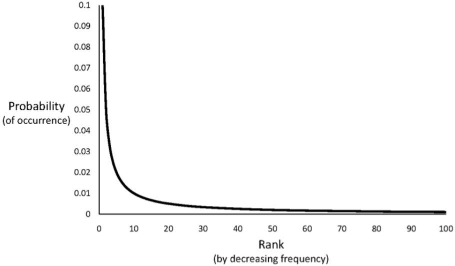
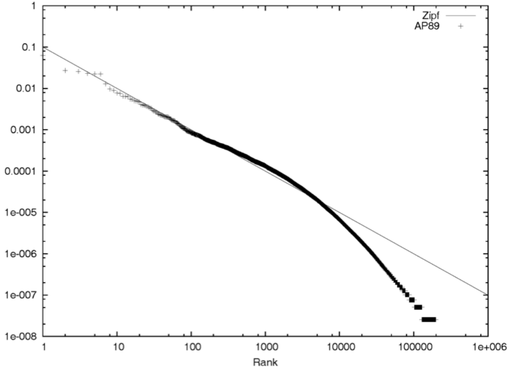
Basic Tenets of Zipf's Law

- Words are not even distributed
  - 10% of words in text documents are made of "the" and "of"



- Zipf's Law describes the relationship between ranking (importance) and frequency
  - The rank (r) of a word times its frequency (f) is approximately a constant(k) $\longrightarrow k \approx rf$
  - The probability of a word's occurrence times its rank is approximately a constant (c) $\longrightarrow c \approx rP$ [for English $c \approx 0.1$]



On a log-log plot of actual data agains the theoretical Zipf's Law plot, we observe that data falls apart for lower frequency (higher rank) words. Since $P = c/r$ and by taking the log of both sides we now have: $\log(P) = \log(c) - \log(r)$. If c is a constant then the slope of the line is $-\log(r)$.

Proportion of Words

What is the proportion of words with a given frequency? If you have the words frequency (usually derive from a small sample), can you calculate the ratio of the word to the collection?

- Words occurring n times has rank $r = k/n$
- Number of words with frequency n is

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

- Proportion (ratio) found by dividing by the total number of words
- Proportion with frequency n is $1/(n(n+1))$

Page rank **without** a random jump factor λ

If v is a webpage in the set B_u where

- v contains at least one link to page u
- L_v are the number of links from page v

Then formula for calculating page rank for u is:

$$Pr(u) = \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

In other words, the rank of page u is the sum of the rank each page v divided by the number of its outgoing links. L_v are, of course, free of duplicates.

Let's assume that our document collection is as followed:

- Document A has links to documents B and C
- Document B has a link to document C but not document A and therefore does not contribute to A's ranking
- Document C has a link to document A but not to document B and therefore does not contribute B's ranking

Let's calculate the ranking in k iterations. For the first iteration, we assume all pages have equal rank ≈ 0.33. Then using results from this step to calculate the next iteration. Repeat until convergence.

First iteration

Calculate page rank for A. Because B does not link to A so we do not consider it:

$$PR(A) = \frac{PR(C)}{1} = \frac{0.33}{1} = 0.33$$

Calculate page rank for B. Because C does not link to B so we do not consider it:

$$PR(B) = \frac{PR(A)}{2} = \frac{0.33}{2} \approx 0.17$$

Calculate page rank for C:

$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1} = \frac{0.33}{2} + \frac{0.33}{1} \approx 0.50$$

Second iteration

Substitute the above results for $PR(A)$, $PR(B)$, and $PR(C)$:

$$PR(A) = \frac{PR(C)}{1} = \frac{0.50}{1} = 0.50$$

$$PR(B) = \frac{PR(A)}{2} = \frac{0.33}{2} \approx 0.17$$

$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1} = \frac{0.33}{2} + \frac{0.17}{1} \approx 0.33$$

Third iteration

Substitute the above results for $PR(A)$, $PR(B)$, and $PR(C)$:

$$PR(A) = \frac{PR(C)}{1} = \frac{0.33}{1} = 0.33$$

$$PR(B) = \frac{PR(A)}{2} = \frac{0.50}{2} = 0.25$$

$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1} = \frac{0.50}{2} + \frac{0.17}{1} \approx 0.42$$

Continue these steps until the numbers converge. In other words, the next iteration no longer updates the PR values or the updated values are less than some small value ϵ

Page rank **with** a random jump factor λ

If we take random page jump into account and the chance of going to any page when $r < \lambda$, then formula for calculating page rank for u is now:

$$Pr(u) = \frac{\lambda}{N} + (1 - \lambda) \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

Note N is the total number of pages in the system, λ is typically a constant that is statistically calculated to be 0.15, and r is the probability of visiting a page.

A similar iterative approach can be taken to calculate that the page ranking with random jump factor.