Statistical Properties of Text

• **Zipf’s Law** models the distribution of terms in a corpus:
  – How many times does the $k^{\text{th}}$ most frequent word appears in a corpus of size $N$ words?
  – Important for determining index terms and properties of compression algorithms.

• **Heap’s Law** models the number of words in the vocabulary as a function of the corpus size:
  – What is the number of unique words appearing in a corpus of size $N$ words?
  – This determines how the size of the inverted index will scale with the size of the corpus.
Word Distribution

• **A few words are very common:**
  – The 2 most frequent words (e.g. “the”, “of”) can account for about 10% of word occurrences.

• **Most words are very rare:**
  – Half the words in a corpus appear only once, called *hapax legomena* (Greek for “read only once”)

• **A “heavy tailed” or “long tailed” distribution:**
  – Since most of the probability mass is in the “tail” compared to an exponential distribution.
Word Distribution

Frequency vs. rank for all words in Moby Dick

Lecture 01
Moby Dick:
- 44% hapax legomena
- 17% dis legomena

“Honorificabilitudinitatibus”:
- Shakespeare’s hapax legomenon
- longest word with alternating vowels and consonants
Zipf’s Law

• Rank all the words in the vocabulary by their frequency, in decreasing order.
  – Let $r(w)$ be the rank of word $w$.
  – Let $f(w)$ be the frequency of word $w$.

• Zipf (1949) postulated that frequency and rank are related by a power law:

$$ f(w) = \frac{c}{r(w)} $$

  – $c$ is a normalization constant that depends on the corpus.
Zipf’s Law

• If the most frequent term (the) occurs $f_1$ times:
  – Then the second most frequent term (of) occurs $f_1 / 2$ times.
  – The third most frequent term (and) occurs $f_1 / 3$ times, …

• **Power Laws**: $y = cx^k$
  – Zipf’s Law is a power law with $k = -1$.
  – Linear relationship between log($y$) and log($x$):
    • $\log(y) = \log c + k \log(x)$
    • on a log scale, power laws give a straight line with slope $k$.

• Zipf is quite accurate, except for very high and low rank.
Zipf’s Law Fit to Brown Corpus

\[ f(w) = \frac{100000}{r(w)} \]
Mandelbrot’s Distribution

- The following more general form gives a bit better fit:
  \[ f = c / (r + \rho)^K \]

- When fit to Brown corpus:
  - \( c = 105.4 \)
  - \( K = -1.15 \)
  - \( \rho = 100 \)
Mandelbrot’s Law Fit to Brown Corpus

Mandelbrot’s function on Brown corpus
Zipf’s Law Impact on IR

- **Good News:**
  - Stopwords will account for a large fraction of text, so eliminating them greatly reduces inverted-index storage costs.
  - Postings list for most remaining words in the inverted index will be short since they are rare, making retrieval fast.

- **Bad News:**
  - For most words, gathering sufficient data for meaningful statistical analysis is difficult since they are extremely rare.
    - for correlation analysis for query expansion.
    - for ML estimation in language modeling.
Vocabulary vs. Collection Size

• How big is the term vocabulary?
  – That is, how many distinct words are there?

• Can we assume an upper bound?
  – Not really upper-bounded due to proper names, typos, etc.

• In practice, the vocabulary will keep growing with the collection size.
Heap’s Law

• **Given:**
  - \( M \) is the size of the vocabulary.
  - \( T \) is the number of tokens in the collection.

• **Then:**
  - \( M = kT^b \)
  - \( k, b \) depend on the collection type:
    - typical values: \( 30 \leq k \leq 100 \) and \( b \approx 0.5 \) (square root).
    - in a log-log plot of \( M \) vs. \( T \), Heaps’ law predicts a line with slope of about \( \frac{1}{2} \).
Heap’s Law Fit to Reuters RCV1

- For RCV1, the dashed line \( \log_{10} M = 0.49 \log_{10} T + 1.64 \) is the best least squares fit.

- Thus, \( M = 10^{1.64} T^{0.49} \) so \( k = 10^{1.64} \approx 44 \) and \( b = 0.49 \).

- For first 1,000,020 tokens:
  - Law predicts 38,323 terms;
  - Actually, 38,365 terms.
  \( \Rightarrow \) Good empirical fit for RCV1!
Explanations

• **Zipf’s Law:**
  – Zipf’s explanation was his “principle of least effort”:
    • Balance between speaker’s desire for a small vocabulary and hearer’s desire for a large one.
  – Herbert Simon’s explanation is “rich get richer.”
  – Li (1992) shows that just random typing of letters including a space will generate “words” with a Zipfian distribution.

• **Heaps’ Law:**
  – Can be derived from Zipf’s law by assuming documents are generated by randomly sampling words from a Zipfian distribution.