Zipf's Law

Phil Shinn
Mobile Voice 2014
Text Statistics

• Huge variety of words used in text but
• Many statistical characteristics of word occurrences are predictable
  – e.g., distribution of word counts
• Retrieval models and ranking algorithms depend heavily on statistical properties of words
  – e.g., important words occur often in documents but are not high frequency in collection
Professor George Kingsley Zipf

- Although he was a contemporary of Turing, there is no evidence the two ever met.
- Zipf died young too, at the age of 48, in 1950, only four years before Turing, but of natural causes.
- Independently wealthy, he paid human computers to count up the frequencies of words.
Zipf’s Law

• Distribution of word frequencies is very skewed
  – a few words occur very often, many words hardly ever occur
  – e.g., two most common words (“the”, “of”) make up about 10% of all word occurrences in text documents

• Zipf’s “law”:
  – observation that rank \( r \) of a word times its frequency \( f \) is approximately a constant \( k \)

assuming words are ranked in order of decreasing frequency
Log-Log Plot
Kucera & Francis, 1967
Zipf’s Law

- What is the proportion of words with a given frequency?
  - Word that occurs $n$ times has rank $r_n = k/n$
  - Number of words with frequency $n$ is
    - $r_n - r_{n+1} = k/n - k/(n+1) = k/n(n+1)$
  - Proportion found by dividing by total number of words = highest rank = $k$
  - So, proportion with frequency $n$ is $1/n(n+1)$
# Top 50 Words in Google Books

Corpus of 97,565 distinct words, which were mentioned 743,842,922,321

Norvig [http://norvig.com/mayzner.html](http://norvig.com/mayzner.html)

<table>
<thead>
<tr>
<th>WORD</th>
<th>COUNT</th>
<th>PERCENT</th>
<th>Bar Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>53.10</td>
<td>7.14%</td>
<td>the</td>
</tr>
<tr>
<td>of</td>
<td>30.97</td>
<td>4.16%</td>
<td>of</td>
</tr>
<tr>
<td>and</td>
<td>22.63</td>
<td>3.04%</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>19.35</td>
<td>2.60%</td>
<td>to</td>
</tr>
<tr>
<td>in</td>
<td>16.89</td>
<td>2.27%</td>
<td>in</td>
</tr>
<tr>
<td>a</td>
<td>15.31</td>
<td>2.06%</td>
<td>a</td>
</tr>
<tr>
<td>is</td>
<td>8.38</td>
<td>1.13%</td>
<td></td>
</tr>
<tr>
<td>that</td>
<td>8.00</td>
<td>1.08%</td>
<td>that</td>
</tr>
<tr>
<td>for</td>
<td>6.55</td>
<td>0.88%</td>
<td>for</td>
</tr>
<tr>
<td>it</td>
<td>5.74</td>
<td>0.77%</td>
<td>it</td>
</tr>
<tr>
<td>as</td>
<td>5.70</td>
<td>0.77%</td>
<td>as</td>
</tr>
<tr>
<td>was</td>
<td>5.50</td>
<td>0.74%</td>
<td>was</td>
</tr>
<tr>
<td>with</td>
<td>5.18</td>
<td>0.70%</td>
<td>with</td>
</tr>
<tr>
<td>be</td>
<td>4.82</td>
<td>0.65%</td>
<td>be</td>
</tr>
<tr>
<td>by</td>
<td>4.70</td>
<td>0.63%</td>
<td>by</td>
</tr>
<tr>
<td>on</td>
<td>4.59</td>
<td>0.62%</td>
<td>on</td>
</tr>
<tr>
<td>not</td>
<td>4.52</td>
<td>0.61%</td>
<td>not</td>
</tr>
<tr>
<td>he</td>
<td>4.11</td>
<td>0.55%</td>
<td>he</td>
</tr>
<tr>
<td>i</td>
<td>3.88</td>
<td>0.52%</td>
<td>i</td>
</tr>
<tr>
<td>this</td>
<td>3.83</td>
<td>0.51%</td>
<td>this</td>
</tr>
<tr>
<td>are</td>
<td>3.70</td>
<td>0.50%</td>
<td>are</td>
</tr>
<tr>
<td>or</td>
<td>3.67</td>
<td>0.49%</td>
<td>or</td>
</tr>
<tr>
<td>his</td>
<td>3.61</td>
<td>0.49%</td>
<td>his</td>
</tr>
<tr>
<td>from</td>
<td>3.47</td>
<td>0.47%</td>
<td>from</td>
</tr>
<tr>
<td>at</td>
<td>3.41</td>
<td>0.46%</td>
<td>at</td>
</tr>
<tr>
<td>which</td>
<td>3.14</td>
<td>0.42%</td>
<td>which</td>
</tr>
<tr>
<td>but</td>
<td>2.79</td>
<td>0.38%</td>
<td>but</td>
</tr>
<tr>
<td>have</td>
<td>2.78</td>
<td>0.37%</td>
<td>have</td>
</tr>
<tr>
<td>an</td>
<td>2.73</td>
<td>0.37%</td>
<td>an</td>
</tr>
<tr>
<td>had</td>
<td>2.62</td>
<td>0.35%</td>
<td>had</td>
</tr>
<tr>
<td>they</td>
<td>2.46</td>
<td>0.33%</td>
<td>they</td>
</tr>
<tr>
<td>you</td>
<td>2.34</td>
<td>0.31%</td>
<td>you</td>
</tr>
<tr>
<td>were</td>
<td>2.27</td>
<td>0.31%</td>
<td>were</td>
</tr>
<tr>
<td>their</td>
<td>2.15</td>
<td>0.29%</td>
<td>their</td>
</tr>
<tr>
<td>one</td>
<td>2.15</td>
<td>0.29%</td>
<td>one</td>
</tr>
<tr>
<td>all</td>
<td>2.06</td>
<td>0.28%</td>
<td>all</td>
</tr>
<tr>
<td>we</td>
<td>2.06</td>
<td>0.28%</td>
<td>we</td>
</tr>
<tr>
<td>can</td>
<td>1.67</td>
<td>0.22%</td>
<td>can</td>
</tr>
<tr>
<td>her</td>
<td>1.63</td>
<td>0.22%</td>
<td>her</td>
</tr>
<tr>
<td>has</td>
<td>1.63</td>
<td>0.22%</td>
<td>has</td>
</tr>
<tr>
<td>there</td>
<td>1.62</td>
<td>0.22%</td>
<td>there</td>
</tr>
<tr>
<td>been</td>
<td>1.62</td>
<td>0.22%</td>
<td>been</td>
</tr>
<tr>
<td>if</td>
<td>1.56</td>
<td>0.21%</td>
<td>if</td>
</tr>
<tr>
<td>more</td>
<td>1.55</td>
<td>0.21%</td>
<td>more</td>
</tr>
<tr>
<td>when</td>
<td>1.52</td>
<td>0.20%</td>
<td>when</td>
</tr>
<tr>
<td>will</td>
<td>1.49</td>
<td>0.20%</td>
<td>will</td>
</tr>
<tr>
<td>would</td>
<td>1.47</td>
<td>0.20%</td>
<td>would</td>
</tr>
<tr>
<td>who</td>
<td>1.46</td>
<td>0.20%</td>
<td>who</td>
</tr>
<tr>
<td>so</td>
<td>1.45</td>
<td>0.19%</td>
<td>so</td>
</tr>
<tr>
<td>no</td>
<td>1.40</td>
<td>0.19%</td>
<td>no</td>
</tr>
</tbody>
</table>
Letter Frequencies

Wikipedia

Relative frequencies ordered by frequency.
General Contemporary Chinese Corpus (GCCC), developed by the State Language Commission of China. The size of GCCC is around 1 billion Chinese characters with the diachronic materials from 1919 to 2005. 

U.S. City Populations
Other Countries


Figure 1: Zipf’s Law in Contemporary City Distributions
Firm Size – Number of Employees

http://www.brookings.edu/es/dynamics/papers/zipf/zipf.PDF

Figure 2: Distribution of U.S. firm sizes (by employees) for 1997, data combined from Census/SBA and Compustat together with self-employment data
Firm Size – By Revenue

**Figure 3:** Distribution of U.S. firm sizes (by revenue, in $ million) for 1997, data from Census
Wealth

Figure 2: Zipf's Law for the Richest Billionaires in the United States.

The richest 390 persons in the US are billionaires whose wealth we plot against their rank as the uppermost set of points (the first 195 richest being grey circles, the second 195 poorest being red circles). The second set is the sub-sample that we translate to the original ranks and plot as the set of red circle points below the diagonal straight line which is the pure Zipf plot associated with $x(k) = x_M/k$. The inset is a pure Zipf plot dimensioned to the entire set of 390 billionaires and the poorest sub-sample of 195. (Source: Forbes List http://www.forbes.com/).
Earthquakes
Books Sold, Phone Calls Received

Lunar Craters, Solar Flare Intensity

![Diagram of lunar craters and solar flare intensity](image)
U.S. Net Worth, Surnames

(j) net worth in US dollars

(k) name frequency
Figure 1. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links, measured in 1997.
Alicebot

http://www.alicebot.org/articles/wallace/zipf.html
Alicebot Exchanges

- 8024 YES
- 5184 NO
- 2268 OK
- 2006 WHY
- 1145 BYE
- 1101 HOW OLD ARE YOU
- 946 HI
- 934 HOW ARE YOU
- 846 WHAT
- 663 GOOD
- 645 WHY NOT
- 584 OH
- 553 REALLY
- 544 YOU
- 531 WHAT IS YOUR NAME
- 525 COOL
- 516 I DO NOT KNOW
- 488 FUCK YOU
- 486 THANK YOU
- 416 SO
- 414 ME TOO
- 403 LOL
- 403 THANKS
- 381 NICE TO MEET YOU TOO
- 375 SORRY
- 374 ALICE
- 353 OKAY
- 352 WHAT IS MY NAME
- 349 WHERE DO YOU LIVE
- 340 NOTHING
- 309 I KNOW
- 303 WHO ARE YOU
- 300 NOPE
- 297 SHUT UP
- 296 I LOVE YOU
- 288 SURE
- 286 HELLO ALICE
- 277 HOW
- 262 WHAT DO YOU MEAN
- 261 MAN
- 251 WOW
- 239 SMILE
- 233 ME
- 227 WHAT DO YOU LOOK LIKE
Our experiments with ALICE indicate that the number of choices for the "first word" is more than ten, but it is only about two thousand. Specifically, 1800 words covers 95% of all the first words input to ALICE. The number of choices for the second word is only about two. To be sure, there are some first words ("I" and "You" for example) that have many possible second words, but the overall average is just under two words. The average branching factor decreases with each successive word.

http://www.alicebot.org/articles/wallace/zipf.html
Explaining Zipf

“It is not known why Zipf's law holds for most languages... However, it may be partially explained by the statistical analysis of randomly generated texts. Wentian Li has shown that in a document in which each character has been chosen randomly from a uniform distribution of all letters (plus a space character), the "words" follow the general trend of Zipf's law (appearing approximately linear on log-log plot)...Zipf himself proposed that neither speakers nor hearers using a given language want to work any harder than necessary to reach understanding, and the process that results in approximately equal distribution of effort leads to the observed Zipf distribution.” - Wikipedia

Original goal of this work was to replicate Li’s work.
Process

Randomly generate a sequence of one letter, then generate another. If the next letter is a blank, you're done. Otherwise, choose another symbol. Continue until you have at least 2 (or 4) symbols.
Randomly Generated
Randomly Generated with Bias

2 symbol sequence
  prob blank = 0.3
  prob symbol 1 = 0.47
  prob symbol 2 = 0.2

4 symbol sequence
  prob blank = 0.2
  prob symbol 1 = 0.5
  prob symbol 2 = 0.13
  prob symbol 3 = 0.1
  prob symbol 4 = 0.07
Li Concludes

“In conclusion, Zipf's law is not a deep law of natural language as one might have first thought. It is very much related to the representation one has chosen, i.e., rank as an independent variable.”
But...

“Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution,” Ramon Ferrer-i-Cancho, Brita Elvevåg, Published: March 09, 2010, PLOS One

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009411
Figure 1. The rank histograms of English texts versus that of random texts ()

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0009411
Figure 2. The rank histograms of English texts versus that of random texts (A).
Figure 3. The rank histograms of English texts versus that of random texts ( ).

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0009411
Conclusion

The good fit of random texts to real Zipf's law-like rank distributions has not yet been established. Therefore, we suggest that Zipf's law might in fact be a fundamental law in natural languages.
Bibliography


“Random Texts Do Not Exhibit the Real Zipf’s Law-Like Rank Distribution,” Ramon Ferrer-i-Cancho, Brita Elvevåg, Published: March 09, 2010, PLOSOne


Hila Riemer, Suman Mallik, Devanathan Sudharshan, “Market Shares Follow the Zipf Distribution,”
http://www.business.illinois.edu/Working_Papers/papers/02-0125.pdf


Matthieu Cristelli, Michael Batte & Luciano Pietronero, “There is More than a Power Law in Zipf”