Text Technologies for Data Science
INFR11145

Laws of Text

Instructor:
Walid Magdy

26-Sep-2018

Lecture Objectives

• Learn about some text laws
  • Zipf’s law
  • Benford’s law
  • Heap’s law
  • Clumping/contagion

• This lecture is practical
You can try with me ...

- Shell commands: `cat`, `sort`, `uniq`, `grep`
- Perl (or alternative)
- Excel (or alternative)
- Download the following:
  - Bible: [http://www.gutenberg.org/cache/epub/10/pg10.txt](http://www.gutenberg.org/cache/epub/10/pg10.txt)

Words’ nature

- Word → basic unit to represent text
- Certain characteristics are observed for the words we use!
- These characteristics are very consistent, that we can apply laws for them
- These laws apply for:
  - Different languages
  - Different domains of text
Frequency of words

- Some words are very frequent e.g. “the”, “of”, “to”
- Many words are less frequent e.g. “schizophrenia”, “bazinga”
- ~50% terms appears once
- Frequency of words has hard exponential decay

Zipf’s Law:

- For a given collection of text, ranking unique terms according to their frequency, then:

  \[ r \times P_r \cong const \]

- \( r \), rank of term according to frequency
- \( P_r \), probability of appearance of term

- \( P_r \cong \frac{\text{const}}{r} \rightarrow f(x) \cong \frac{1}{x} \)
Zipf’s Law:

Wikipedia abstracts → 3.5M En abstracts

\[ r \times P_r \cong \text{const} \]

\[ r \times freq_r \cong \text{const} \]

<table>
<thead>
<tr>
<th>Term</th>
<th>Rank</th>
<th>Frequency</th>
<th>( r \times \text{freq} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1</td>
<td>5,134,790</td>
<td>5,134,790</td>
</tr>
<tr>
<td>of</td>
<td>2</td>
<td>3,102,474</td>
<td>6,204,948</td>
</tr>
<tr>
<td>in</td>
<td>3</td>
<td>2,607,875</td>
<td>7,823,625</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
<td>2,492,328</td>
<td>9,969,312</td>
</tr>
<tr>
<td>is</td>
<td>5</td>
<td>2,181,502</td>
<td>10,907,510</td>
</tr>
<tr>
<td>and</td>
<td>6</td>
<td>1,962,326</td>
<td>11,773,956</td>
</tr>
<tr>
<td>was</td>
<td>7</td>
<td>1,159,088</td>
<td>8,113,616</td>
</tr>
<tr>
<td>to</td>
<td>8</td>
<td>1,088,396</td>
<td>8,707,168</td>
</tr>
<tr>
<td>by</td>
<td>9</td>
<td>766,656</td>
<td>6,899,904</td>
</tr>
<tr>
<td>an</td>
<td>10</td>
<td>566,970</td>
<td>5,669,700</td>
</tr>
<tr>
<td>it</td>
<td>11</td>
<td>557,492</td>
<td>6,132,412</td>
</tr>
<tr>
<td>for</td>
<td>13</td>
<td>493,374</td>
<td>5,970,456</td>
</tr>
<tr>
<td>as</td>
<td>14</td>
<td>480,277</td>
<td>6,413,862</td>
</tr>
<tr>
<td>on</td>
<td>15</td>
<td>471,544</td>
<td>6,723,878</td>
</tr>
<tr>
<td>from</td>
<td>16</td>
<td>412,785</td>
<td>7,073,160</td>
</tr>
</tbody>
</table>

Distribution of first digit in frequencies?

1) Uniform →

2) Exp decay →

3) Normal →
**Benford’s Law:**

- First digit of a number follows a Zipf’s like law!
  - Terms frequencies
  - Physical constants
  - Energy bills
  - Population numbers

- Benford’s law:
  \[ P(d) = \log(1 + \frac{1}{d}) \]

**Heap’s Law:**

- While going through documents, the number of new terms noticed will reduce over time
- For a book/collection, while reading through, record:
  - \( n \): number of words read
  - \( v \): number of new words (unique words)

- Vocabulary growth:
  \[ v(n) = k \times n^b \]
  where, \( b < 1 \)
  typically, \( 0.4 < b < 0.7 \)
Heap’s Law: shouldn’t it saturate?

- $n = 80+$ million, but still growing
- Think about:
  - spelling errors
  - names
  - emails
  - codes
- Accurate for most collections, but different $k, b$
- Not very accurate when $n$ is small

Clumping/Contagion in text

- From Zipf’s law, we notice:
  - Most words do not appear that much!
  - Once you see a word once $\rightarrow$ expect to see again!
  - Words are like:
    - Rare contagious disease
    - Not, rare independent lightening
- Words are rare events, but they are contagious
Clumping/Contagion in text

- Wiki abstract collection
  - Identify terms appeared only twice
  - Measure distance between the two occurrences of the terms:
    \[ d = n_{occurrence\,2} - n_{occurrence\,1} \]
  - Plot density function of \( d \)

- Majority of terms appearing only twice appear close to each other.

Applying the laws

- Given a collection of 20 billion terms,
- What is the number of unique terms?
  
  Heap’s law: \( v(n) = k \times n^b \), assume \( k = 0.25 \), \( b = 0.5 \)
  
  \[ v(n) = 0.25 \times (20B)^{0.5} \approx 35M \]

- What is the number of terms appearing once?
  
  Zipf’s law \( \Rightarrow \sim 17M \) appeared only once
Summary

• Text follows well-known phenomena
• Text Laws:
  • Zipf
  • Heap
  • Contagion in text

Recourses

• Text book:
  • Search engines: IR in practice → chapter 4
• Videos:
  • Zipf’s law, Vsouce:
    https://www.youtube.com/watch?v=fCn8zs912OE
  • Benford’s law, Numberphile:
    https://www.youtube.com/watch?v=XXjIR2Qk1kM
• Tools:
  • Unix commands for windows
    https://sourceforge.net/projects/unxutils