Text and words

- **Word** – basic unit for representing text
  - reflects meaning
  - BOW retains context

- **Word occurrence governed by statistical laws**
  - **Zipf’s law**: frequency and rarity of words
  - **Heaps’ law**: rate at which new words will appear
  - **Dependence**: clumpy / contagious nature of words

---

Zipf’s law

- **Observation**: frequent words and rare words
  - “of” and “the” make up 10% of all occurrences
  - hardly ever see “aardvark”
- **Rank words by frequency**
- **Zipf’s law**:
  - rank of the word times its probability (frequency) is approximately a constant
  - \( r \times P_r = \text{const} \)

---

Mandelbrot’s law

- **Improvement to Zipf’s law**: \( P_r = c / (r+q)^b \)
  - \( q, b \): parameters tuned to fit the data
  - \( b = 1 \) surprisingly often
  - allows infinite vocabulary (under Zipf: \( \Sigma P_r = \infty \))
  - better fit to frequent words
- **General family**:
  - power-law distributions
  - Zipf / Mandelbrot, Benford, Laval et al. → Pareto, Cauchy, Zeta distribution (\( q=0 \))

---

Heaps’ law

- **Experiment**: read a book / newspaper / website
  - record every time you see a new word
  - \( n \)… number of new words seen, \( n \)… total words
  - plot \( n \) against \( n \)
- **Vocabulary growth**
  - \( v = k \times n^b \)
  - \( b = 0.5 \)
  - should level off eventually…right?

---

Heaps’ law: to infinity

- still seeing new words after 30 million
- spelling errors, products, person/company names, email, code

- accurate for many collections
  - different parameters \( k, b \)
  - sometimes poor for small \( n \)
Clumping / contagion

- Word occurrences: rare but "contagious" events
  - a-priori, you're very unlikely to see a given word
  - see it once → much more likely to see again
- "rare contagious diseases" rather than "lightning strikes"

*said* in Brown Corpus

<table>
<thead>
<tr>
<th>Brown Corpus</th>
<th>Binomial Prediction for &quot;said&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Applying the laws

- Given a 200Gb crawl of English web pages
  - what can we guess about resulting index size?

- Heaps:
  - ~35m unique index terms (~20b total words)
- Zipf / Mandelbrot:
  - ~17m will have one entry, "the" will have >10m
- Clumping:
  - how often we should expect >1 entry for rare words

Estimating index size

- How many pages does engine X have?
  - X allows us to run a few queries, reports counts
    - let $a, b$ = words known to occur independendty
    - run queries: \{a\}, \{b\}, \{a, b\}, get counts: $n_a$, $n_b$, $n_{ab}$
    - since $a, b$ independent:
      $$\frac{n_{ab}}{n} = P(a, b) = P(a)P(b) = \frac{n_a}{n} \times \frac{n_b}{n} \Leftrightarrow \frac{n - n_{ab}}{n} = \frac{n_a}{n} \cdot \frac{n_b}{n}$$
    - repeat many times, with frequent $a, b$
- Use sample $c$ to guess
  - correlation between $a, b$

Estimating Bing index size

- right, walk: $2,060M \times 678M / 328M = 4.3B$
- green, john: $1,190M \times 1,130M / 188M = 7.1B$
- big, weather: $1,820M \times 673M / 155M = 7.9B$
- black, hawaii: $1,600M \times 251M / 67M = 5.9B$
- first, water: $2,740M \times 1,330M / 729M = 4.9B$
- walk, january: $865M \times 669M / 109M = 5.3B$
- fly, pink: $436M \times 507M / 52M = 4.3B$
- cool, bed: $768M \times 808M / 96M = 6.5B$

→ probably contains around 6B English pages

Summary

- Nature of text
  - Zipf's law: \( r \times P_r = \text{const} \)
  - Mandelbrot's law: \( P_r = c / (r+q)^k \)
  - Heaps law: \( v = k \times n^p \)
  - clumping: words ~ rare infectious diseases
  - index size: \( n = n_a n_b / n_{ab} \)