Text Collections

Padmini Srinivasan
School of Library and Information Science
Management Sciences
University of Iowa
Outline

• Datasets and DTDs.
• Sentences and words
• Normalization, document representation (preliminary)
• Characteristics of texts
  – Zipf’s law
  – Mandelbrot’s version
  – Heap’s Law
Sample Text Collections

TREC (http://trec.nist.gov)

TREC datasets: worf.info-science.uiowa.edu/TREC (View with netscape)

DTDs
What’s a sentence?

Tentative boundary at . ? ! (maybe ; : - )

- move boundary after following quotation marks.
- disqualify boundary if:
  - preceded by known abbreviation that is commonly followed by proper name Prof. vs.
  - preceded by known abbreviation not followed by capital etc. Jr.
- disqualify boundary with ? or ! if followed by lower case letter
- rest are sentence boundaries.

MXTERMINATOR: http://www.cis.upenn.edu/~adwait/statnlp.html
Project?: Tailoring/training MXTERMINATOR for MEDLINE.

What’s a word?

“a string of contiguous alphanumerical characters with spaces on either side, may include hyphens and apostrophes, but no other punctuation mark”. [Kucera and Frances, 1967]

Tokens and Types.

- $22.50
- :-) 
- etc. Calif.
- they’re going
- boys’ toys
• A-1-plus or A+ co-operate take-it-or-leave-it
• splitting words at the end of the line
• inconsistencies: data-base data base database
• multiple forms of the same lexeme (single meaning)
• whitespace not "term"/word boundary: New York, in spite of, work out. work the answer out.
Normalizing terms

- Stopwords
- Capitalization
- Stemming
- Semantic mappings
Document Representation

- Entries
- Weights
- Positional data
- Frequencies

Modify/Expand representations

- Thesaurus
- Definitions
- Statistically related terms
Characteristics of Texts

Zipf’s law of word occurrences in a text collection

Mandelbrot’s formula

Heap’s law of vocabulary growth
Zipf’s Law

George Kingsley Zipf (1902-1950)

A small number of words occur very often

Many words occur rarely

Principle of least effort: Difficult to make up a new word each time for the same concept

Balance between speakers/writer’s desire for a small vocabulary and hearers/readers desire for a large one.
Sort vocabulary by frequency of occurrence. Most frequent gets rank 1 and least frequent gets last rank.

\[ \text{Rank} \times \text{Frequency} \approx K \]

\[ \text{Rank} \times \frac{\text{Frequency}}{N} \approx \frac{K}{N} \text{ where } N = \text{total number of word occurrences.} \]

\[ \text{Rank} \times \text{Probability} \approx \text{Constant} \]

\[ \text{Constant} \approx 0.1 \]
### Zipf Table (from Allen and Croft, U. Mass)

#### Examples of Zipf

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq</th>
<th>Freq</th>
<th>Pr(%)</th>
<th>r^Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>2,420,778</td>
<td>1</td>
<td>6.488</td>
<td>0.0649</td>
</tr>
<tr>
<td>of</td>
<td>1,045,733</td>
<td>2</td>
<td>2.803</td>
<td>0.0561</td>
</tr>
<tr>
<td>to</td>
<td>968,882</td>
<td>3</td>
<td>2.597</td>
<td>0.0779</td>
</tr>
<tr>
<td>a</td>
<td>892,429</td>
<td>4</td>
<td>2.392</td>
<td>0.0957</td>
</tr>
<tr>
<td>and</td>
<td>865,644</td>
<td>5</td>
<td>2.32</td>
<td>0.116</td>
</tr>
<tr>
<td>in</td>
<td>847,825</td>
<td>6</td>
<td>2.272</td>
<td>0.1363</td>
</tr>
<tr>
<td>said</td>
<td>504,593</td>
<td>7</td>
<td>1.352</td>
<td>0.0947</td>
</tr>
<tr>
<td>for</td>
<td>363,865</td>
<td>8</td>
<td>0.975</td>
<td>0.078</td>
</tr>
<tr>
<td>that</td>
<td>347,072</td>
<td>9</td>
<td>0.93</td>
<td>0.0837</td>
</tr>
<tr>
<td>was</td>
<td>293,027</td>
<td>10</td>
<td>0.785</td>
<td>0.0785</td>
</tr>
<tr>
<td>on</td>
<td>291,947</td>
<td>11</td>
<td>0.783</td>
<td>0.0861</td>
</tr>
<tr>
<td>he</td>
<td>250,919</td>
<td>12</td>
<td>0.673</td>
<td>0.0807</td>
</tr>
<tr>
<td>is</td>
<td>245,843</td>
<td>13</td>
<td>0.659</td>
<td>0.0857</td>
</tr>
<tr>
<td>with</td>
<td>223,846</td>
<td>14</td>
<td>0.6</td>
<td>0.084</td>
</tr>
<tr>
<td>at</td>
<td>210,064</td>
<td>15</td>
<td>0.563</td>
<td>0.0845</td>
</tr>
<tr>
<td>by</td>
<td>209,586</td>
<td>16</td>
<td>0.562</td>
<td>0.0899</td>
</tr>
<tr>
<td>it</td>
<td>195,621</td>
<td>17</td>
<td>0.524</td>
<td>0.0891</td>
</tr>
<tr>
<td>from</td>
<td>189,451</td>
<td>18</td>
<td>0.508</td>
<td>0.0914</td>
</tr>
<tr>
<td>as</td>
<td>181,714</td>
<td>19</td>
<td>0.487</td>
<td>0.0925</td>
</tr>
<tr>
<td>be</td>
<td>157,300</td>
<td>20</td>
<td>0.422</td>
<td>0.0843</td>
</tr>
<tr>
<td>were</td>
<td>153,913</td>
<td>21</td>
<td>0.413</td>
<td>0.0866</td>
</tr>
<tr>
<td>an</td>
<td>152,576</td>
<td>22</td>
<td>0.409</td>
<td>0.09</td>
</tr>
<tr>
<td>have</td>
<td>149,749</td>
<td>23</td>
<td>0.401</td>
<td>0.0923</td>
</tr>
<tr>
<td>his</td>
<td>142,285</td>
<td>24</td>
<td>0.381</td>
<td>0.0915</td>
</tr>
<tr>
<td>but</td>
<td>140,880</td>
<td>25</td>
<td>0.378</td>
<td>0.0944</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq</th>
<th>Freq</th>
<th>Pr(%)</th>
<th>r^Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>has</td>
<td>136,007</td>
<td>26</td>
<td>0.365</td>
<td>0.0948</td>
</tr>
<tr>
<td>are</td>
<td>130,322</td>
<td>27</td>
<td>0.349</td>
<td>0.0943</td>
</tr>
<tr>
<td>not</td>
<td>127,493</td>
<td>28</td>
<td>0.342</td>
<td>0.0957</td>
</tr>
<tr>
<td>who</td>
<td>116,364</td>
<td>29</td>
<td>0.312</td>
<td>0.0904</td>
</tr>
<tr>
<td>they</td>
<td>111,024</td>
<td>30</td>
<td>0.298</td>
<td>0.0893</td>
</tr>
<tr>
<td>its</td>
<td>111,021</td>
<td>31</td>
<td>0.298</td>
<td>0.0922</td>
</tr>
<tr>
<td>had</td>
<td>103,943</td>
<td>32</td>
<td>0.279</td>
<td>0.0892</td>
</tr>
<tr>
<td>will</td>
<td>102,949</td>
<td>33</td>
<td>0.276</td>
<td>0.0911</td>
</tr>
<tr>
<td>would</td>
<td>99,503</td>
<td>34</td>
<td>0.267</td>
<td>0.0907</td>
</tr>
<tr>
<td>about</td>
<td>92,983</td>
<td>35</td>
<td>0.249</td>
<td>0.0872</td>
</tr>
<tr>
<td>i</td>
<td>92,005</td>
<td>36</td>
<td>0.247</td>
<td>0.0888</td>
</tr>
<tr>
<td>been</td>
<td>88,786</td>
<td>37</td>
<td>0.238</td>
<td>0.0881</td>
</tr>
<tr>
<td>this</td>
<td>87,286</td>
<td>38</td>
<td>0.234</td>
<td>0.0889</td>
</tr>
<tr>
<td>their</td>
<td>84,638</td>
<td>39</td>
<td>0.227</td>
<td>0.0885</td>
</tr>
<tr>
<td>new</td>
<td>83,449</td>
<td>40</td>
<td>0.224</td>
<td>0.0895</td>
</tr>
<tr>
<td>or</td>
<td>81,796</td>
<td>41</td>
<td>0.219</td>
<td>0.0899</td>
</tr>
<tr>
<td>which</td>
<td>80,385</td>
<td>42</td>
<td>0.215</td>
<td>0.0905</td>
</tr>
<tr>
<td>we</td>
<td>80,245</td>
<td>43</td>
<td>0.215</td>
<td>0.0925</td>
</tr>
<tr>
<td>more</td>
<td>76,388</td>
<td>44</td>
<td>0.205</td>
<td>0.0901</td>
</tr>
<tr>
<td>after</td>
<td>75,165</td>
<td>45</td>
<td>0.201</td>
<td>0.0907</td>
</tr>
<tr>
<td>us</td>
<td>72,045</td>
<td>46</td>
<td>0.193</td>
<td>0.0888</td>
</tr>
<tr>
<td>percent</td>
<td>71,956</td>
<td>47</td>
<td>0.193</td>
<td>0.0906</td>
</tr>
<tr>
<td>up</td>
<td>71,082</td>
<td>48</td>
<td>0.191</td>
<td>0.0915</td>
</tr>
<tr>
<td>one</td>
<td>70,266</td>
<td>49</td>
<td>0.188</td>
<td>0.0923</td>
</tr>
<tr>
<td>people</td>
<td>68,988</td>
<td>50</td>
<td>0.185</td>
<td>0.0925</td>
</tr>
</tbody>
</table>

Top 50 words from 84,678 Associated Press 1989 articles
(37,309,114 word occurrences, lowercased, punctuation removed, 266MB)
Figure 2.4. A plot of the hyperbolic curve relating $f$, the frequency of occurrence and $r$, the rank order (Adapted from Schultz, page 120)
Plotting Zipf on a log log scale

\[ \text{Rank} \times \text{Frequency} \approx K \]

\[ \text{Frequency} \approx \frac{K}{\text{Rank}} \] (or \( \text{Frequency} \approx K \times \text{Rank}^{-1} \) - power law)

\[ \log_{10}\text{Frequency} \approx \log_{10}(\frac{K}{\text{Rank}}) \]

\[ \log_{10}\text{Frequency} \approx \log_{10}K - \log_{10}\text{Rank} \]

\[ \log_{10}\text{Frequency} \approx \text{Constant} - \log_{10}\text{Rank} \]

When \( Y = \text{Constant} + BX \) is the form, then plotting \( Y \) against \( X \) gives

slope = \( B \) and intercept = \( \text{Constant} \) Here slope = \(-1\) and Constant = 5
(found empirically for Brown corpus)

\[ \log_{10}K = 5, \text{ i.e., } K = 10^5 = 100,000 \]
Logs (Review)

\[ \log_{\text{base}} x = y \implies \text{base}^y = x \]

(eg. \( \log_2 8 = 3 \implies 2^3 = 8 \))

Converting between logs, say you only know logs with base \( b \):

\[ \log_a X = \frac{\log_b X}{\log_b a} \]

Also:

\[ \log_b (X^Y) = Y \log_b X \]

\[ \log_b (XY) = \log_b X + \log_b Y \]

\[ \log_b \left( \frac{X}{Y} \right) = \log_b X - \log_b Y \]
Zipf and the Brown Corpus
[from R. Mooney, U.T. Austin]

$k = 100,000 \quad [Rank=1000, \log(1000)=3]$

Li (1992): random letters w/ space generates words with Zipfian distribution
Mandelbrot’s formula

\[ \text{Rank} \times \text{Frequency} \approx K \text{ (Zipf’s law)} \]

\[ K = 100,000 \text{ for Brown’s corpus} \]

\[ (\text{Rank} + \rho)^B \times \text{Frequency} \approx P \text{ (Mandelbrot’s adjustment)} \]

\[ P = 10^{5.4} \quad B = 1.15 \text{ and } \rho = 100 \text{ for Brown’s corpus} \]
Mandelbrot on log log plot

\[(\text{Rank} + \rho)^B \times \text{Frequency} \approx P \text{ (or } \text{Frequency} \approx P \times (\text{Rank} + \rho)^{-B} \text{ - power law)}\]

\[
\text{Frequency} \approx \frac{P}{(\text{Rank} + \rho)^B}
\]

\[
\log_{10} \text{Frequency} \approx \log_{10} \left[ \frac{P}{(\text{Rank} + \rho)^B} \right]
\]

\[
\log_{10} \text{Frequency} \approx \log_{10} P - \log_{10} (\text{Rank} + \rho)^B
\]

\[
\log_{10} \text{Frequency} \approx \log_{10} P - B \times \log_{10} (\text{Rank} + \rho)
\]

\[
\log_{10} \text{Frequency} \approx \text{Constant} - B \times \log_{10} (\text{Rank} + \rho)
\]

\[Y = \text{intercept} + (\text{slope})X\]

\[\text{slope} = -B = 1.15 \text{ and intercept } = \text{Constant} = \log_{10}(10^{5.4})\]
Mandelbrot and the Brown Corpus
[from R. Mooney, U.T. Austin]

Mandelbrot’s function on Brown corpus

\[ P = 10^{5.4}, B = 1.15, \rho = 100 \]
A few stopwords account for most of the word occurrences. So can reduce storage cost by eliminating these.

Unfortunately, for most words meaningful statistical analysis is tough since these are rare.

Up to half the words in a corpus appear only once, called hapax legomena (Greek for read only once)

Number of meanings of a word inversely related to frequency

Frequency of a word is inversely related to its length

Library book checkout patterns

Incoming Web page requests

Outgoing Web page requests (Cunha and Crovella)

Web page size
Luhn’s approach (1958): Resolution power

Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz [4] page 120)
Heap’s law: Vocabulary Growth

Vocabulary (V): the unique words in the vocabulary (Types versus tokens)

Growth - typos, new names, addition from other languages

N: size of corpus i.e., number of words in the corpus.

\[ V = K N^\beta \] where \( K \) and \( \beta \) depend upon the text collection.

\[ \beta \approx 0.4 - 0.6 \] (approximately square root)

\[ K \approx 10 - 100 \]

Basically vocabulary grows sublinearly with collection size in a proportion close to its square root.

Applies to both individual texts and to collections. Also to the web.
Heaps’ Law

[from R. Mooney, U.T. Austin]
Summary: We have looked at some of the basic features of text collections. Next week, more on word occurrences, and then we move on to phrases.