Information Retrieval and Web Search

Text properties

Instructor: Rada Mihalcea

(Note: some of the slides in this set have been adapted from a course taught by Prof. James Allan at U.Massachusetts, Amherst)
Statistical Properties of Text

• **Zipf’s Law** models the distribution of terms in a corpus:
  - How is the frequency of different words distributed?
  - How many times does the $k^{th}$ 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:
  - How fast does vocabulary size grow with the size of a corpus?
  - 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.
  - 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”)

- Called a “heavy tailed” distribution, since most of the probability mass is in the “tail”
### Sample Word Frequency Data

<table>
<thead>
<tr>
<th>Frequent Word</th>
<th>Number Occurrences</th>
<th>Percentage of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>7,398,934</td>
<td>5.9</td>
</tr>
<tr>
<td>of</td>
<td>3,893,790</td>
<td>3.1</td>
</tr>
<tr>
<td>to</td>
<td>3,364,653</td>
<td>2.7</td>
</tr>
<tr>
<td>and</td>
<td>3,320,687</td>
<td>2.6</td>
</tr>
<tr>
<td>in</td>
<td>2,311,785</td>
<td>1.8</td>
</tr>
<tr>
<td>is</td>
<td>1,559,147</td>
<td>1.2</td>
</tr>
<tr>
<td>for</td>
<td>1,313,561</td>
<td>1.0</td>
</tr>
<tr>
<td>The</td>
<td>1,144,860</td>
<td>0.9</td>
</tr>
<tr>
<td>that</td>
<td>1,066,503</td>
<td>0.8</td>
</tr>
<tr>
<td>said</td>
<td>1,027,713</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Frequencies from 336,310 documents in TREC 3
125,720,891 total words, 508,209 unique words (B. Croft)
Zipf’s Law

- **Rank** \((r)\): The numerical position of a word in a list sorted by decreasing frequency \((f)\).

- Zipf (1949) “discovered” that:

\[
 f \cdot r = k \quad \text{(for constant } k) \]

- If probability of word of rank \(r\) is \(p_r\) and \(N\) is the total number of word occurrences:

\[
 p_r = \frac{f}{N} = \frac{A}{r} \quad \text{for corpus const. } A
\]
Predicting Occurrence Frequencies

• By Zipf, a word appearing \( f \) times has rank \( r_f = \frac{AN}{f} \).

• Several words may occur \( f \) times, assume rank \( r_f \) applies to the last of these.

• Therefore, \( r_f \) words occur \( f \) or more times and \( r_{f+1} \) words occur \( f+1 \) or more times.

• So, the number of words appearing exactly \( f \) times is:

\[
I_f = r_f - r_{f+1} = \frac{AN}{f} - \frac{AN}{f+1} = \frac{AN}{f(f+1)}
\]
Predicting Word Frequencies (cont’d)

• Assume highest ranking term occurs once and therefore has rank $D = AN/1$

• Fraction of words with frequency $f$ is:

$$\frac{I_f}{D} = \frac{1}{f(f + 1)}$$

• Fraction of words appearing only once is therefore $\frac{1}{2}$. 
Occurrence Frequency Data

<table>
<thead>
<tr>
<th>Number Occurrences</th>
<th>Predicted Proportion $1/f(f+1)$</th>
<th>Actual Proportion</th>
<th>Actual Number of Words Occuring $f$ Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.500</td>
<td>.402</td>
<td>204,357</td>
</tr>
<tr>
<td>2</td>
<td>.167</td>
<td>.132</td>
<td>67,082</td>
</tr>
<tr>
<td>3</td>
<td>.083</td>
<td>.069</td>
<td>35,083</td>
</tr>
<tr>
<td>4</td>
<td>.050</td>
<td>.046</td>
<td>23,271</td>
</tr>
<tr>
<td>5</td>
<td>.033</td>
<td>.032</td>
<td>16,332</td>
</tr>
<tr>
<td>6</td>
<td>.024</td>
<td>.024</td>
<td>12,421</td>
</tr>
<tr>
<td>7</td>
<td>.018</td>
<td>.019</td>
<td>9,766</td>
</tr>
<tr>
<td>8</td>
<td>.014</td>
<td>.016</td>
<td>8,200</td>
</tr>
<tr>
<td>9</td>
<td>.011</td>
<td>.014</td>
<td>6,907</td>
</tr>
<tr>
<td>10</td>
<td>.009</td>
<td>.012</td>
<td>5,893</td>
</tr>
</tbody>
</table>

Frequencies from 336,310 documents in TREC 3
125,720,891 total words, 508,209 unique words (B. Croft)
Does Real Data Fit Zipf's Law?

• A law of the form $y = kx^c$ is called a power law.
• Zipf's law is a power law with $c = -1$
• On a log-log plot, power laws give a straight line with slope $c$.

$$\log(y) = \log(kx^c) = \log k + c \log(x)$$

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

\[ k = 100,000 \]
Explanations for 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.
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.

• **Bad News**: For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.
Zipf’s Law on the Web

• Length of web pages has a Zipfian distribution.
• Number of hits to a web page has a Zipfian distribution.
Exercise

• Assuming Zipf’s Law with a corpus constant $A=0.1$, what is the fewest number of most common words that together account for more than 25% of word occurrences (i.e. the minimum value of $m$ such as at least 25% of word occurrences are one of the $m$ most common words)
Vocabulary Growth

• How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?
• This determines how the size of the inverted index will scale with the size of the corpus.
• Vocabulary not really upper-bounded due to proper names, typos, etc.
Heaps’ Law

- If $V$ is the size of the vocabulary and the $n$ is the length of the corpus in words:

$$V = Kn^\beta \text{ with constants } K, \ 0 < \beta < 1$$

- Typical constants:
  - $K \approx 10-100$
  - $\beta \approx 0.4-0.6$ (approx. square-root)
Heaps’ Law Data
Exercise

• We want to estimate the size of the vocabulary for a corpus of 1,000,000 words. However, we only know statistics computed on smaller corpora sizes:
  – For 100,000 words, there are 50,000 unique words
  – For 500,000 words, there are 150,000 unique words

  – Estimate the vocabulary size for the 1,000,000 words corpus
  – How about for a corpus of 1,000,000,000 words?