Recall: Best-Match Algorithm (Co-ordinated Search)

For each document I, Score (I) = 0;

For each query term

    Search the vocabulary list

    Pull out the postings list

    For each document J in the list,

        Score(J) = Score(J) +1

This basically counts how many terms are in common between a document and a query

Known as Best-Match approach or Co-ordinated Search

3

Contrasting Assumptions

- A term is present in a document or not
  - 0 or 1 (A binary flag)
- It doesn't consider the degree of association between a term and a document
  - How much a document talks 'about' the 'term'?
    - The aboutness notion
      - This captures the semantics
      - If a term appears **often** in a document, then the document is **likely** to be **about** that 'term'
  - Indexing should capture this information

4

Text Statistics

- Huge variety of words used in text **but:**
- Many statistical characteristics of word occurrences are predictable
  - e.g., distribution of word counts
- Retrieval models and ranking algorithms depend heavily on statistical properties of words
  - Document representation: Transforming raw text of documents into a form that represents their meaning
  - Determination of a word's semantic utility based on its statistical properties
  - How can we find meaning in text? What does the distribution of frequency occurrences tell us about the pattern of their use?
    - e.g., **important words occur often in documents but are not high frequency in collection**

5

Plotting Word Frequency by Rank

- Main idea: Count (Frequency)
  - How many times tokens occur in the text
    - Over all documents in the collection
- Now rank these according to how often they occur.

6

Most and Least Frequent Terms

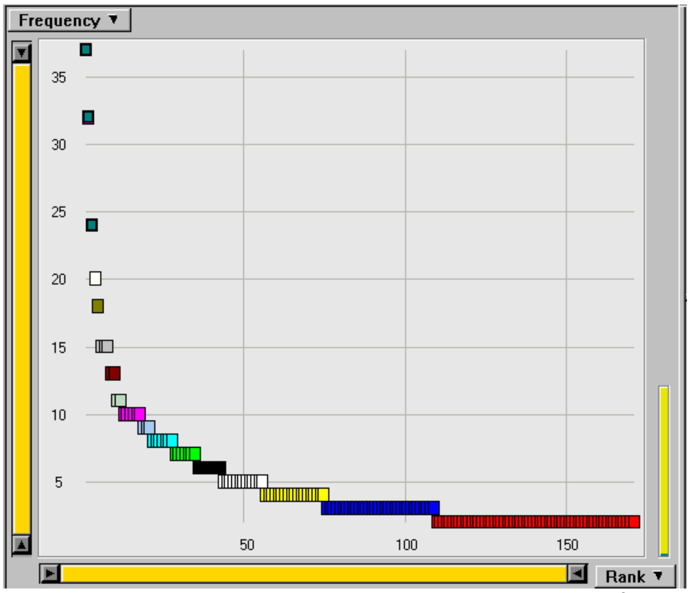
Rank	Freq	Term
1	37	system
2	32	knowledg
3	24	base
4	20	problem
5	18	abstract
6	15	model
7	15	languag
8	15	implem
9	13	reason
10	13	inform
11	11	expert
12	11	analysi
13	10	rule
14	10	program
15	10	oper
16	10	evalu
17	10	comput
18	10	case
19	9	gener
20	9	form

150	2	enhanc
151	2	energi
152	2	emphasi
153	2	detect
154	2	desir
155	2	date
156	2	critic
157	2	content
158	2	consider
159	2	concern
160	2	compon
161	2	compar
162	2	commerci
163	2	clause
164	2	aspect
165	2	area
166	2	aim
167	2	affect

7

The Corresponding Curve

Rank	Freq	
1	37	system
2	32	knowledg
3	24	base
4	20	problem
5	18	abstract
6	15	model
7	15	languag
8	15	implem
9	13	reason
10	13	inform
11	11	expert
12	11	analysi
13	10	rule
14	10	program
15	10	oper
16	10	evalu
17	10	comput
18	10	case
19	9	gener
20	9	form



8

Zipf's Law

- Distribution of word frequencies is very *skewed*
  - A few words occur very often, many words hardly ever occur
  - e.g., two most common words ("the", "of") make up about 10% of all word occurrences in English documents

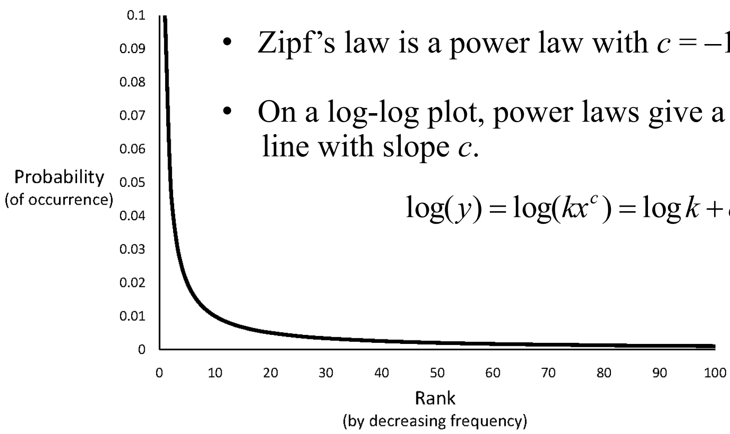
- Zipf's "law":
  - Observation that rank (r) of a word times its frequency (f) is approximately a constant (k)
    - Assuming words are ranked in order of decreasing frequency
  - i.e., $r.f \approx$ k or $r.P_r \approx A$, where P_r is probability of word occurrence and $A \approx 0.1$ for English

9

Zipf's Law

- A law of the form $y = kx^c$ is called a **power law**.
- Zipf's law is a power law with $c = -1$
- On a log-log plot, power laws give a straight line with slope c.

$$\log(y) = \log(kx^c) = \log k + c\log(x)$$



10

Zipf's Distribution

- The important points:
  - A few elements occur *very frequently*
  - A medium number of elements have medium frequency
  - Many elements occur *very infrequently*

11

News Collection (AP89) Statistics

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Word	*Freq.*	*r*	*Pr(%)*	*r.Pr*
assistant	5,095	1,021	.013	0.13
sewers	100	17,110	2.56×10^{-4}	0.04
toothbrush	10	51,555	2.56×10^{-5}	0.01
hazmat	1	166,945	2.56×10^{-6}	0.04

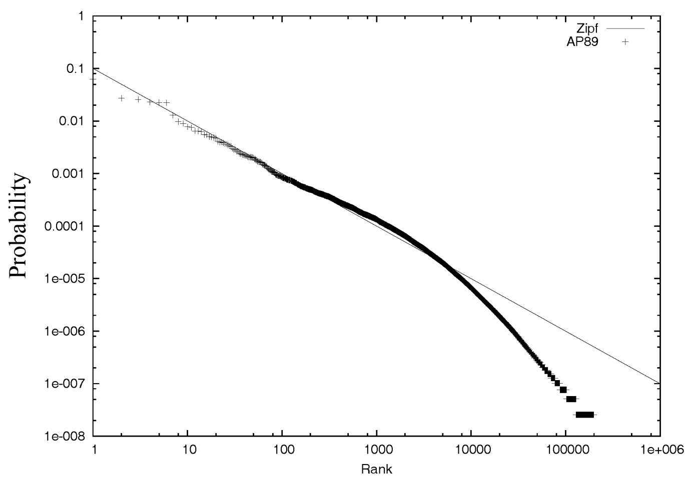
12

Top 50 Words from AP89

Word	*Freq.*	*r*	P_r(%)	$r.P_r$	*Word*	*Freq*	*r*	P_r(%)	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

13

Zipf's Law for AP89



Zipf is quite accurate except for very high and low rank.

- Note problems at high and low frequencies
  - See the Mandelbrot (1954) Correction:

$$(r + \beta)^{\alpha} \cdot P_r = \gamma$$ where β, α, and γ are parameters

14

Zipf's Law

- What is the proportion of words with a given frequency?
  - Word that occurs n times has rank $r_n = k/n$
  - Number of words with frequency n is
    - $r_n - r_{n+1} = k/n - k/(n+1) = k/n(n+1)$
  - This proportion can be found by dividing by the total number of words = highest rank $= k/1 = k$
  - So, proportion with frequency n is $1/n(n+1)$

Rank	*Word*	*Frequency*
1000	concern	5,100
1001	spoke	5,100
1002	summit	5,100
1003	bring	5,099
1004	star	5,099
1005	immediate	5,099
1006	chemical	5,099
1007	african	5,098

15

Zipf's Law

- Example word frequency ranking

Rank	*Word*	*Frequency*
1000	concern	5,100
1001	spoke	5,100
1002	summit	5,100
1003	bring	5,099
1004	star	5,099
1005	immediate	5,099
1006	chemical	5,099
1007	african	5,098

- To compute number of words with frequency 5,099
  - rank of "chemical" minus the rank of "summit"
  - 1006 − 1002 = 4

16

Example

Number of Occurrences (n)	Predicted Proportion ($1/n(n+1)$)	Actual Proportion	Actual Number of Words
1	.500	.402	204,357
2	.167	.132	67,082
3	.083	.069	35,083
4	.050	.046	23,271
5	.033	.032	16,332
6	.024	.024	12,421
7	.018	.019	9,766
8	.014	.016	8,200
9	.011	.014	6,907
10	.009	.012	5,893

- Proportions of words occurring n times in 336,310 TREC documents
- Vocabulary size is 508,209

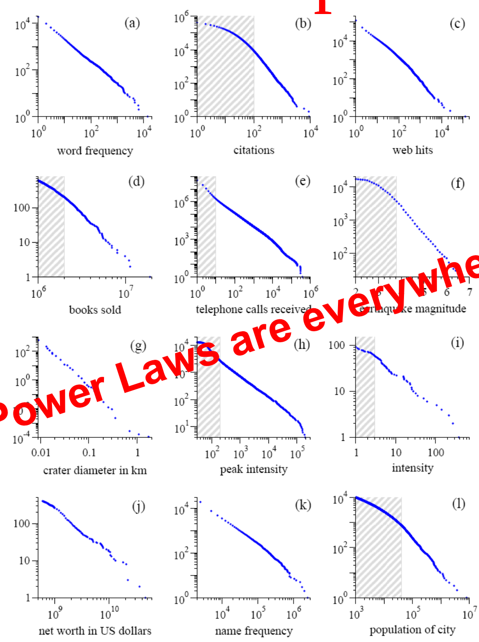


Figure from: Newman, M. E. J. (2005) "Power laws, Pareto distributions and Zipf's law." Contemporary Physics 46:323–351.

Consequences of Zipf Law

- There are always a few very frequent tokens that are not good discriminators. Referred to as "stopwords" in IR
  - Correspond to linguistic notion of "closed-class" words
    - English examples: *to, from, on, and, the, ...*
    - Grammatical classes that don't take on new members.
  - Eliminating them greatly reduces inverted-index storage costs
  - Postings list for most remaining words in the inverted index will be short since they are rarer, making retrieval fast
- There are always a large number of tokens that occur once
  - These words do not describe the content of documents
- Medium frequency words are the most descriptive

19

Vocabulary Growth

- How big is the term vocabulary?
  - How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?
- This determines how the size of the lexicon within the inverted index will scale with the size of the corpus.
- Vocabulary not really upper-bounded due to proper names, typos, invented words (e.g. product, company names), email addresses, etc.

20

Heaps' Law

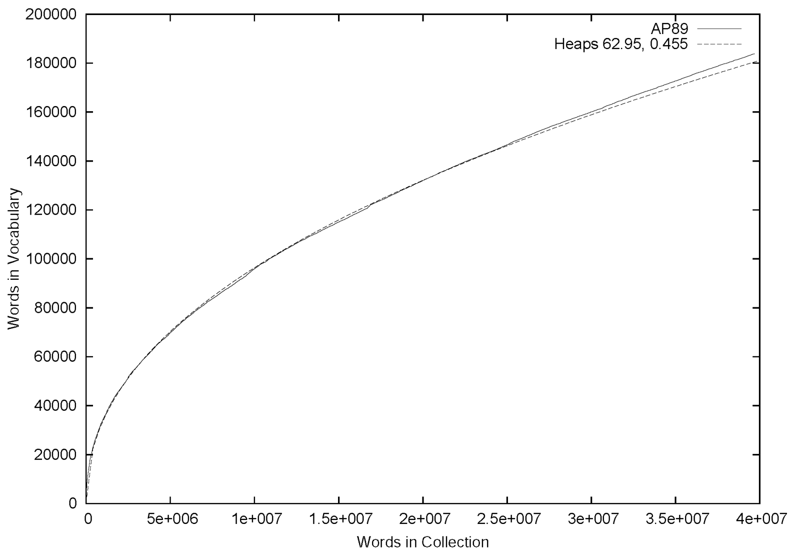
- If V is the size of the vocabulary (number of unique words) and n is the number of words in corpus:

$$V = Kn^{\beta} \quad \text{with constants } K,\ 0 < \beta < 1$$

- Typical constants:
  - $K \approx 10\text{–}100$
  - $\beta \approx 0.5$

21

AP89 Example



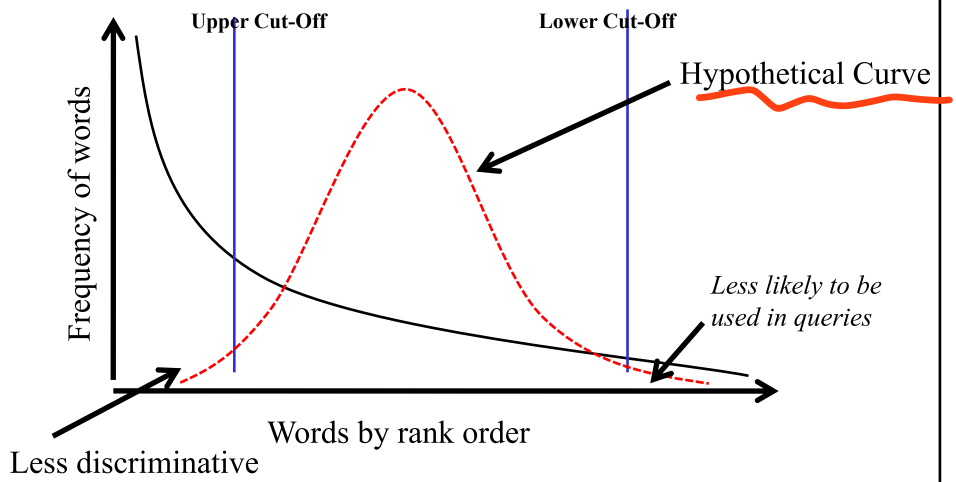
22

Automatic Document Representation

- Content Analysis: transforming raw text into more computationally useful forms
  - The objective is to use a set of terms to describe the document
- So far we looked into properties of word occurrences
  - Word frequencies follow a Zipf distribution
    - Stopwords vs Content words
  - Word co-occurrences exhibit dependencies
    - E.g. Microsoft -> Windows; Microsoft -> Office
- Let's revisit what sorts of words are useful for representing document aboutness

23

Word Frequency vs. Resolving Power



The frequency of a word occurrence in an article furnishes a useful measurement of word significance [Luhn 1957]

24

Resolving Power

- Why some words occur more frequently and how such statistics can be exploited when **automatically measuring aboutness?**
  - …. the frequency of a word occurrence in an article furnishes a useful measurement of word significance [Luhn 1957]

- **Two critical factors**
  - Word frequency within a document
  - Collection frequency

25

Term Weighting

- More effective approaches score documents based on:
  - How many query terms they contain
  - How discriminative each of those terms are

- Not all terms are equally useful, so we weight them
  - Weight describes/quantifies the relationship between a keyword and a document.
  - Generality
    - Use terms with significant weights
    - Binary is a special case
    - A retrieval method can exploit these weights.

26

Notations

OUTPUT

- w_{kd} = weight of k^{th} word in document d

INPUTS

- f_{kd} = number of occurrences of k^{th} word in document d (term frequency)
- N = Number of documents in the collection
- D_k = number of documents containing k^{th} word

27

How to Weight Terms

- Two demands on the weight
  - The degree to which a particular document is about a topic (or a particular keyword)
    - **Repetition** is an indication of emphasis
  - Degree to which a keyword **discriminates** documents in the collection
    - Let us call it ***discrim***$_k$

$$w_{kd} \propto f_{kd} \times discrim_k$$

28

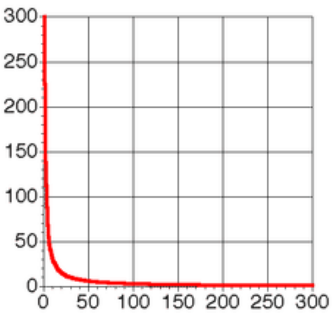
Inverse Document Frequency
(IDF)

- From a discriminating point of view:
  - Queries use rather broadly defined, frequently occurring terms
  - It is the more specific terms that are particularly important in identifying relevant material

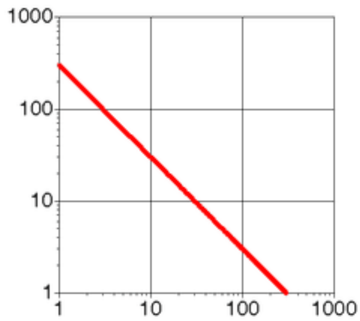
$$idf(t_k) = \log \frac{N}{D_k}$$

K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval" Journal of Documentation, 1972

Linear and log scale



$$\frac{N}{D_k}$$



$$\log \frac{N}{D_k}$$

TF – IDF Weighting Schemes

$$w_{kd} = f_{kd}\left(\log \frac{N}{D_k}\right)$$

Problem if D_k is zero

$$w_{kd} = f_{kd}\left(\log \frac{N+1}{D_k+1}\right) \; **$$

$$W_{kd} = \underbrace{(1 + \log(f_{kd}))}_{\text{TF}} \underbrace{\left(\log \frac{(N - D_k) + 0.5}{D_k + 0.5}\right)}_{\text{IDF}}$$

31

Best Match Retrieval Algorithm (with weights)

- Given what we know now about term weighting, we can update the Best-Match retrieval algorithm from before:

- Best-Match

  for each document I, Score(I) =0; I =1 to N
    for each query term t_k
      Search the vocabulary list
      Pull out the postings list
      for each document J in the list,
        Score(J) = Score(J) + w_{kd}

文档d被命中概率与Wkd成正相关

32