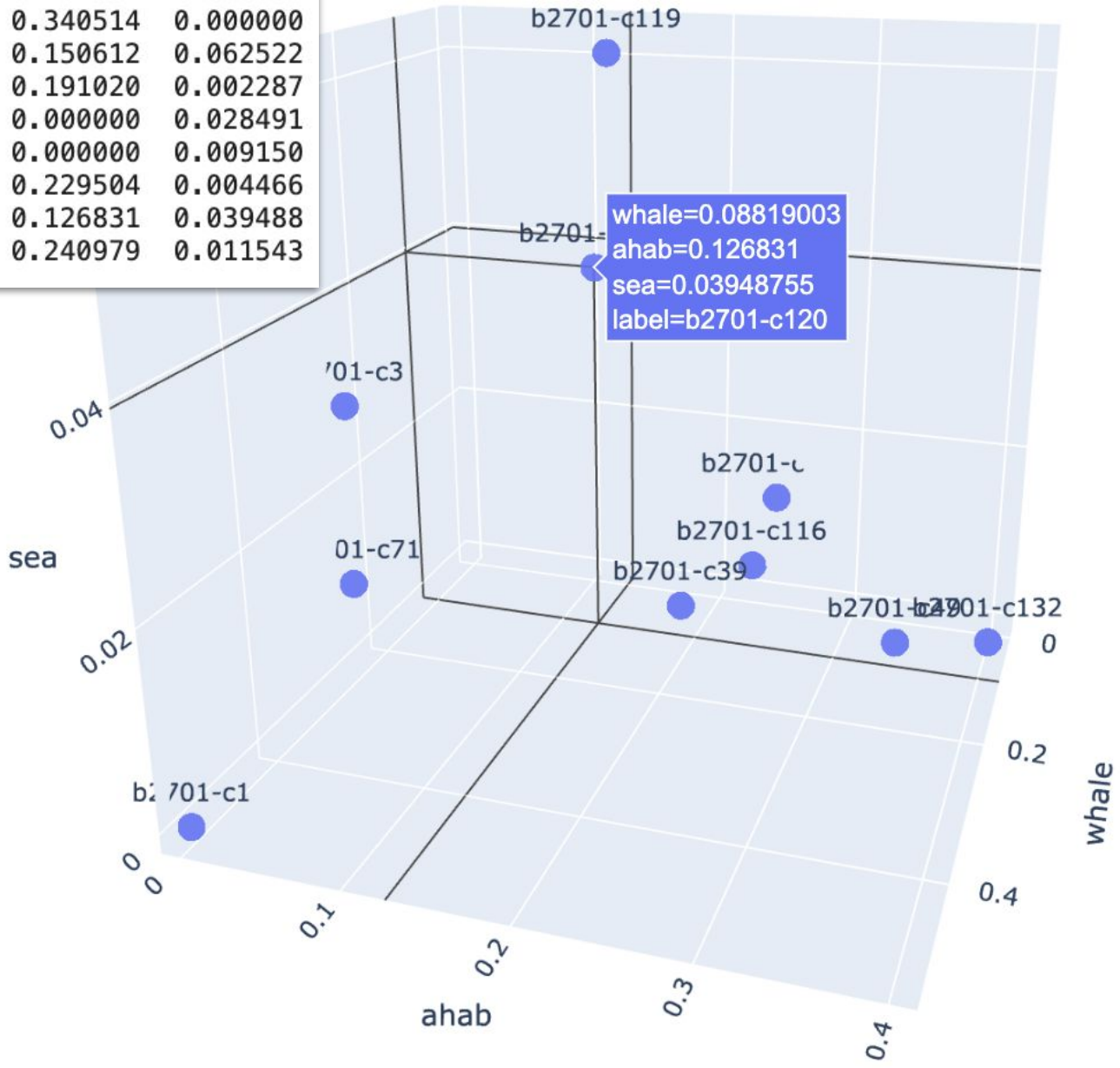


Here is an example of a vector space representation of Jane Austen's *Persuasion*. **Each point represents a document** as a coordinate of three dimensions -- the terms "anne", "wentworth", and "spoke"

book_id	chap_num	whale	ahab	sea
2701	1	0.523628	0.000000	0.000000
2701	132	0.046545	0.401632	0.000000
2701	49	0.072853	0.340514	0.000000
2701	119	0.139634	0.150612	0.062522
2701	39	0.097063	0.191020	0.002287
2701	3	0.250988	0.000000	0.028491
2701	71	0.255428	0.000000	0.009150
2701	116	0.029922	0.229504	0.004466
2701	120	0.088190	0.126831	0.039488
2701	112	0.000000	0.240979	0.011543

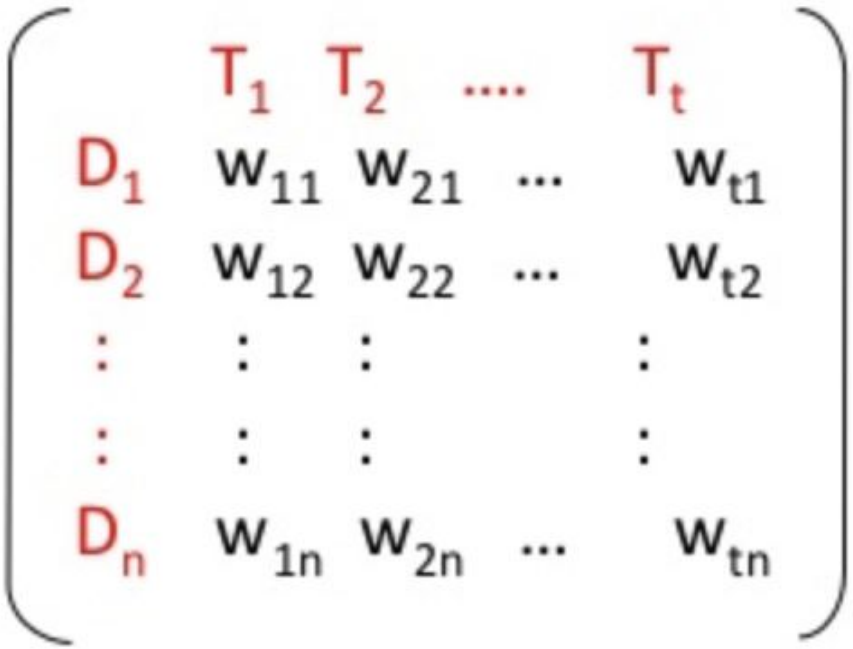
This is a vector space representation of Melville's *Moby Dick*. The dimensions are the terms "ahab", "whale", and "sea"

Note use of OHCO to define our "document"



Vector Spaces as Matrices

Mathematically VSMs are of course just **matrices**:



Varieties of VSM for Text

Axis of Similarity	Matrix Type	Hypothesis	Methods
Documents	Term-Document	Bag of words	LSI, Topic Models
Terms	Word-Context	Bag of words	LSA
Relations	Pair-Pattern	Extended distributional	
Documents (sentences)	Token-Document	Distributional	Word Embedding

	Ever tried. Ever failed.	No matter. Try again.	Fail again. Fail better.
Ever	1	0	0
tried	1	0	0
Ever	1	0	0
failed	1	0	0
No	0	1	0
matter	0	1	0
Try	0	1	0
again	0	1	0
Fail	0	0	1
again	0	0	1
Fail	0	0	1
better	0	0	1

A **token-document matrix**, where rows are features, which are tokens in sequential order

Cells are **binary or one-hot** (as they must be), signalling the occupation of a position in the text by a term instance (i.e. a token)

	Ever tried. Ever failed.	No matter. Try again.	Fail again. Fail better.
Ever	2	0	0
tried	1	0	0
failed	1	0	0
No	0	1	0
matter	0	1	0
Try	0	1	0
again	0	1	1
Fail	0	0	2
better	0	0	1

A **term-document matrix** (a kind of word-context matrix) in which feature rows are terms

Cell values are **counts**, in the bag-of-words model

A Token-Time Matrix

Tokens vs **Narrative Time**
Each column is a token position in the text

	0	1	2	3	4	5	6	7	8	9	...	88939	88940	88941	88942	88943	88944	88945	88946	88947	88948
sir	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
walter	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
elliot	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
of	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0	1	0	0	0	0
kellynch	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
finis	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
gutenberg	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
ebook	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
jane	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
austen	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

anne

wentworth

Uses of VSMs

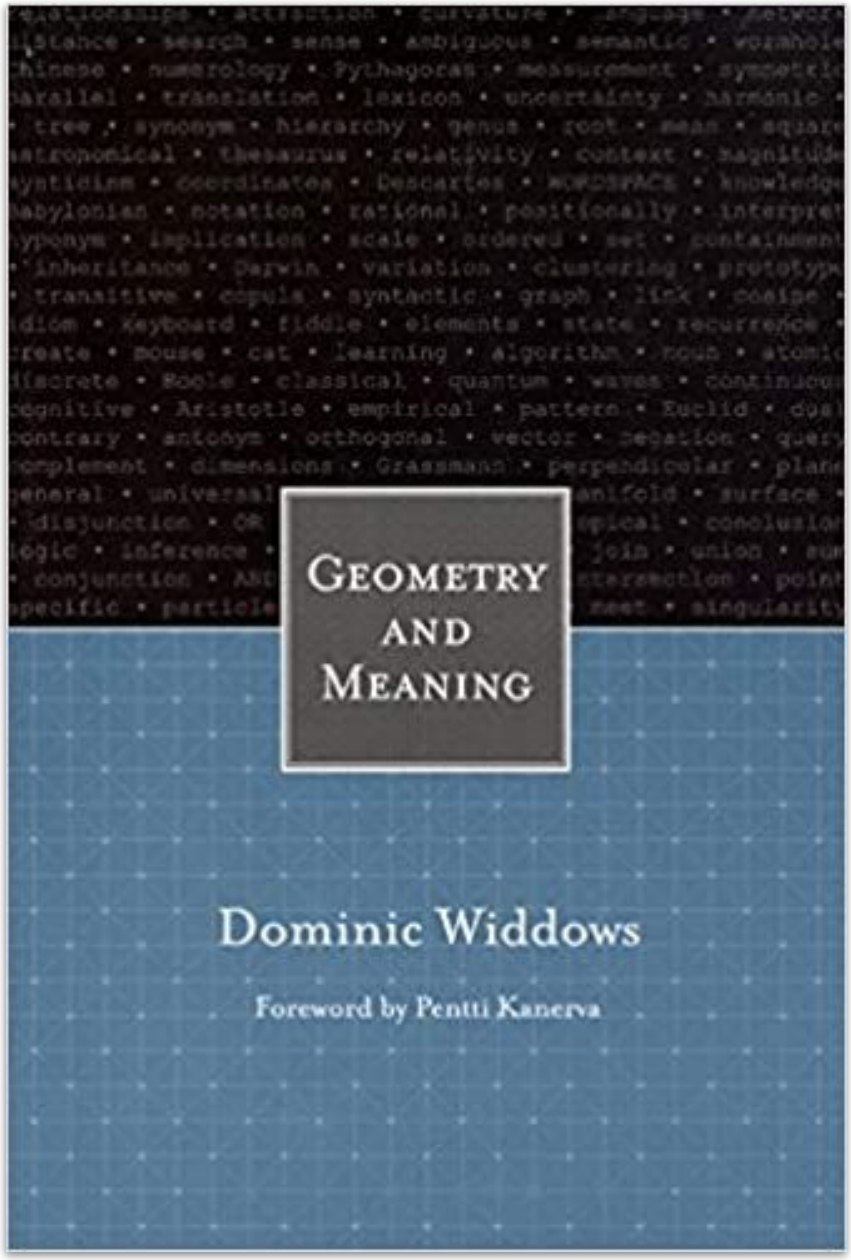
A simple and robust way to convert so-called **unstructured** data into **structured** data

- In data science, structured = tabular

Can easily compute **probabilities** and **statistics** and perform **matrix operations,** e.g. eigendecomposition

Allows for a **geometry of meaning** (Widdows)

- Infer **similarity and difference of documents** by means of geometric **distance measures** within vector space



From Pythagoras's harmonic sequence to Einstein's theory of relativity, geometric models of position, proximity, ratio, and the underlying properties of physical space have provided us with powerful ideas and accurate scientific tools. Currently, similar geometric models are being applied to another type of space—**the conceptual space of information and meaning**, where the contributions of Pythagoras and Einstein are a part of the landscape itself. **The rich geometry of conceptual space can be glimpsed, for instance, in internet documents**: while the documents themselves define a structure of visual layouts and point-to-point links, search engines create an additional structure by matching keywords to nearby documents in a spatial arrangement of content. What the Geometry of Meaning provides is a much-needed exploration of computational techniques to represent meaning and of the conceptual spaces on which these representations are founded.

We get from **frequency** to **meaning**

by way of **geometry**

In **document vector space**, the coordinate system has as many **dimensions** as there are terms in our **vocabulary**

We can **reduce** these dimensions by selecting only significant terms (e.g. by TFIDF)

Then, we can infer the **similarity** of all document **pairs** and then **cluster** documents based on these similarities

We'll look in depth about how to do in the next module

Summary

Zipf's law begins to show how frequency is related to meaning

The **bag of words** (BOW) hypothesis introduces **structure** to frequency which will be the foundation for topic modeling, etc.

TF-IDF takes advantage of OHCO and BOW and estimates significance of words within documents

> **Aggregate** TF-IDF can shed light on **corpus-level** term significance

The **Vector Space Model** represents BOW as a **matrix**

> Documents are represented as coordinates in word space
> Cell values are variable -- raw counts, binary counts, TF-IDF, etc.