Text Statistics

Frequencies of Words

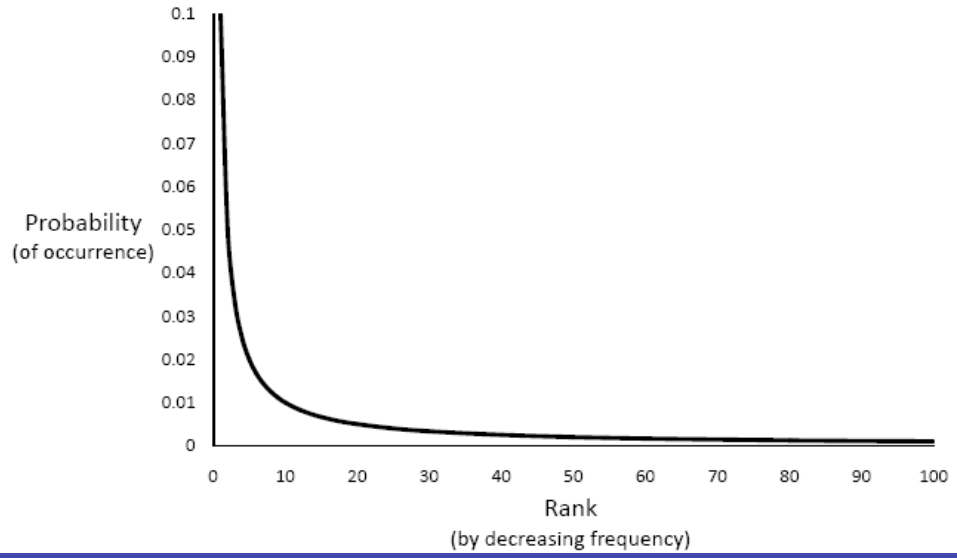
- Some words occur much more frequently than others in documents describing one topic
  - An essential observation by Luhn in 1958
  - The importance of a word in a document depends on its frequency in the document
- The distribution of word frequencies is very skewed
  - The two most frequent words in English (“the”, “of”) account for about 10% of all word occurrences
  - The most frequent 6 words account for 20% of occurrences
  - The most frequent 50 words are about 40% of all text
  - About one-half of all words only occur once

Zipf's Law/Distribution

- The frequency of the r-th most common word is inversely proportional to r
  - The rank of a word times its probability of occurrence is approximately a constant

$$r \times P(r) = c$$

  - For English, $c \approx 0.1$



Example

- AP89: the Associated Press collection of news stories from 1989

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Word occurring > 1,000 times	4,169
Words occurring once	70,064

50 Most Frequent Words in 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

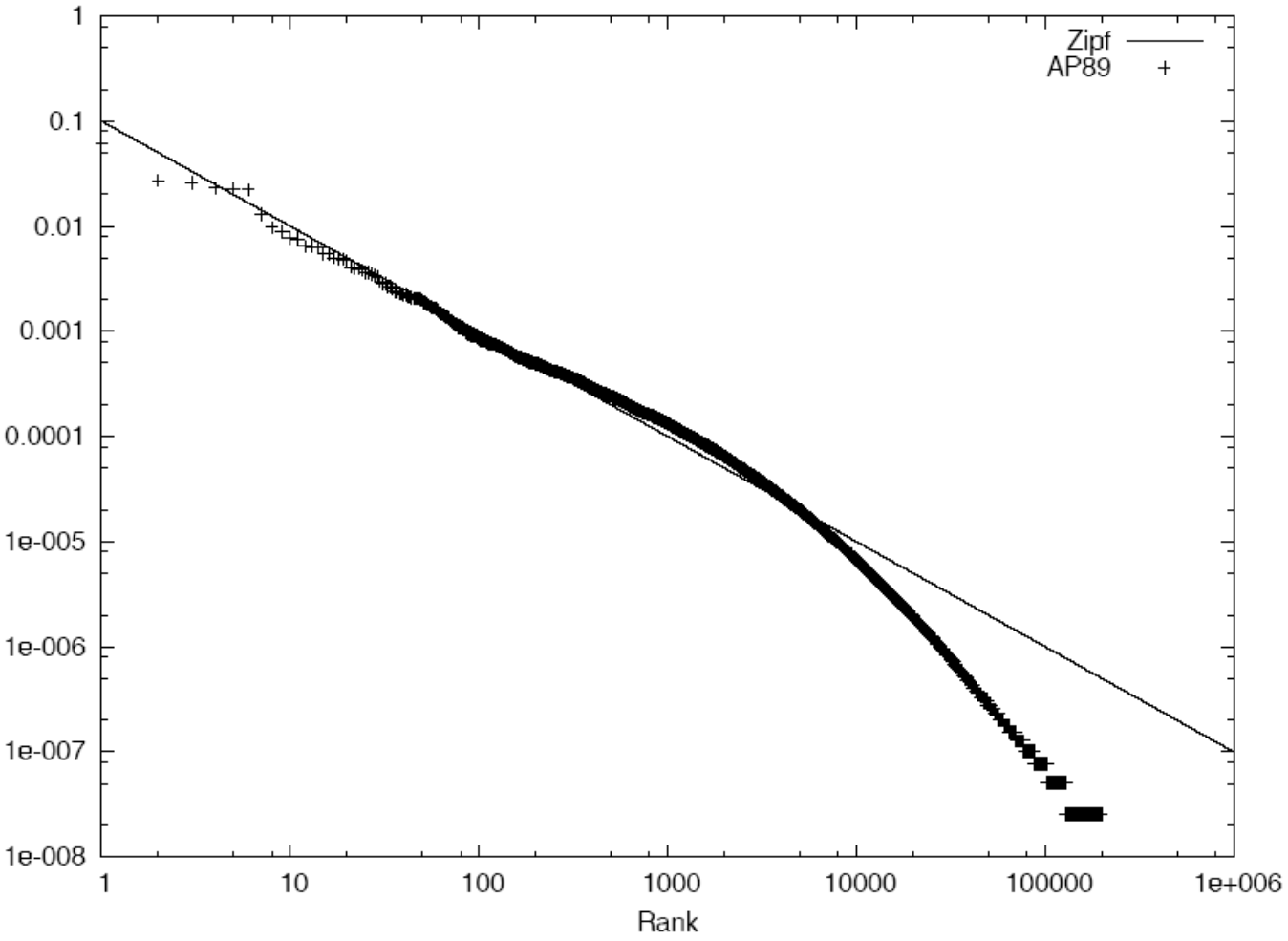
Some low frequency words

Word	*Freq.*	r	$P_r(\%)$	$r.P_r$
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

A Log-Log Plot

Constant

$\log Pr = \log (c \times r^{-1}) = \log c - \log r$



Words with the Same Frequency

- A word that occurs n times has the rank r_n = k / n, where k = c x m, and m is the total number of word occurrences
- But, more than one word may have the same frequency – we assume that the rank r_n is associated with the **last** of the group of words with the same frequency
- The number of words with the same frequency n is $r_n - r_{n+1}$

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

Using Zipf's Law in Prediction

- The proportion of words with frequency n is given by 1/n(n +1)
  - $r_n - r_{n+1} = k/n - k/(n+1) = k/n(n+1)$
  - The rank of the last word with frequency 1 is k
- ½ of the words in the vocabulary will occur only once

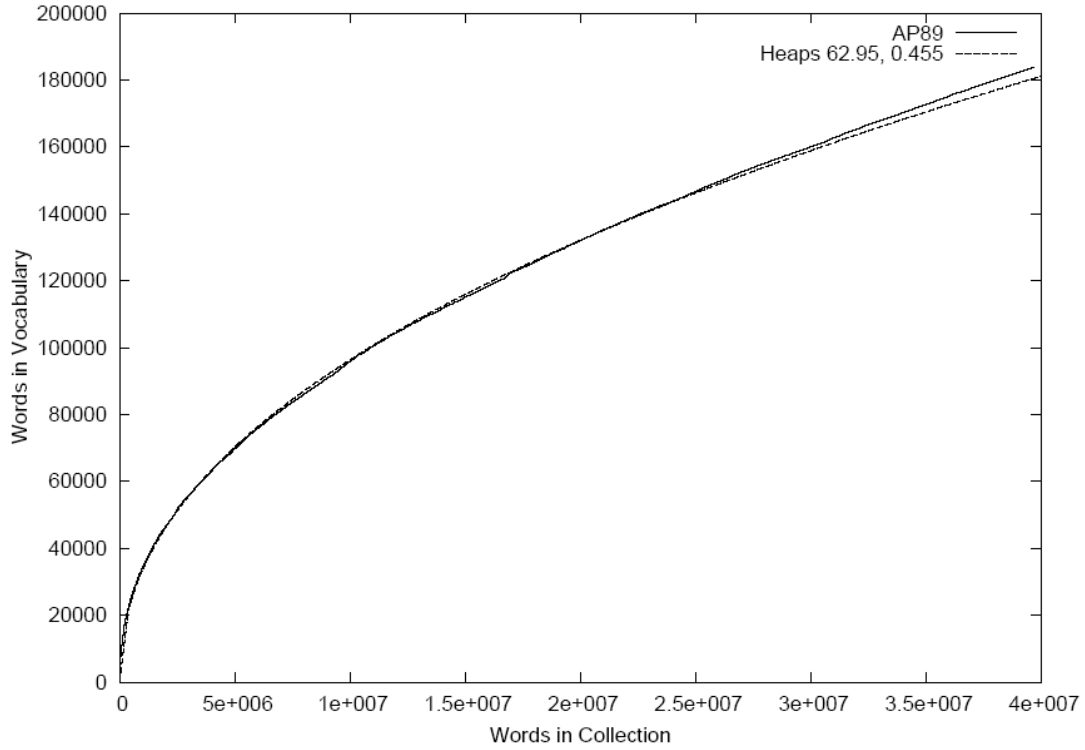
Number of Occurrences (n)	*Predicted Proportion (1/n(n+1))*	*Actual Proportion*	*Actual Number of Words*
1	.500	.402	204,357
2	.167	.132	67,082
3	.083	.069	35,083
4	.050	.046	23,271
5	.033	.032	16,332
6	.024	.024	12,421
7	.018	.019	9,766
8	.014	.016	8,200
9	.011	.014	6,907
10	.009	.012	5,893

Vocabulary Growth

- The size of corpus grows, new words occur
- The number of new words occurring in a given amount of new text decreases as the size of the corpus increases
  - Sources of new words: invented words (e.g., drug names, start-up company names), spelling errors, product numbers, people's names, email addresses, …
- Heaps' Law: empirically, the size of the corpus and the size of the vocabulary is $v = k \times n^{\beta}$
  - The vocabulary size v
  - Corpus size n
  - Values of k and β vary from each collection
  - Often, $10 \leq k \leq 100$, $\beta \approx 0.5$

Prediction Using Heaps' Law

- The number of new words will increase very rapidly when the corpus is small and would continue to increase indefinitely, but at a slower rate for larger corpus



Estimating Result Size

- How many documents in the corpus are there containing all the words in a query?
  - An estimate appears in most of the search engines
- Independency assumption: if keywords are independent, then $P(a \cap b \cap c) = P(a) \times P(b) \times P(c)$
  - A search engine always have access to the number of documents a word occurs in
  - $P(a) = f_a/N$, $P(b) = f_b/N$, $P(c) = f_c/N$
  - $f_{abc} = N \times f_a/N \times f_b/N \times f_c/N = (f_a \times f_b \times f_c) / N^2$

Independent Keyword?

Word(s)	*Document Frequency*	*Estimated Frequency*
tropical	120,990	
fish	1,131,855	
aquarium	26,480	
breeding	81,885	
tropical fish	18,472	5,433
tropical aquarium	1,921	127
tropical breeding	5,510	393
fish aquarium	9,722	1,189
fish breeding	36,427	3,677
aquarium breeding	1,848	86
tropical fish aquarium	1,529	6
tropical fish breeding	3,629	18

Using Word Co-occurrences

- A search engine may collect the frequencies of word co-occurrences
- $P(a \cap b \cap c) = P(a \cap b) \times P(c \mid (a \cap b))$
  - $P(c \mid (a \cap b))$ can be approximated by $\max\{P(c \mid a), P(c \mid b)\}$
- Example: $f_{tropical \cap fish \cap aquarium} = f_{tropical \cap aquarium} \times f_{fish \cap aquarium} / f_{aquarium} = 1921 \times 9722 / 26480 = 705$
  - Closer to the real frequency 1529
  - Still low

Estimation Using the Current Results

- Search engines do not rank all documents that contain the query words
  - They rank a much smaller subset of documents that are likely to be the most relevant
- Estimate the result size as C/s
  - C: the number of documents found that contain all query words in the documents ranked
  - s: the proportion of the total documents that have been ranked
- Example: after processing 3,000 out of the 26,480 documents that contain “aquarium”, if the number of documents containing all three keywords is 258, we can estimate the result size as 258 / 3000 x 26480 = 2277

Summary

- Zipf's Law and application in prediction
- Heaps' Law and application in prediction
- Estimation of the result size

To-Do List

- Sections 4.1-4.2