Power Law in Natural Languages and Random Text

E6083: lecture 6
Prof. Predrag R. Jelenković

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
predrag@ee.columbia.edu

February 21, 2007

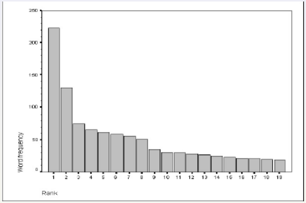
Outline

1. Power Law in Natural Languages
   - Entropy Optimization Formulation
2. Power Law in Random Text
   - Equal Probability Case
   - Unequal Probability Case

Power Law in Language

Discovery of Power Law in Language

Zipf found Power Law by analyzing the distribution of words in English



Explanation:

- Least Effort: a universal property of mind, the principle of least effort to balance between uniformity and diversity
- Least Cost: entropy-optimization formulation

Outline

Entropy Optimization Model (Mandelbrot 1953)

- W words, cost of transmitting *jth* most frequent word is C_j, cost of space is 0
- Average information per word is the entropy

$$H = -\sum_{j=1}^{W} f_j \log_2 f_j,$$

and average cost per word is

$$C = \sum_{j=1}^{W} f_j C_j.$$

Entropy-optimization Formulation

Objective: optimize the average amount of information per unit transmitting cost

$$A = \frac{C}{H}.$$

Take a derivative

$$\frac{\partial A}{\partial f_j} = \frac{C_j H + C \log_2(e f_j)}{H^2}$$

Natural cost

Number of letters plus the additional space for a space. Hence,

$$\log_N j \leq C_j \leq \log_N j + 1$$

because the word with k letters have frequency ranks from $1 + (N^k - 1)/(N - 1)$ to $(N^{k+1} - 1)/(N - 1)$.

Entropy-optimization Formulation

Solution

$f_j = e^{-1}2^{-HC_j/C}$, which implies, for cost $\log_N j \le C_j \le \log_N j + 1$

$$(2^{-H/C}e^{-1})j^{-H(\log_N 2)/C} \le f_j \le e^{-1}j^{-H(\log_N 2)/C}$$

i.e., a power law.

Question

Can Zipf's Law still hold without an intentionally least effort principle? Let's do an experiment!

A Fascinating Problem on Monkey Typing

Basic setting

- N letters and a space
- Hit space with probability p
- Hit other letters with probability $p_i, 1 \leq i \leq N$

The Result of Random Typing

Q: What is the rank-frequency distribution of words?

A: power law!

Definition

We call frequency f_j follows a power law in j if $c_1 j^{-\alpha} \leq f_j \leq c_2 j^{-\alpha}$ for large j

Outline

Equal Letter Probability Case (Miller 1957)

$p_i = \frac{1-p}{N}$, then

- N^k possible words of length k
- The words with k letters have frequency ranks from $1 + (N^k - 1)/(N-1)$ to $(N^{k+1} - 1)/(N-1)$
- Each word with k letters occurs with probability

$$p_k = \left(\frac{1-p}{N}\right)^k p$$

- The word with rank-frequency j occurs with probability f_j

$$\left(\frac{1-p}{N}\right)^{\log_N j+1} p \leq f_j \leq \left(\frac{1-p}{N}\right)^{\log_N j} p$$

Outline

Unequal Letter Probability Case

A Simple Example with Only Two Letters

- Letter "a" appears with probability q, "b" with q^2, space $1-q-q^2$
- Every word with pseudorank k occurs with probability $q^k(1-q-q^2)$

- The number of words with pseudorank k is the $(k+1)$th Fibonacci number $F_{k+1} = \Phi^{k+1}/\sqrt{5} + o(1)$, $\Phi = (1+\sqrt{5})/2$
- When $F_{k+1} - 1 < j \leq F_{k+3} - 1$, the *jth* most frequent word has psudorank k,

$$k + 2 \leq \log_\Phi(\sqrt{5}(j+1)) < k + 3.$$

Therefore f_j satisfies

$$q^{\log_\Phi(\sqrt{5}(j+1))-2}(1-q-q^2) < f_j \leq q^{\log_\Phi(\sqrt{5}(j+2))-3}(1-q-q^2)$$

General Case

- $0 < p_1 < p_2 < \cdots, p_i = p_1^{a_i}, a_i \in \mathcal{R}$, w_i letters are struck with p_i
- prob of space is $1 - \sum w_i p_i$
- Q: How many words occurs with probability greater or equal than $p_1^{\nu}(1 - \sum w_i p_i), \nu \in \mathcal{R}$?
- A: the quantity c_v: how many ways v can be expressed as a sum of elements a_i. Algebraically,

$$\frac{1}{1 - \sum_{i=1}^{n} w_i x^{a_i}} = \sum c_v x^v.$$

Main Theorem (Condrad & Mitzenmacher '04)

It is proved using complex analysis that, for x_0 that satisfies $\sum_{i=1}^{n} w_i x_0^{a_i} = 1$,

$$\liminf_{t\to\infty} \frac{\sum_{v<t} c_v}{(1/x_0)^t} = \liminf_{t\to\infty} \frac{\sum_{v\le t} c_v}{(1/x_0)^t} = A$$

$$\limsup_{t\to\infty} \frac{\sum_{v<t} c_v}{(1/x_0)^t} = \limsup_{t\to\infty} \frac{\sum_{v\le t} c_v}{(1/x_0)^t} = A'$$

- if all $a_i/a_{i'} \in Q$ (rational-ratio case), $A = x_0^{1/r} A'$ with $r = D/a_1$, D is the least common multiple of the denominators of the ratios a_i/a_1
- if some $a_i/a_{i'}$ is irrational, $A = A'$

Derivation of Power Law Using the Main Theorem

Using the asymptotic result to estimate the rank frequencies $f_j = p_1^{t(j)} \left(1 - \sum_{i=1}^n w_i p_i\right)$.

1. Pick $0 < L \leq L'$ such that $L(1/x_0)^t \leq \sum_{v<t} c_v \leq \sum_{v \leq t} c_v \leq L'(1/x_0)^t$
2. $\sum_{v<t} c_v < j \leq \sum_{v \leq t} c_v \Longrightarrow \frac{\log j - \log L'}{\log(1/x_0)} \leq t(j) < \frac{\log j - \log L}{\log(1/x_0)} \Longrightarrow$

$$p_1^{(\log j - \log L)/\log(1/x_0)} \left(1 - \sum w_i p_i\right) \leq f_j \leq p_1^{(\log j - \log L')/\log(1/x_0)} \left(1 - \sum w_i p_i\right)$$

Probabilistic Arguments for Monkeys Typing Randomly

Keyboard has N Letters with hitting probabilities $p_1 \geq p_2 \geq \cdots \geq p_N$ and a space with hitting probability p. Define the set $W_k = \{\text{all words of length } k\}$.

If $p_1 = p_2 = \cdots = p_N$, then, the words of longer length are less likely and hence occur lower in the rank order of word frequency. Thus, $W_1 \prec W_2 \prec \cdots \prec W_\infty$ where $a \prec b$ means a has a lower rank than b. The words with k letters have frequency ranks from $1 + \frac{N^k - 1}{N-1}$ to $\frac{N^{k+1}-1}{N-1}$.

Now, if $p_1, p_2, \cdots, p_N$ are not equal, then, the rank of the set W_k will stretch. In other words, some words of shorter length will have a higher rank than some words of longer length.

Estimate the spread of elements in W_x to the ones in W_y

- We want to study under what conditions $W_x \prec W_y$ for $x < y$. If $p_n^x > p_1^y \Leftrightarrow y > \frac{\log p_n}{\log p_1} x$, then, $\Rightarrow W_x \prec W_y$.
- Therefore, $W_{\frac{\log p_1}{\log p_n} x} \prec W_x \prec W_{\frac{\log p_n}{\log p_1} x}$.

Suppose that the word with rank j belongs to $W_{C(j)}$, and we obtain

$$\sum_{i=1}^{\frac{\log p_1}{\log p_N} C(j)} N^i \leq j \leq \sum_{i=1}^{\frac{\log p_n}{\log p_1} C(j)} N^i.$$

Thus

$$\frac{\log p_1}{\log p_N} \log_N j \leq C(j) \leq \frac{\log p_N}{\log p_1} \log_N j + c.$$

Rough bounds of the frequency show power laws

The word with rank-frequency j satisfies

$$p_N^{\frac{\log p_1}{\log p_N} \log_N j} p \geq f_j \geq p_1^{\frac{\log p_N}{\log p_1} \log_N j + c} p,$$

which implies

$$\frac{\log p_1}{\log N} \geq \frac{\log f_j}{\log j} \geq \frac{\log p_N}{\log N}.$$

More refined sample path analysis may lead to the exact power law exponent.

The Mathematics of Monkeys and Shakespeare

Infinite monkey experiments

A monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type or create a particular chosen text, such as the complete works of William Shakespeare.

However, the probability is too small! It is quite clearly impossible for even a trivial fragment of Shakespeare's work to have arisen by chance.

Monkeys Produce Hamlet: Feasibility Study

Keys	Chances (one in...)
1	32
2	$32 * 32 = 1024$
3	$32 * 32 * 32 = 32768$
4	$32 * 32 * 32 * 32 = 1048576$
5	$32^5 = 33554432$
6	$32^6 = 1073741824$
7	$32^7 = 34359738368$
8	$32^8 = 1099511627776$
9	$32^9 = 3.518437208883e + 013$
10	$32^{10} = 1.125899906843e + 015$
...	
20	$32^{20} = 1.267650600228e + 030$
...	
30	$32^{30} = 1.427247692706e + 045$
...	
41	$32^{41} = 5.142201741629e + 061$