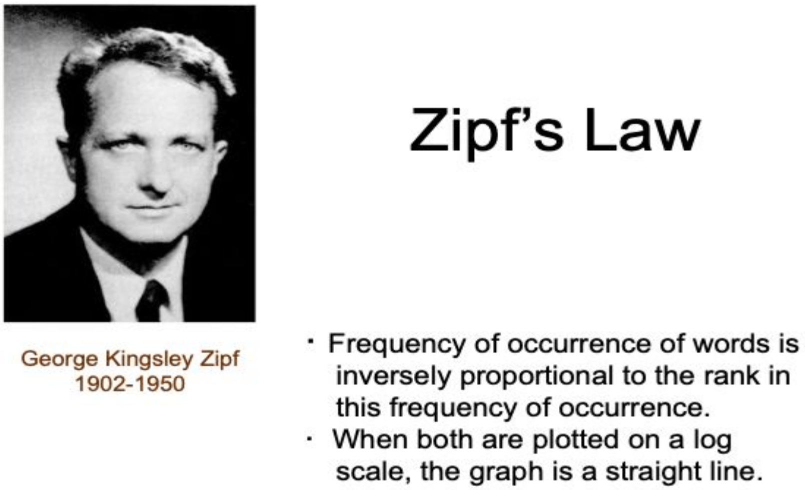
ICA 1 Zipf’s Law

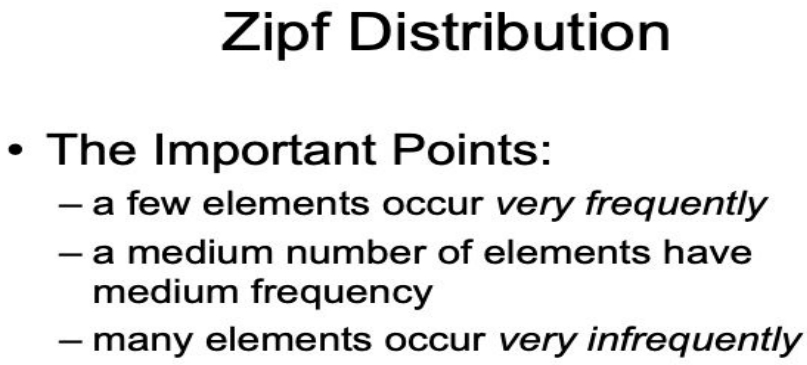
(September 9th)



Zipf’s Law

George Kingsley Zipf
1902-1950

- Frequency of occurrence of words is inversely proportional to the rank in this frequency of occurrence.
- When both are plotted on a log scale, the graph is a straight line.



Zipf Distribution

- The Important Points:
  - a few elements occur *very frequently*
  - a medium number of elements have medium frequency
  - many elements occur *very infrequently*

In this activity, you will explore Zipf’s Law.

From NLTK import gutenberg corpus and open up the book *Alice in Wonderland* by Lewis Carol.

Step 0: Preprocessing - strip the text of punctuation, and lower case the words - obtaining the original text as a list of words.

Step 1: Now write a function that counts words, unique words, and frequency of occurrence.

Step 2: For the most frequent 25 words and for the most frequent 25 *additional* words that start with the letter *c* (a total of 50 words), print the word, the number of times it occurs, its rank in the overall list of words, the probability of occurrence, and the product of the rank and the probability.

Step 3: Also indicate the total number of words and the total number of unique words that you found.

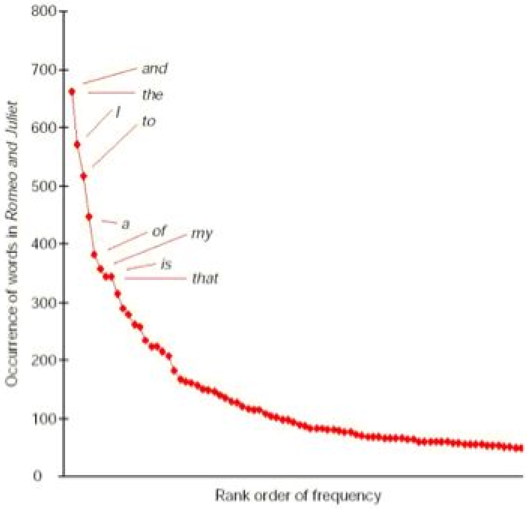
Step 4: Discuss how this text satisfies Zipf's Law (or why it does not).

Zipf Distribution

The product of the frequency of words (f) and their rank (r) is approximately constant

Rank = order of words' frequency of occurrence

$$f = C * 1/r$$
$$C \cong N/10$$



Zipf Distribution (Same curve on linear and log scale)

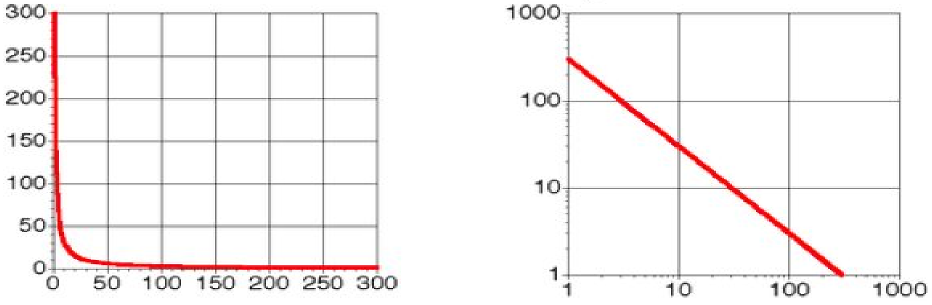


Illustration by Jacob Nielsen