Full text Indexing: index all text with uncontrolled vocabulary

- Advantages
  - (Possibly) Keep all the information within the text
  - Often no human efforts; easy to build
- Disadvantages
  - Difficult to cross vocabulary gap (e.g., “cancer” in query, “neoplasm” in document)
  - Large storage space

How to build full text Index:

- What are the candidates in the word vocabulary? Are they effective to represent semantic meanings
- How to bridge small vocabulary gap (e.g., car and cars)
Text Representation: Indexing
Statistical Properties of Text

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1130021</td>
<td>market</td>
<td>52110</td>
</tr>
<tr>
<td>of</td>
<td>547311</td>
<td>bank</td>
<td>47940</td>
</tr>
<tr>
<td>to</td>
<td>516636</td>
<td>stock</td>
<td>47401</td>
</tr>
<tr>
<td>a</td>
<td>464736</td>
<td>trade</td>
<td>47310</td>
</tr>
<tr>
<td>in</td>
<td>390819</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>and</td>
<td>387703</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Statistics collected from Wall Street Journal (WSJ), 1987
Text Representation: Indexing
Statistical Properties of Text

- Observations from language/corpus independent features
- A few words occur very frequently (High Peak)
  - Top 2 words: 8%-15% (e.g., words that carry no semantic meanings like “the”, “to”)
- Most words occur rarely (Heavy Tail)
- Representative words often in the middle
  - e.g., market and stock for WSJ
- Rules formally describe word occurrence patterns: Zipf’s law, Heaps’ Law

---

Zipf’s law: relate a term’s frequency to its rank

- Rank all terms with their frequencies in descending order, for a term at a specific rank (e.g., \( r \)) collects and calculates
  
  \[
  f_r : \text{term frequency} \quad p_r = \frac{f_r}{N} : \text{relative term frequency}
  \]

- Zipf’s law (by observation):
  
  \[
  p_r = \frac{A}{r} \quad A \approx 0.1
  \]

  So \( p_r = \frac{f_r}{N} \Rightarrow rf_r = AN \Rightarrow \log(r) = -\log(f_r) + \log(AN) \)

  So Rank X Frequency = Constant
Text Representation: Indexing
Statistical Properties of Text

Statistics collected from Wall Street Journal (WSJ), 1987

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>$r \cdot p_r$</th>
<th>Word</th>
<th>Frequency</th>
<th>$r \cdot p_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>113021</td>
<td>0.059</td>
<td>market</td>
<td>52110</td>
<td>0.101</td>
</tr>
<tr>
<td>of</td>
<td>547311</td>
<td>0.058</td>
<td>bank</td>
<td>47940</td>
<td>0.109</td>
</tr>
<tr>
<td>to</td>
<td>516636</td>
<td>0.082</td>
<td>stock</td>
<td>47401</td>
<td>0.110</td>
</tr>
<tr>
<td>a</td>
<td>464736</td>
<td>0.098</td>
<td>trade</td>
<td>47310</td>
<td>0.112</td>
</tr>
<tr>
<td>in</td>
<td>390819</td>
<td>0.103</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>and</td>
<td>387703</td>
<td>0.122</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Term Rank

Term Frequency
Text Preprocessing:

- Extract representative index terms
- Parse query/document for useful structure
  - E.g., title, anchor text, link, tag in xml
- Tokenization
  - For most western languages, words separated by spaces; deal with punctuation, capitalization, hyphenation
  - For Chinese, Japanese: more complex word segmentation
- Remove stopwords: (remove “the”, “is”, ..., existing standard list)
- Morphological analysis (e.g., stemming):
  - Stemming: determine stem form of given inflected forms
- Other: extract phrases; decompounding for some European languages

rörelseuppskattningssökningsintervallsinställningar

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 the</td>
<td>1</td>
<td>1 at</td>
<td>1</td>
<td>1 different</td>
<td>1</td>
<td>1 may</td>
<td>1</td>
<td>1 step</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 and</td>
<td>1</td>
<td>1 basal</td>
<td>1</td>
<td>1 exchange</td>
<td>1</td>
<td>1 nontarget</td>
<td>1</td>
<td>1 substance</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 by</td>
<td>1</td>
<td>1 be</td>
<td>1</td>
<td>1 exogenous</td>
<td>1</td>
<td>1 not</td>
<td>1</td>
<td>1 suggests</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 steroids</td>
<td>1</td>
<td>1 been</td>
<td>1</td>
<td>1 fluorescent</td>
<td>1</td>
<td>1 may</td>
<td>1</td>
<td>1 target</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 centrioles</td>
<td>1</td>
<td>1 bodies</td>
<td>1</td>
<td>1 from</td>
<td>1</td>
<td>1 of</td>
<td>1</td>
<td>1 technique</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 in</td>
<td>1</td>
<td>1 can</td>
<td>1</td>
<td>1 growth</td>
<td>1</td>
<td>1 precise</td>
<td>1</td>
<td>1 two</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 affect</td>
<td>1</td>
<td>1 at</td>
<td>1</td>
<td>1 has</td>
<td>1</td>
<td>1 receptor</td>
<td>1</td>
<td>1 unexpected</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 already</td>
<td>1</td>
<td>1 cell</td>
<td>1</td>
<td>1 identity</td>
<td>1</td>
<td>1 regularly</td>
<td>1</td>
<td>1 vitally</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Although</td>
<td>1</td>
<td>1 cells</td>
<td>1</td>
<td>1 level</td>
<td>1</td>
<td>1 reveal</td>
<td>1</td>
<td>1 way</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 antibodies</td>
<td>1</td>
<td>1 cilia-bearing</td>
<td>1</td>
<td>1 localization</td>
<td>1</td>
<td>1 Specific</td>
<td>1</td>
<td>1 with</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24 stopwords out of total 61 words
Text Representation: Bag of Words

The simplest text representation: “bag of words”

- Query/document: a bag that contains words in it
- Order among words is ignored

```
steroids  substance  growth  ......  steroids
......  exchange  ......  nontarget
step  ......  bodies  two  precise  ......
```

| 3 steroids | 1 cell-bending | 1 precise | 1 two |
| 2 centrioles | 1 different | 1 receptor | 1 unexpected |
| 1 affect | 1 exchange | 1 regularly | 1 slightly |
| 1 already | 1 exceptional | 1 reveal | 1 way |
| 1 Although | 1 fluorescent | 1 Specific |
| 1 antibody | 1 growth | 1 step |
| 1 basal | 1 identity | 1 substance |
| 1 bodies | 1 level | 1 suggests |
| 1 cell | 1 localization | 1 target |
| 1 cells | 1 nontarget | 1 technique |

Text Representation: Phrases

- Single word/stem indexing may not be sufficient
  - e.g., “hit a home run yesterday”
- More complicated indexing includes phrases (thesaurus classes)
- How to automatically identify phrases
  - Dictionary
  - Find the most common N word phrases by corpus statistics (be careful of stopwords)
  - Syntactic analysis, noun phrases
  - More sophisticated segmentation algorithm like “Hidden Markov Model”
Text Representation: Word Stemming

- **Word Stemming**
  - Associate morphological variants of words into a single form
    - E.g., plurals, adverbs, inflected word forms
    - May lose the precise meaning of a word

- Different types of stemming algorithms
  - Rule-based systems: Porter Stemmer, Krovetz Stemmer
  - Porter Stemmer Example: describe/describes -> describ
  - Statistical method: Corpus-based stemming

---

**Porter Stemmer**

- Based on a pattern of vowel-consonant sequence
  - [C](VC)m[V], m is an integer
- Rules are divided into steps and examined in sequence
  - Step 1a: ies → i; s → ; …
    - cares → care
  - Step 1b: if m>0 eed → ee
    - agreed → agree
  - … Step 5a, Step 5b
- Pretty aggressive:
  - nativity → native
Text Representation: Word Stemming

K Stemmer: based on morphological rules
• If word occurs in a dictionary, do not stem it
• For all other words
  – Remove inflectional endings: plurals to singular; past tense to present tense; remove “ing”
  – Remove derivational endings by a sequence of rules: may make mistake when suffixes indicate different meanings like “sign” to “signify”

Examples of Stemming:
• Original Text:
  – Information retrieval deals with the representation, storage, organization of, and access to information items

• Porter Stemmer (Stopwords removed):
  – Online example:
    http://facweb.cs.depaul.edu/mobasher/classes/csc575/porter.html
  – Inform retrieve deal represent storag organ access inform item
Problems with Rule-based Stemming

• Rule-based stemming may be too aggressive
  – e.g., execute/executive, university/universe

• Rule-based stemming may be too conservative
  – e.g., European/Europe, matrices/matrix

• Difficult to understand the meaning the stems
  – e.g., Iteration/iter, general/gener

Corpus-Based Stemming

• Hypothesis: Word variants that should be considered equally often co-occur in documents (passages or text windows) in the corpus
  – Collect the statistics of co-occurrence of words in the corpus and form the connected graph
  – Cut the graph by different methods and find the connected subgraphs to form equivalence classes
Text Representation: Word Stemming

(Xu & Croft, 1998)