Day 2: Textual Data, Sampling, and Working with Texts

Kenneth Benoit

Essex Summer School 2013

July 23, 2013

Day 2 Basic Outline

- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts
- Selecting features
- Practical issues working with texts
- Demonstrations
- Examples

BUILDING BLOCKS

Some key basic concepts

Some key basic concepts

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

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

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

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.

Some key basic concepts

Some key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

Some key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

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

Some key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

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

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

Some key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

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

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

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

VALIDITY OF FEATURE FREQUENCY APPROACHES

Word frequency as an indicator of substantive content

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

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)

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)

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

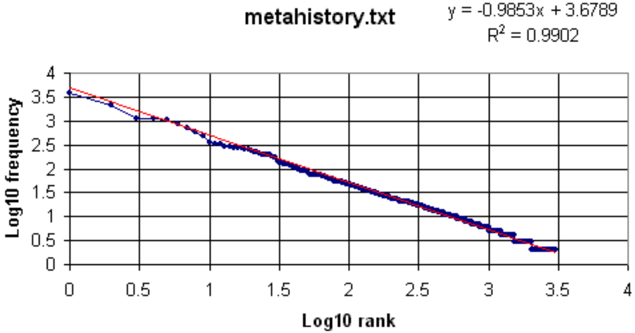
- Formulaically: if a word occurs f times and has a rank r in a list of frequencies, then for all words $f = \frac{a}{r^b}$ where a and b are constants and b is close to 1

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, $\log(f) = \log(a) - b\log(r)$

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

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

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

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

Word frequency examples

- Variations use vocabulary diversity analysis (e.g. Labbé et. al. 2004)

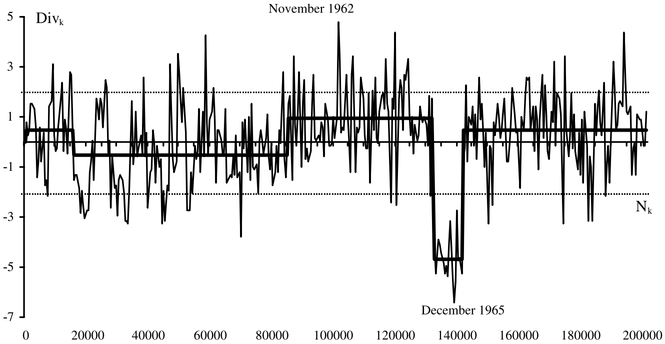
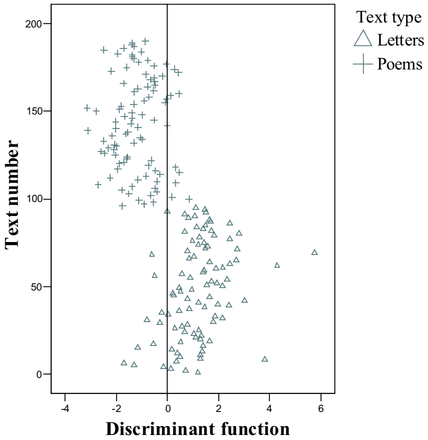


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Examples continued

- Word *length* (defined as number of syllables) can be indicative of genre, if not necessarily authorship (Kelih et. al. 2004)



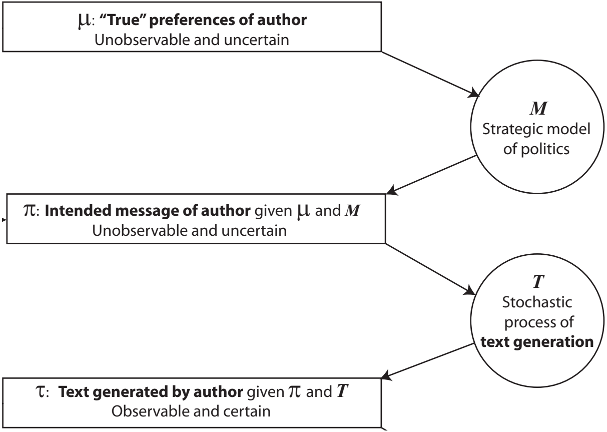
SELECTING TEXTS AND UNITS

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Themes
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Sample v. "population"

- Basic Idea: Observed text is a stochastic realization
- Systematic features shape most of observed verbal content
- Non-systematic, random features also shape verbal content



μ: **"True" preferences of author**
Unobservable and uncertain
M
Strategic model of politics
π: **Intended message of author** given μ and *M*
Unobservable and uncertain
T
Stochastic process of **text generation**
τ: **Text generated by author** given π and *T*
Observable and certain

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive
  - random sampling
  - non-random sampling
- Key is to make sure that what is being analyzed is a valid representation of the phenomenon as a whole – a question of research design

Random versus “Constructed” Sampling

- Based on a study by Riffe, Aust and Lacy (1993), who compared sampling from newspaper articles randomly versus “constructed”
- Either randomly sample 7 consecutive days, or between 2–4 consecutive weeks, and compare to “known” quantities
- Study showed that constructed sampling is much more efficient
- Why? Because cyclic variation in newspaper content occurs according to the day of the week – not every day contains equal proportions of different content

SELECTING FEATURES

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- purposive selection Use of a dictionary of words or phrases
- deliberate disregard Use of “stop words”: words excluded because they represent linguistic connectors of no substantive content

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

- But no list should be considered universal

A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, aint, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, arent, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, cmon, cs, came, can, cant, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldnt, course, currently, definitely, described, despite, did, didnt, different, do, does, doesnt, doing, dont, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadnt, happens, hardly, has, hasnt, have, havent, having, he, hes, hello, help, hence, her, here, heres, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, id, ill, im, ive, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isnt, it, itd, itll, its, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, lets, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldnt, since, six, so, some, somebody,

Strategies for feature *weighting*: 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 $0.016 * 0.916 = 0.0147$

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).

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

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

Stemming words

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.

Stemming words

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.

both convert the morphological variants into stem or root terms

Stemming words

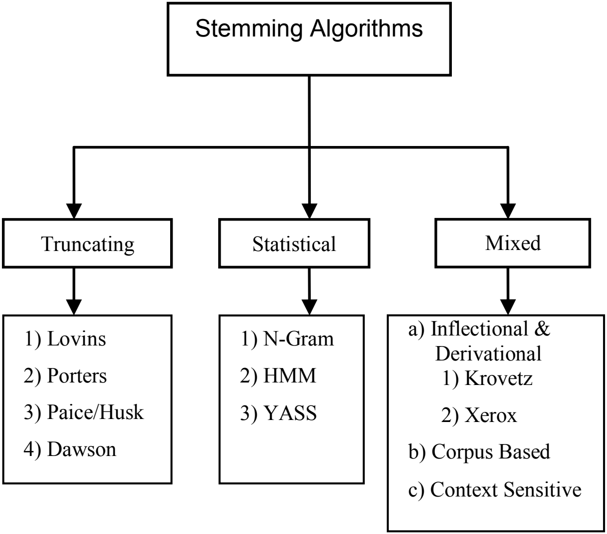
Lemmatization refers to the algorithmic process of converting words to their lemma forms.

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.

both convert the morphological variants into stem or root terms

example: `produc` from `production, producer, produce, produces, produced`

Varieties of stemming algorithms



Issues with stemming approaches

- The most common is proably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`
- Other corpus-based, statistical, and mixed appraoches designed to overcome these limitations (good review in Jirvani article)
- Key for you is to be careful through inspection of morphological variants and their stemmed versions

Selecting more than words: collocations

collocations bigrams, or trigrams e.g. *capital gains tax*

how to detect: pairs occuring more than by chance, by measures of χ^2 or *mutual information* measures

example:

Summary Judgment	Silver Rudolph	Sheila Foster
prima facie	COLLECTED WORKS	Strict Scrutiny
Jim Crow	waiting lists	Trail Transp
stare decisis	Academic Freedom	Van Alstyne
Church Missouri	General Bldg	Writings Fehrenbacher
Gerhard Casper	Goodwin Liu	boot camp
Juan Williams	Kurland Gerhard	dated April
LANDMARK BRIEFS	Lee Appearance	extracurricular activities
Lutheran Church	Missouri Synod	financial aid
Narrowly Tailored	Planned Parenthood	scored sections

Table 5: Bigrams detected using the mutual information measure.

PRACTICAL ISSUES WORKING WITH TEXT

Practical issues working with texts

Practical issues working with texts

File formats How the electronic text is formatted

Practical issues working with texts

File formats How the electronic text is formatted

Conversion Converting files from one format to another

Practical issues working with texts

File formats How the electronic text is formatted

Conversion Converting files from one format to another

Pre-analysis text processing Considering inflected forms as equivalent, through lemmatization and/or stemming

Practical issues working with texts

File formats How the electronic text is formatted

Conversion Converting files from one format to another

Pre-analysis text processing Considering inflected forms as equivalent, through lemmatization and/or stemming

dropping infrequent words as they may not be informative

Practical issues working with texts

File formats How the electronic text is formatted

Conversion Converting files from one format to another

Pre-analysis text processing Considering inflected forms as equivalent, through lemmatization and/or stemming

dropping infrequent words as they may not be informative

stop lists for most frequent words

Practical issues working with texts: Generating “datasets”

Practical issues working with texts: Generating "datasets"

- Raw data is always the text file
  - This is the "corpus" in textual form, prior to conversion to a quantitative feature matrix
  - Sometimes, we wish also to preserve the pre-text formatted file containing the text (e.g., pdf)
- We need to preserve the rules used to create the quantitative matrix from the text files

DEMONSTRATIONS

Software preview

- QDAMiner/Wordstat
- `quanteda` in R