N-Grams and Smoothing

Manish Shrivastava

Basic Idea:

- Examine short sequences of words
- How likely is each sequence?
- “Markov Assumption” – word is affected only by its “prior local context” (last few words)

Example

- The boy ate a chocolate
- The girl bought a chocolate
- The girl then ate a chocolate
- The boy bought a horse

- Can we figure out how likely is the following sentence
  - The boy bought a chocolate

“Shannon Game”

- **Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.**
- **Predict the next word, given (*n-1*) previous words**
- **Determine probability of different sequences by examining training corpus**

Forming Equivalence Classes (Bins)

- "*n-gram*" = sequence of n words
  - bigram
  - trigram
  - four-gram or quadrigram
- Probabilities of n-grams
  - Unigram $$p(w) = \frac{c(w)}{N}$$
  - Bigram $$P(w_i \mid w_{i-1}) = \frac{c(w_i, w_{i-1})}{c(w_{i-1})}$$
  - Trigram $$P(w_i \mid w_{i-1}, w_{i-2}) = \frac{c(w_i, w_{i-1}, w_{i-2})}{c(w_{i-1}, w_{i-2})}$$

Maximum Likelihood Estimation (MLE)

- The boy bought a chocolate
  - Unigram Probabilities
    - (4/16)*(2/16)*(2/16)*(4/16)*(3/16)
    - (4*2*2*4*3)/21^5 = 0.000047
  - Bi-gram Probabilities
    - <The boy> <boy bought> <bought a> <a chocolate>
    - (2/4)*(2/4)*(2/2)*(3/4) = 0.1875
- Data
  - The boy ate a chocolate
  - The girl bought a chocolate
  - The girl then ate a chocolate
  - The boy bought a horse

Reliability vs. Discrimination

“large green ___________”

tree? mountain? frog? car?

“swallowed the large green ________”

pill? candy?

Reliability vs. Discrimination

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)

Selecting an *n*

Vocabulary (V) = 20,000 words

n	Number of bins
2 (bigrams)	400,000,000
3 (trigrams)	8,000,000,000,000
4 (4-grams)	1.6×10^{17}

Statistical Estimators

- Given the observed training data …
  - How do you develop a model (probability distribution) to predict future events?
  - Language Modeling
    - Predict Likelihood of sequences

Maximum Likelihood Estimation

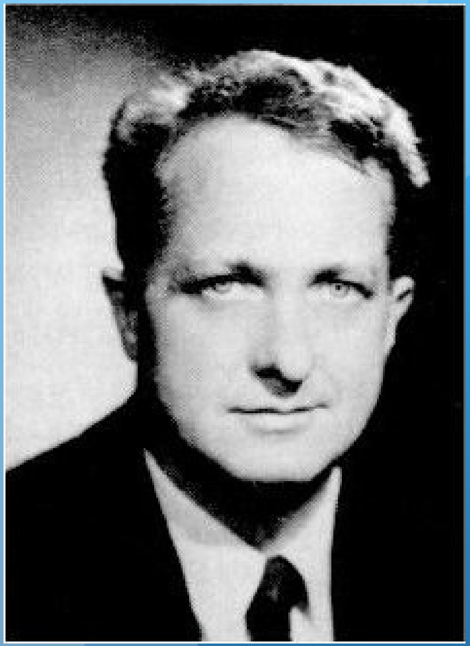
- $P_{MLE}(w_n|w_1 \dots w_{n-1}) = \frac{C(w_1 \dots w_n)}{C(w_1 \dots w_{n-1})}$

- Estimate sequence probabilities using "counts" or frequencies of sequences
- Problems
  - Sparseness
  - What do you do when unknown words are seen??

Example

- Data
  - The boy ate a chocolate
  - The girl bought a chocolate
  - The girl then ate a chocolate
  - The horse bought a boy

- The boy bought a chocolate
  - Unigram Probabilities
    - (4/16)*(2/16)*(2/16)*(4/16)*(3/16)
    - (4*2*2*4*3)/21^5 = 0.000047
  - Bi-gram Probabilities
    - <The boy> <boy bought> <bought a> <a chocolate>
    - (2/4)*(0/4)*(2/2)*(3/4) = **0**



George Kingsley Zipf
1902-1950

Zipf’s Law

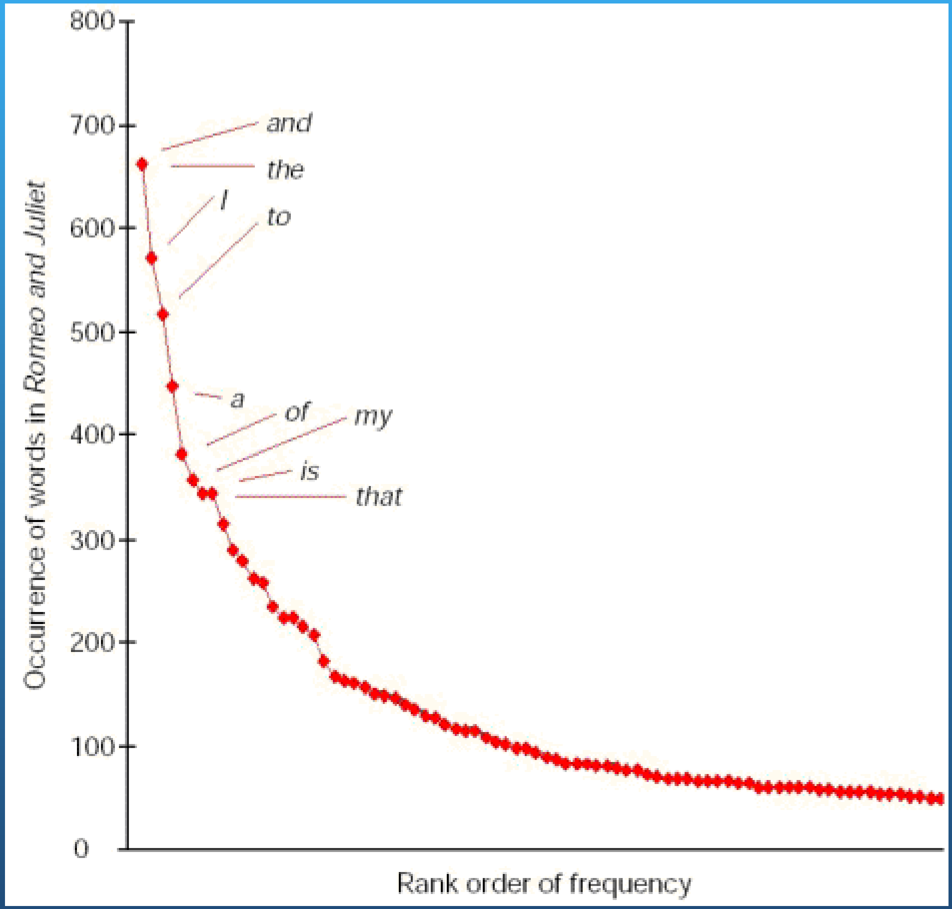
- Frequency of occurrence of words is inversely proportional to the rank in this frequency of occurrence.
- When both are plotted on a log scale, the graph is a straight line.

Zipf Distribution

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

Zipf Distribution

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

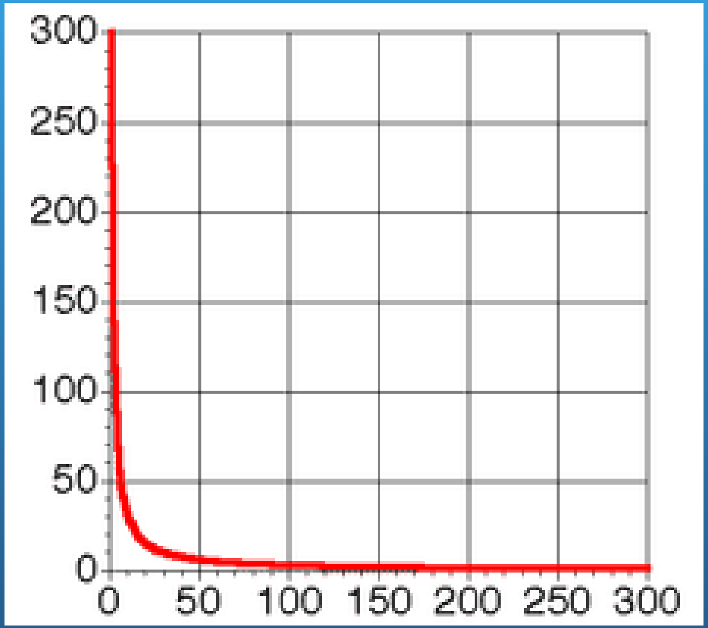


Rank = order of words' frequency of occurrence

$$f = C * 1/r$$

$$C \cong N/10$$

Zipf Distribution
(Same curve on linear and log scale)



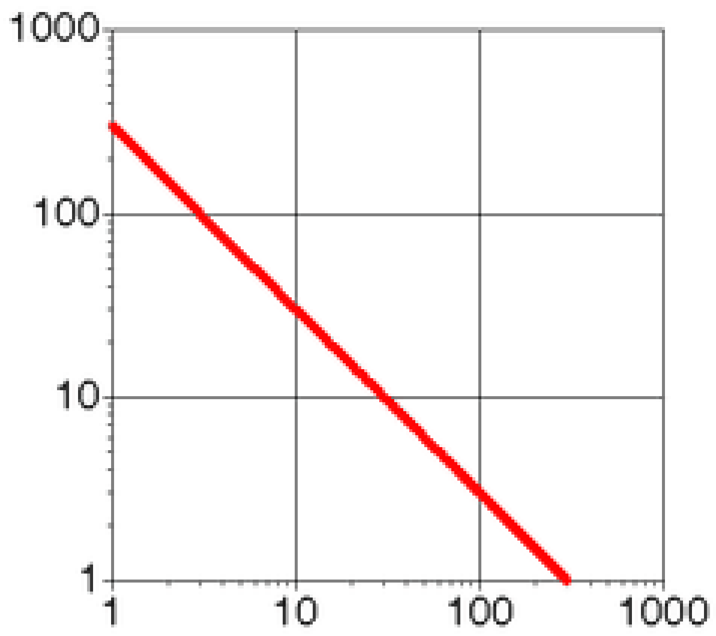
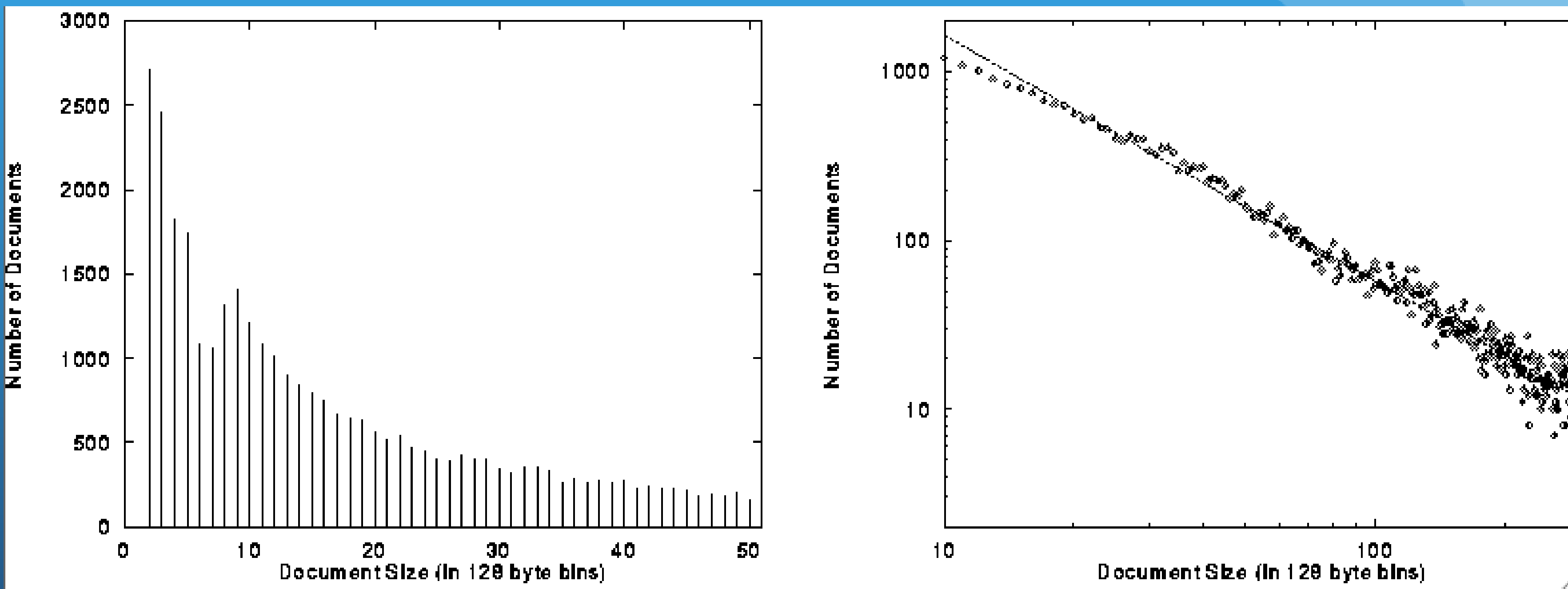


Illustration by Jacob Nielsen

What Kinds of Data Exhibit a Zipf Distribution?

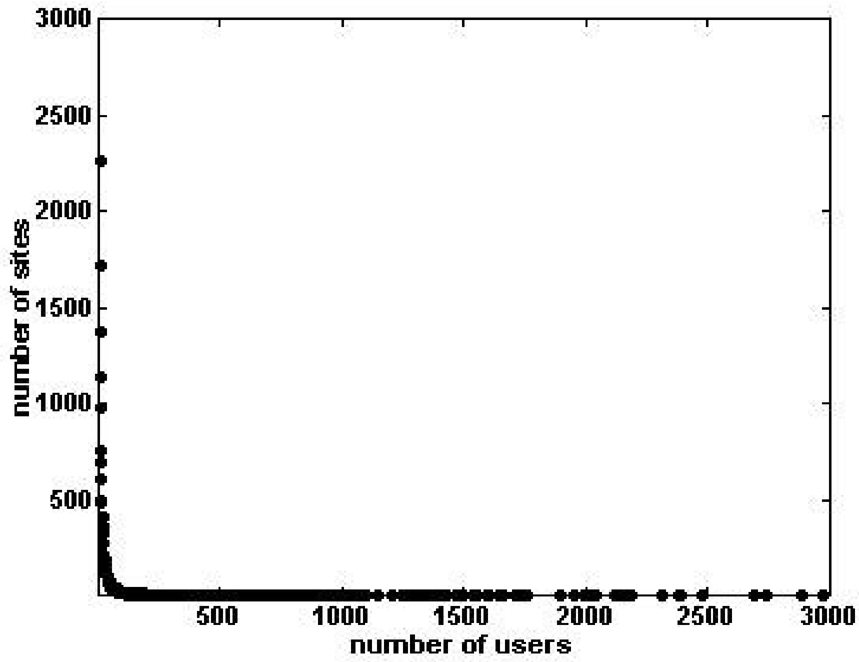
- Words in a text collection
  - Virtually any language usage
- Library book checkout patterns
- Incoming Web Page Requests (Nielsen)
- Outgoing Web Page Requests (Cunha & Crovella)
- Document Size on Web (Cunha & Crovella)

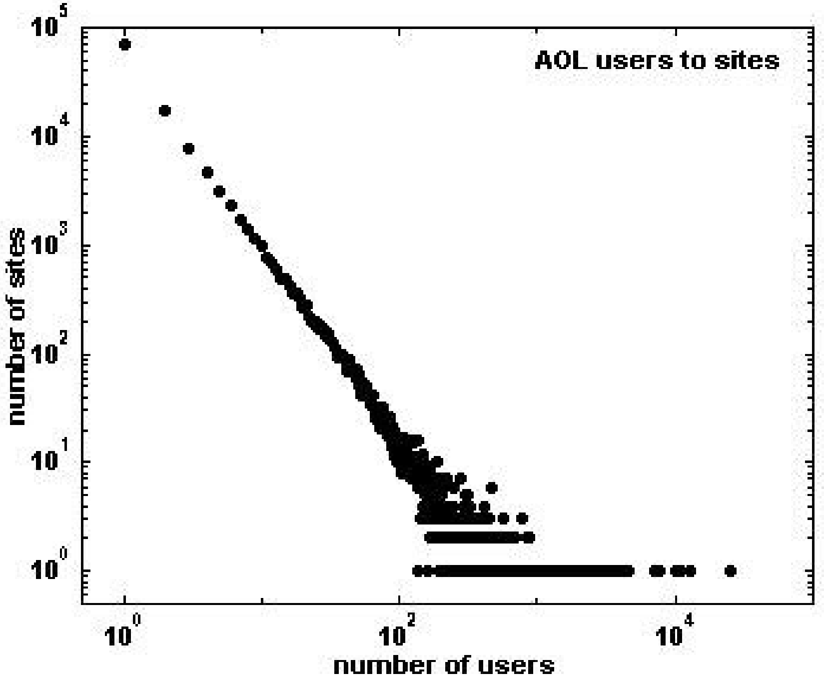
Characteristics of WWW Client-based Traces



Zipf's Law Applied To WWW Documents

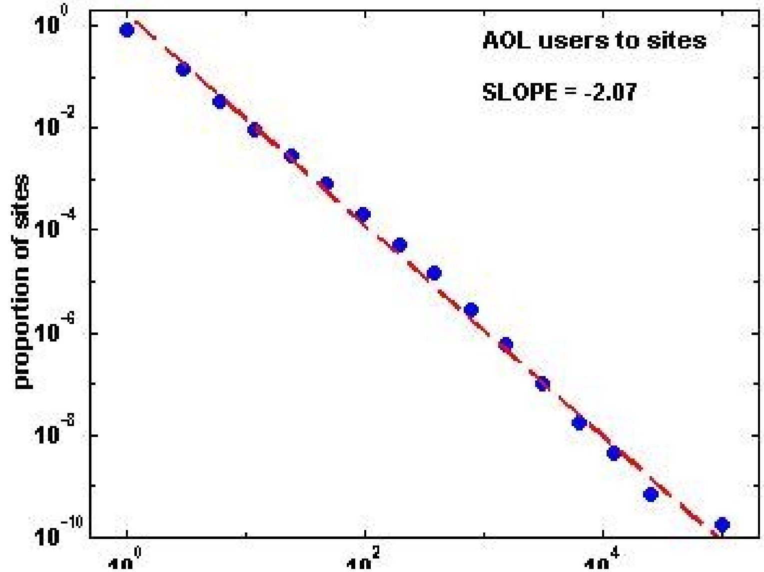
Distribution of users among web sites





Binned distribution of users to sites

Exponentially increasing bins



Cumulative distribution of users to sites