Eigenschaften von Texten

Statistische Eigenschaften von Text

- Wie ist die Häufigkeit verschiedener Wörter verteilt?
- Wie schnell wächst die Größe des Vokabulars mit der Größe eines Korpus?
- Solche Faktoren beeinflussen die Performanz des Information Retrieval und können verwendet werden, um angemessene Termgewichte und andere Aspekte eines IR-Systems auszuwählen.

Worthäufigkeiten

- Einige Wörter sind sehr gebräuchlich.
  - 2 der häufigsten Wörter (z.B. “the”, “of”) können ca. 10 % der Wortvorkommen ausmachen.
- Die meisten Wörter sind sehr selten.
  - Die Hälfte der Wörter in einem Korpus erscheint nur einmal, dies nennt man *hapax legomena* (Griechisch für “nur einmal lesen”)
- Dies ergibt eine schiefe Verteilung.

Beispiel: Worthäufigkeiten

(von B. Croft, UMass)

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Zipfs Gesetz

- Rank (r): Die numerische Position eines Wortes in einer nach abnehmender Häufigkeit (f) sortierten Liste
- Zipf (1949) "entdeckte" dass:

$$f \propto \frac{1}{r} \qquad f \cdot r = k \ \text{(für Konstante } k\text{)}$$

- Wenn die Wahrscheinlichkeit der Rangstelle r p_r ist und N die Gesamtzahl der Wortvorkommen:

$$p_r = \frac{f}{N} = \frac{A}{r} \quad \text{für Korpus - unabhängige Konst. } A \approx 0.1$$

Zipf und Termgewichtung

- Luhn (1958) wies darauf hin, dass sowohl sehr häufige als auch sehr seltene Wörter nicht sehr nützlich für das Indexieren sind.

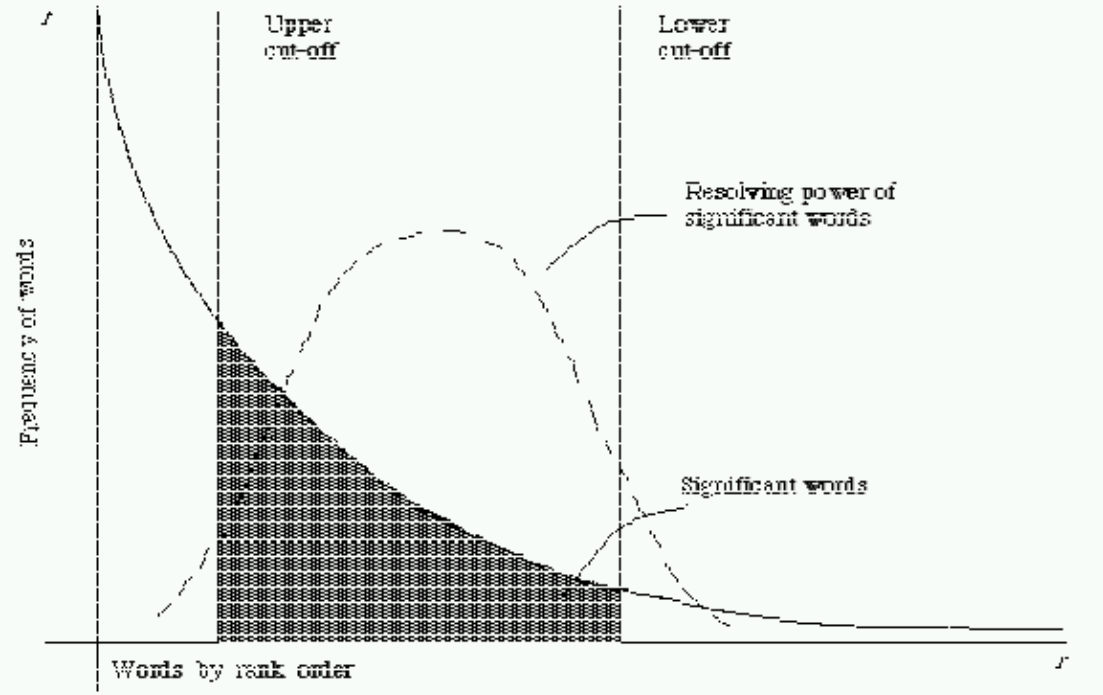


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 121)

Vorhersagen zu Wort-Häufigkeiten

- Gemäß Zipf, hat ein Wort, das *n* mal erscheint, die Rangordnung $r_n = AN/n$
- Mehrere Wörter können *n*-mal auftreten. Wir nehmen an, dass die Rangordnung r_n für das letzte davon gilt.
- Also kommen r_n Wörter *n*-fach oder öfter vor und r_{n+1} Wörter $n+1$-fach oder öfter.
- Dann beträgt die Anzahl der Wörter, die **genau** *n* Mal vorkommen:

$$I_n = r_n - r_{n+1} = \frac{AN}{n} - \frac{AN}{n+1} = \frac{AN}{n(n+1)}$$

Forts. Vorhersagen von Worthäufigkeiten

- Nehmen wir an, dass der am höchsten gerankte Term einmal vorkommt und daher die Rangordnung $D = AN/1$ hat.
- Der Anteil der Wörter mit Häufigkeit n ist dann:

$$\frac{I_n}{D} = \frac{1}{n(n+1)}$$

- Der Anteil der Wörter, die nur einmal vorkommen, ist demzufolge ½.

Daten zu Vorkommens-Häufigkeiten

(von B. Croft, UMass)

Number of Occurrences (n)	Predicted Proportion of Occurrences 1/n(n+1)	Actual Proportion occurring n times I_n/D	Actual Number of Words occurring n times
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

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

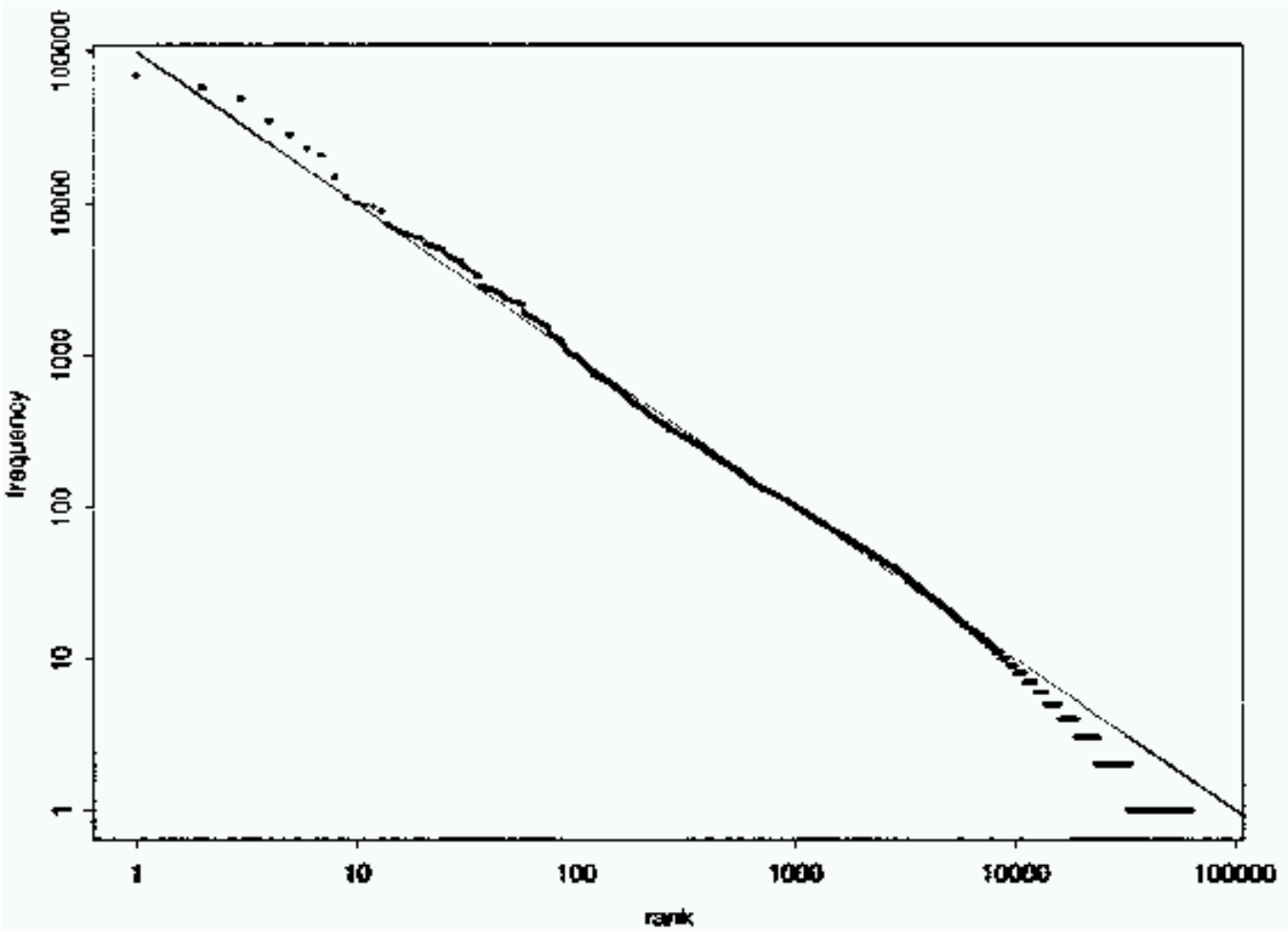
Passen reale Daten zu Zipfs Gesetz?

- Ein Verhältnis der Form $y = kx^c$ wird als Potenzgesetz bezeichnet.
- Zipf's Gesetz ist ein Potenzgesetz mit $c = -1$
- Bei einem log-log Plot ergeben Potenzgesetze eine gerade Linie mit Neigung c.

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

- Zipf ist ziemlich genau, außer bei sehr hoher und sehr niedriger Rangordnung.

Vergleich von Zipf mit Brown-Korpus



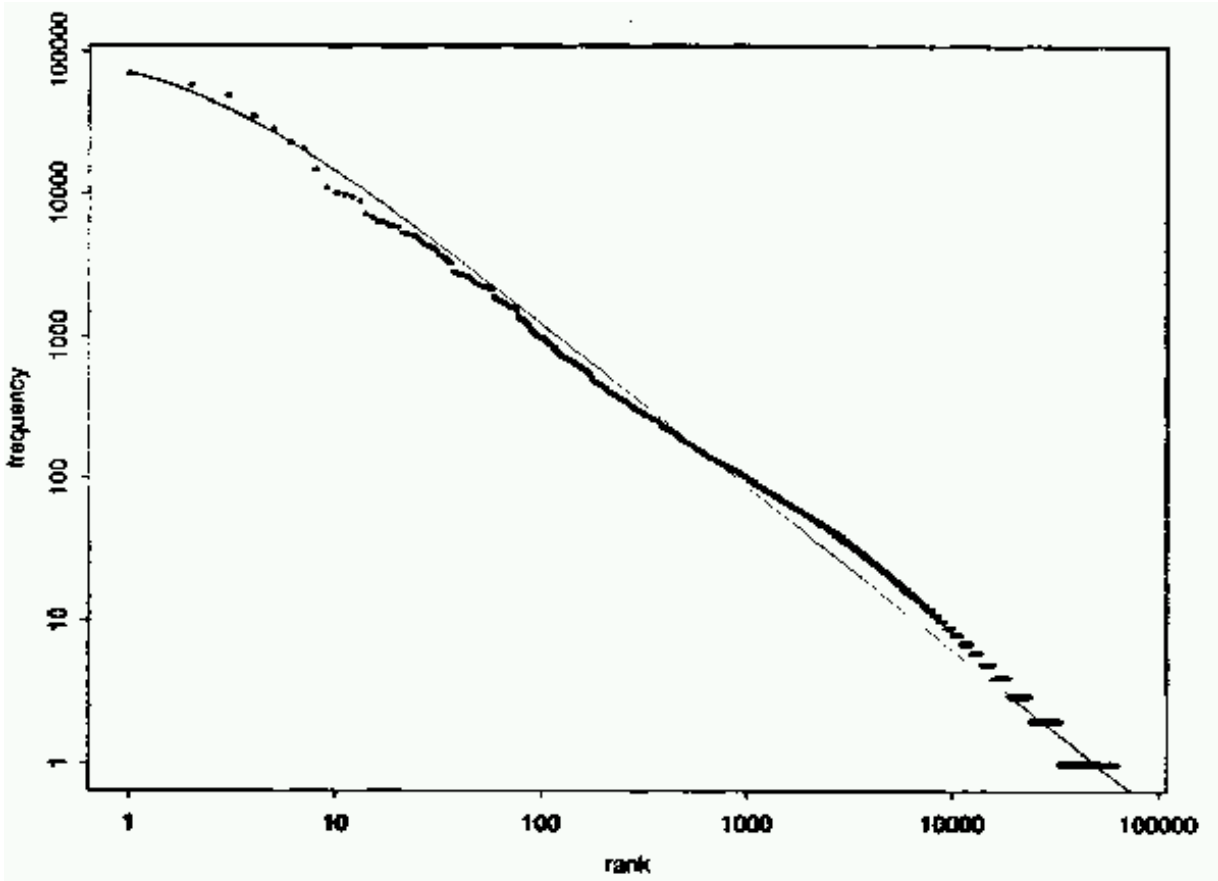
$k = 100{,}000$

Mandelbrot-Korrektur (1954)

- Die folgende – allgemeinere – Form gibt einen etwas besseren Angleich:

$$f = P(r+\rho)^{-B} \qquad \text{für Konstanten } P, B, \rho$$

Mandelbrot-Fit



Mandelbrot's function on Brown corpus

$P = 10^{5.4}$, $B = 1.15$, $\rho = 100$

Erläuterungen zum Zipf-Gesetz

- Zipfs Erläuterung war sein “Prinzip des geringsten Aufwands”: der Ausgleich zwischen dem Wunsch des Sprechers nach wenig und dem des Hörers nach viel Vokabular.
- Debatte (1955-61) über die Erläuterung zwischen Mandelbrot und H. Simon.
- Li (1992) zeigt, dass nur willkürliches Tippen von Buchstaben einschließlich Leerraum “Wörter” mit einer Zipfschen Verteilung erzeugt.
  - http://linkage.rockefeller.edu/wli/zipf/

Die Auswirkung von Zipfs Gesetz auf IR

- Gute Nachricht: Stopwörter machen einen großen Teil des Textes aus, so dass deren Beseitigung die Speicherkosten des invertieren Indexes in großem Umfang reduziert.
- Schlechte Nachricht: Für die meisten Wörter ist es schwierig, ausreichend Daten für eine aussage-fähige statistische Analyse zu sammeln (z.B. für Korelationsanalysen für Anfrageerweiterungen), da sie ziemlich selten sind.

Wachstum des Vokabulars

- Wie wächst die Größe des gesamten Vokabulars (Anzahl eindeutiger Wörter) mit der Größe des Korpus?
- Dies bestimmt, wie die Größe des invertierten Indexes mit der Größe des Korpus wächst.
- Das Vokabular ist aufgrund von Eigennamen, Tippfehlern, etc. nicht wirklich nach oben bgrenzt.

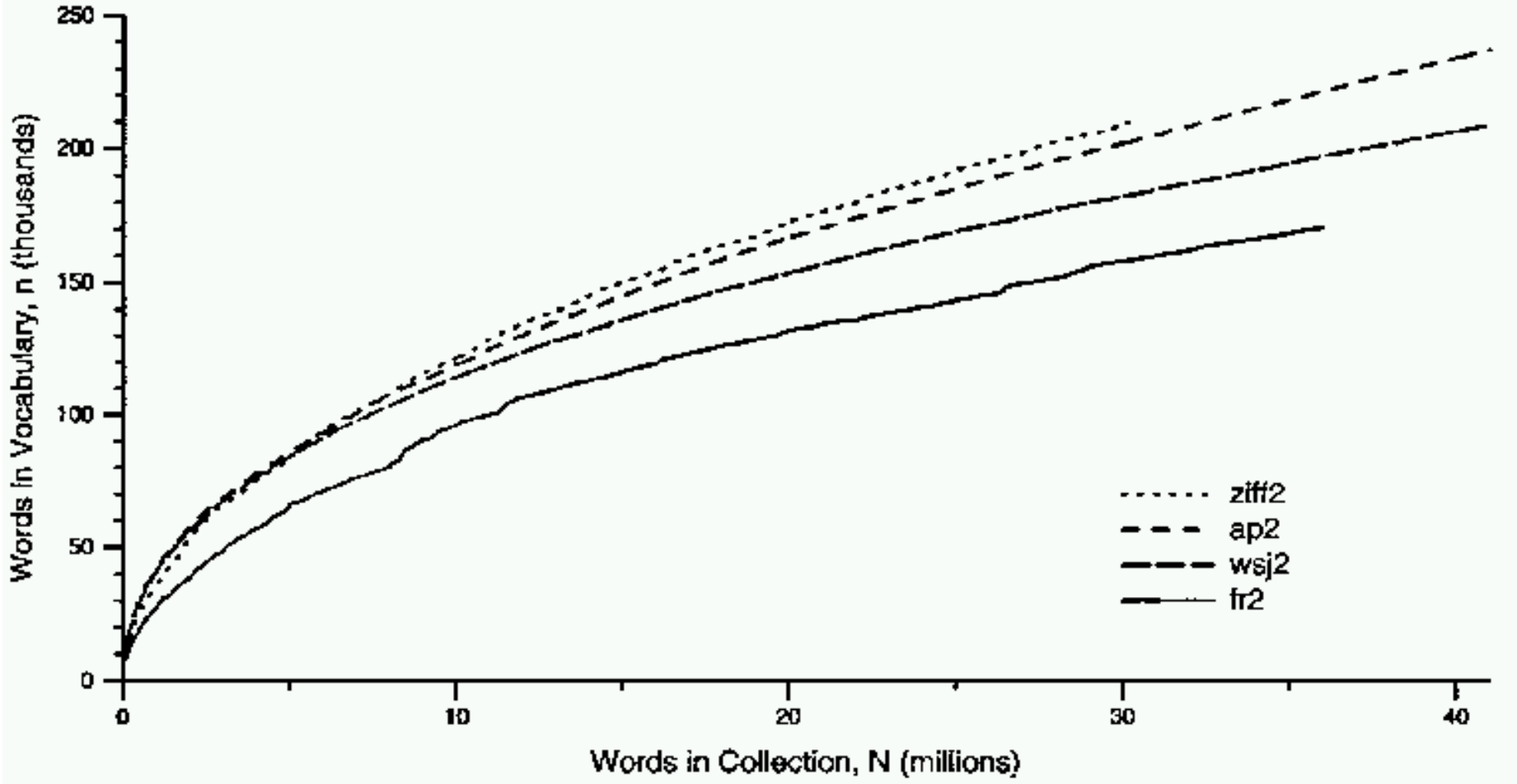
Heaps Gesetz

- Falls *V* die Größe des Vokabulars und *n* die Länge des Korpus in Wörtern ist:

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

- Typische Konstanten:
  - $K \approx 10–100$
  - $\beta \approx 0.4–0.6$ (d.h. ungefähr Quadratwurzel)

Daten, die Heaps Gesetz folgen



Erläuterung zu Heaps Gesetz

- Kann von Zipfs Gesetz abgeleitet werden, wobei angenommen wird, dass Dokumente in einer willkürlichen Wortauswahl aus einer Zipfschen Verteilung erzeugt werden.