Day 3: Descriptive Inference from Text

Kenneth Benoit

Essex Summer School 2011

July 13, 2011

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

word frequency refers to the number of times that words occur in a text or in a *corpus* of texts

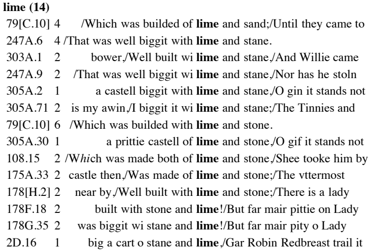
concordance a(n alphabetical) list of the principal words used in a text, with their immediate contexts

lemmas the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached. *Lemmatization* refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

Some key basic concepts

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.



lime (14)

79[C.10]	4	/Which was builded of	**lime**	and sand;/Until they came to
247A.6	4	/That was well biggit with	**lime**	and stane.
303A.1	2	bower,/Well built wi	**lime**	and stane,/And Willie came
247A.9	2	/That was well biggit wi	**lime**	and stane,/Nor has he stoln
305A.2	1	a castell biggit with	**lime**	and stane,/O gin it stands not
305A.71	2	is my awin,/I biggit it wi	**lime**	and stane;/The Tinnies and
79[C.10]	6	/Which was builded with	**lime**	and stone.
305A.30	1	a prittie castell of	**lime**	and stone,/O gif it stands not
108.15	2	/W*hi*ch was made both of	**lime**	and stone,/Shee tooke him by
175A.33	2	castle then,/Was made of	**lime**	and stone;/The vttermost
178[H.2]	2	near by,/Well built with	**lime**	and stone;/There is a lady
178F.18	2	built with stone and	**lime**	!/But far mair pittie on Lady
178G.35	2	was biggit wi stane and	**lime**	!/But far mair pity o Lady
2D.16	1	big a cart o stane and	**lime**	,/Gar Robin Redbreast trail it

stop words Words that are designated for exclusion from any analysis of a text

Some useful linguistic terms

From a field known as *corpus linguistics*

type for our purposes, a unique word

token any word – so token count is total words

hapax legomena (or just *hapax*) are types that occur just once

Some key basic concepts

readability provides estimates of the readability of a text based on word length, syllable length, etc.

- Fog Index, developed by Robert Gunning, indicates the number of years of formal education required to read and understand a passage of text
- Flesch Index, developed in 1940 by Dr. Rudolph Flesch, is based on a 100 point scale, with 100 being easiest to read
- Flesch-Kincaid Index is a refinement to the Flesch Index that relates the score to a U.S. grade level

(more on how these are computed shortly)

complexity A word is considered "complex" if it contains three syllables or more

Some key basic concepts

term frequency is a normalized count of the number of times a particular term appears in a document. The normalization occurs by dividing the term's frequency by the total frequency of all terms in that document

inverse document frequency is the (logarithm) of the total number of documents in the corpus, divided by the total number of documents where a given term appears

tf-idf is the term frequency multiplied by the inverse document frequency, and measured the commonness of words – typically used to filter out the most common words from a corpus prior to analysis

Word concordances on popular web sites

- Amazon word statistics example `http://www.amazon.com/Innovative-Comparative-Methods-Policy-Analysis/dp/0387288287/ref=sr_1_1?ie=UTF8\&s=books\&qid=1249293340\&sr=8-1`

- New York Times inaugural address example: `http://www.nytimes.com/interactive/2009/01/17/washington/20090117_ADDRESSES.html`

Basic descriptive summaries of text

Readability statistics Use a combination of syllables and sentence length to indicate "readability" in terms of complexity

Vocabulary diversity (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Word (relative) frequency

Theme (relative) frequency

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

  $$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

  Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

- Flesch-Kincaid rescales to the US educational grade levels (1–12):

  $$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

Gunning fog index

- Measures the readability in terms of the years of formal education required for a person to easily understand the text on first reading
- Usually taken on a sample of around 100 words, not omitting any sentences or words
- Formula:

$$0.4\left[\left(\frac{\text{total words}}{\text{total sentences}}\right)+100\left(\frac{\text{complex words}}{\text{total words}}\right)\right]$$

  where complex words are defined as those having three or more syllables, not including proper nouns (for example, Ljubljana), familiar jargon or compound words, or counting common suffixes such as -es, -ed, or -ing as a syllable

Word frequency as an indicator of substantive content

- Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Atomic words have been found to be far more informative than *n*-grams in this regard (Benoit and Laver 2003, Midwest paper)
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome (e.g. Hopkins and King 2008)
- Other approaches use frequencies: Poisson, multinomial, and related distributions (e.g. Laver, Benoit and Garry 2003)

Word frequency: Zipf's Law

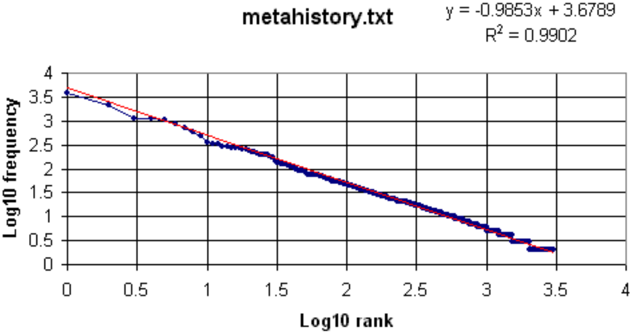
- Zipf's law: Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.
- The simplest case of Zipf's law is a "1/f function". Given a set of Zipfian distributed frequencies, sorted from most common to least common, the second most common frequency will occur $1/2$ as often as the first. The third most common frequency will occur $1/3$ as often as the first. The nth most common frequency will occur $1/n$ as often as the first.
- In the English language, the probability of encountering the the most common word is given roughly by $P(r) = 0.1/r$ for up to 1000 or so
- The assumption is that words and phrases mentioned most often are those reflecting important concerns in every communication

Word frequency: Zipf's Law

- Formulaically: if a word occurs f times and has a rank r in a list of frequencies, then for all words $f = \frac{a}{r^b}$ where a and b are constants and b is close to 1

- So if we log both sides, $\log(f) = \log(a) - b\log(r)$

- If we plot $\log(f)$ against $\log(r)$ then we should see a straight line with a slope of approximately -1.



Word frequency continued

- Some approaches trim low-frequency words or words that are non-discriminating among texts
- Frequently this is based on a measure of word frequency known as *tf-idf*: term frequency-inverse document frequency
- Rationale behind filtering out words based on frequency
  - Substantive: Non-discriminating words (articles, conjunctions, pronouns, etc.) are non-informative
  - Practical: Non-discriminating words may strain computational abilities of particular statistical or computational techniques, esp. those requiring word frequency matrix analysis
  - Substantive: Low-frequency words may simply not be worth bothering about

Computation of tf-idf

- $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
  where $n_{i,j}$ is number of occurences of term t_i in document d_j, k is total number of terms in document d_j

- $idf_i = \ln \frac{|D|}{|\{d_j : t_i \in d_j\}|}$
  where
  - $|D|$ is the total number of documents in the set
  - $|\{d_j : t_i \in d_j\}|$ is the number of documents where the term t_i appears (i.e. $n_{i,j} \neq 0$)

- $tf\text{-}idf_i = tf_{i,j} \cdot idf_i$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$
- If the word had only appeared in 15 of the 100 manifestos, then the *tf-idf* would be 0.0304 (three times higher).
- A high weight in tf-idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; hence the weights hence tend to filter out common terms