Quantitative text analysis overview and fundamentals

Kenneth Benoit

MY 459/559: Quantitative Text Analysis

January 11, 2016

Outline

- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Targets

- Whom this class is for
- Learning objectives
  - fundamentals
  - availability and consequences of *choices*
  - practical ability to work with texts
  - issues of text for social science
- Prequisites
  - quantitative methods
  - familiarity with R
  - ability to use a text editor
  - (optional) ability to process text files in a programming language such as Python

Basic QTA Process: Texts → Feature matrix → Analysis

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

docs	words made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Outline

- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Course resources

- Syllabus: describes class, lists readings, links to reading, and links to exercises and datasets
- Moodle page
  - Contains course handout
  - Slides from class
  - In-class exercises and supporting materials
  - Texts for analysis
  - (links to) Software tools and instructions for use
- Main readings
  - Mainly articles
  - Available on Perusall.com – enroll using code `BENOIT-4372`
  - Some other texts or on-line articles linked to the course handout

Course resources (cont.)

- Software: R
  - the quanteda package
  - additional resources: GitHub issues, StackOverflow channel

Outline

- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Quantitative text analysis requires assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
  - most common is the bag of words assumption
  - many other possible definitions of "features"
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Key feature of quantitative text analysis

1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the doumentary unit of analysis

Key feature of quantitative text analysis (cont.)

4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. Conversion of textual features into a quantitative matrix
6. A quantitative or statistical procedure to extract information from the quantitative matrix
7. Summary and interpretation of the quantitative results

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

docs	words made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

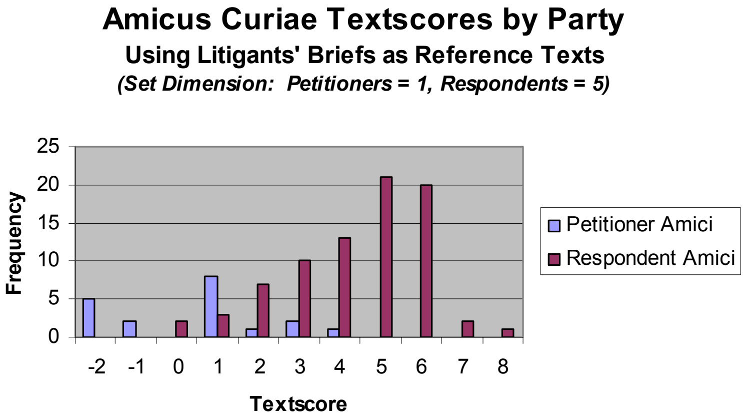
Vocabulary analysis

Sentiment analysis

Outline

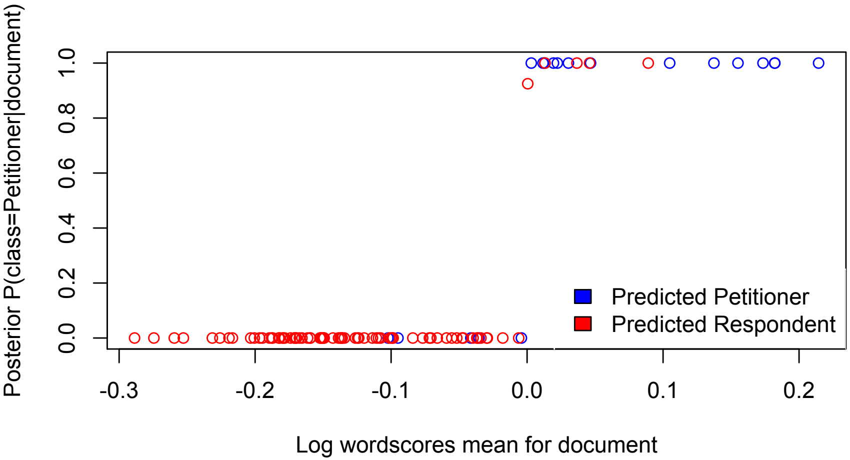
- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Legal document scaling: "Wordscores"

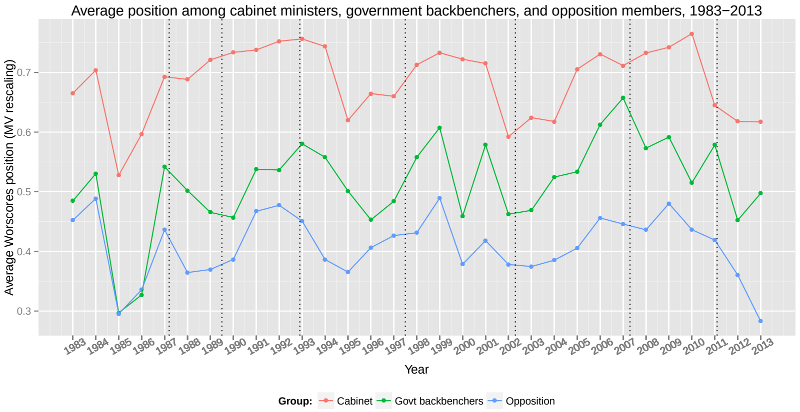


(from Evans et. al. 2007)

Document classification: "Naive Bayes" classifier

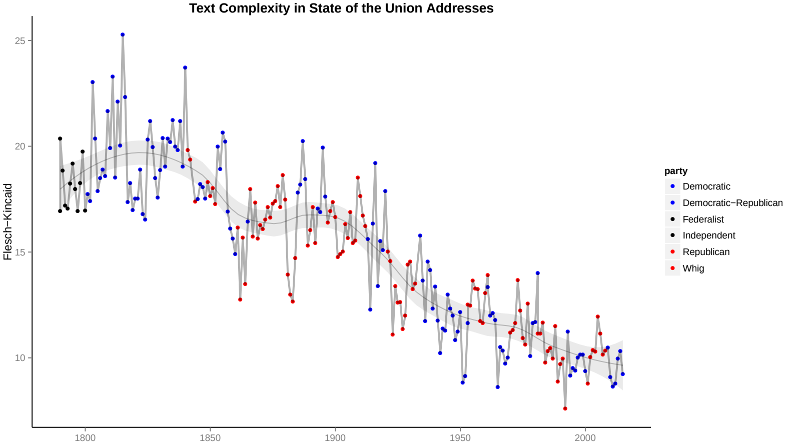


Government v. Opposition in yearly budget debates



(from Herzog and Benoit EPSA 2013)

Reading level of US State-of-the-Union addresses over time



Wordcloud of Tweets from 2014 EP campaign, by list-leading candidate

Outline

- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Basic QTA Process: Texts → Feature matrix → Analysis

```
When I presented the
supplementary budget to
this House last April, I
said we could work our
way through this period
of severe economic
distress. Today, I can
report that
notwithstanding the
difficulties of the past
eight months, we are now
on the road to economic
recovery.
```

```
In this next phase of the
Government's plan we must
stabilise the deficit in
a fair way, safeguard
those worst hit by the
recession, and stimulate
crucial sectors of our
economy to sustain and
create jobs. The worst is
over.
```

```
This Government has the
moral authority and the
well-grounded optimism
rather than the cynicism
of the Opposition. It has
the imagination to create
the new jobs in energy,
agriculture, transport
and construction that
this green budget will
```

docs	words made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Some key basic concepts

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

types
: for our purposes, a unique word

tokens
: any word – so token count is total words
    - hapax legomena (or just *hapax*) are types that occur just once

stems
: words with suffixes removed

lemmas
: canonical word form (the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached)

keys
: such as dictionary entries, where the user defines a set of equivalence classes that group different word types

Some more 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

diversity
: (lexical diversity) A measure of how many types occur per fixed word rate (a normalized vocabulary measure)

Outline

- Motivation for this course
- Logistics
- Issues
- Examples
- Building blocks/foundations of quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations

Word frequencies and their properties

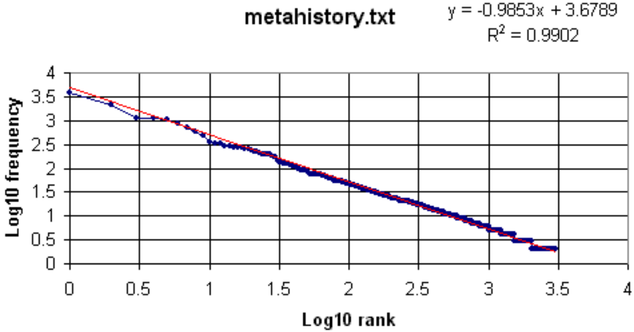
- Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Single tend to be the most informative, as *n*-grams are very rare
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome
- Other approaches use frequencies: Poisson, multinomial, and related distributions

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.



Day 2 Outline

- Selecting texts / defining documents
- Selecting features
- Weighting strategies for features
- Collocations
- Getting texts into `quanteda`
- Detecting collocations
- Exploring texts
- Describing textual data
- Quantifying lexical diversity
- Quantifying the complexity of texts

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

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

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g.
  Rindfleischetikettierungsberwachungsaufgabenbertragungsgesetz
  (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)
  Saunauntensitzer

Defining Features (cont.)

- "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

  莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月
  9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生
  日。生日派对上，莎拉波娃露出了甜美的微笑。

- linguistic features, such as parts of speech
- (if qualitative coding is used) coded or annotated text segments
- linguistic features: parts of speech

Parts of speech

- the Penn “Treebank” is the standard scheme for tagging POS

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Parts of speech (cont.)

- several open-source projects make it possible to tag POS in text, namely Apache's OpenNLP (and R package openNLP wrapper)

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
> sprintf("%s/%s", s[a3w], tags)
 [1] "Pierre/NNP"      "Vinken/NNP"      ",/,"             "61/CD"
 [5] "years/NNS"       "old/JJ"          ",/,"             "will/MD"
 [9] "join/VB"         "the/DT"          "board/NN"        "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"     "Nov./NNP"
[17] "29/CD"           "./."             "Mr./NNP"         "Vinken/NNP"
[21] "is/VBZ"          "chairman/NN"     "of/IN"           "Elsevier/NNP"
[25] "N.V./NNP"        ",/,"             "the/DT"          "Dutch/JJ"
[29] "publishing/NN"   "group/NN"        "./."
```

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
- deliberate disregard Use of "stop words": words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases
- declared equivalency classes Non-exclusive synonyms, what I call a *thesaurus* (lots more on these on Day 4)

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,

Weighting strategies for feature counting

term frequency
: Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency
: Could eliminate words appearing in few documents

inverse document frequency
: Conversely, could weight words more that appear in the most documents

tf-idf
: a combination of term frequency and inverse document frequency, common method for feature weighting

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 $0.016 * 0.916 = 0.0147$
- 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

Other weighting schemes

- the SMART weighting scheme (Salton 1991, Salton et al): The first letter in each triplet specifies the term frequency component of the weighting, the second the document frequency component, and the third the form of normalization used (not shown). Example: *lnn* means log-weighted term frequency, no idf, no normalization

Term frequency		Document frequency	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - \mathrm{df}_t}{\mathrm{df}_t}\}$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$		
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$		

- Note: Mostly used in information retrieval, although some use in machine learning

Stemming words

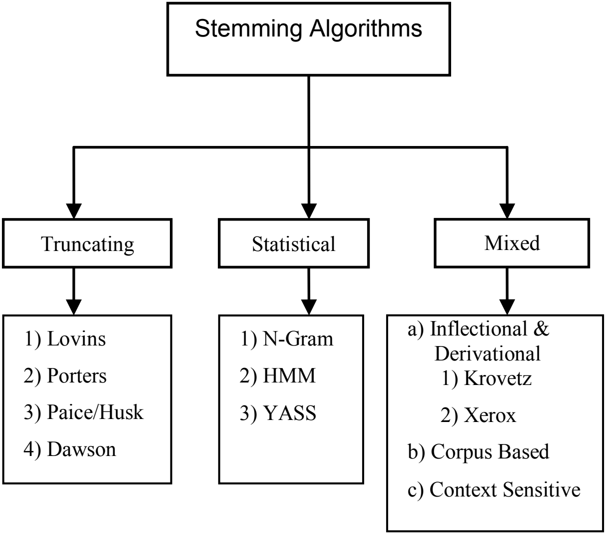
Lemmatization refers to the algorithmic process of converting words to their lemma forms.

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.

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation or "word bigram"
- We can detect these using χ^2 or likelihood ratio measures (Dunning paper)
- Implemented in quanteda as `collocations()`

Getting texts into `quanteda`

- text format issue
  - text files
  - zipped text files
  - spreadsheets/CSV
  - (pdfs)
  - (Twitter feed)
- encoding issue
- metadata and document variable management

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation
- We can detect these using measures of association, such as a likelihood ratio, to detect word pairs that occur with greater than chance frequency, compared to an independence model
- The key is to distinguish "true collocations" from uninteresting word pairs/triplets/etc, such as "of the"
- Implemented in quanteda as `collocations`

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)

Contingency tables for bigrams

Tabulate every token against every other token as pairs, and compute for each token:

	token2	¬token2	Totals
token1	n_{11}	n_{12}	n_{1p}
¬token1	n_{21}	n_{22}	n_{1p}
Totals	n_{p1}	n_{p2}	n_{pp}

Contingency tables for trigrams

		token3	¬token3	Totals
token1	token2	n_{111}	n_{112}	n_{11p}
token1	¬token2	n_{121}	n_{122}	n_{12p}
¬token1	token2	n_{211}	n_{212}	n_{21p}
¬token1	¬token2	n_{221}	n_{222}	n_{22p}
	Totals	n_{pp1}	n_{pp2}	n_{ppp}

computing the “independence” model

- bigrams

$$Pr(\text{token1}, \text{token2}) = Pr(\text{token1})Pr(\text{token2})$$

- trigrams

$$Pr(\text{t1}, \text{t2}, \text{t3}) = Pr(\text{t1})Pr(\text{t2})Pr(\text{t3})$$
$$Pr(\text{t1}, \text{t2}, \text{t3}) = Pr(\text{t1}, \text{t2})Pr(\text{t3})$$
$$Pr(\text{t1}, \text{t2}, \text{t3}) = Pr(\text{t1})Pr(\text{t2})Pr(\text{t3})$$
$$Pr(\text{t1}, \text{t2}, \text{t3}) = Pr(\text{t1}, \text{t3})Pr(\text{t2})$$

more independence models

- for 4-grams, there are 14 independence models
- generally: the number equals the *Bell number* less one, where the Bell number B_n can be computed recursively as:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$$

- but most of these are of limited relevance in collocation mining, as they subsume elements of earlier collocations

statistical association measures

where m_{ij} represents the cell frequency expected according to independence:

G^2 likelihood ratio statistic, computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}}) \tag{1}$$

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

statistical association measures (cont.)

pmi point-wise mutual information score, computed as $\log n_{11}/m_{11}$

dice the Dice coefficient, computed as

$$\frac{n_{11}}{n_{1.} + n_{.1}} \tag{3}$$

Augmenting collocation detection with additional information

- Use parts of speech information

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

- other (machine prediction) tools

Exploring Texts: Key Words in Context

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

Another KWIC Example (Seale et al (2006)

Table 3
Example of Keyword in Context (KWIC) and associated word clusters display

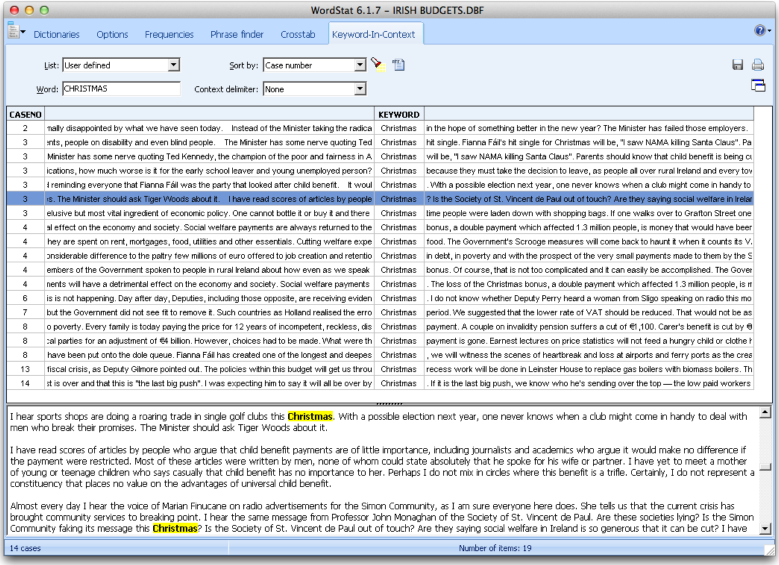
Extracts from Keyword in Context (KWIC) list for the word 'scan'

An MRI **scan** then indicated it had spread slightly
Fortunately, the MRI **scan** didn't show any involvement of the lymph nodes
3 very worrying weeks later, a bone **scan** also showed up clear.
The bone **scan** is to check whether or not the cancer has spread to the bones.
The bone **scan** is done using a type of X-ray machine.
The results were terrific, CT **scan** and pelvic X-ray looked good
Your next step appears to be to await the result of the **scan** and I wish you well there.
I should go and have an MRI **scan** and a bone **scan**

Three-word clusters most frequently associated with keyword 'scan'

N	Cluster	Freq
1	A bone scan	28
2	Bone scan and	25
3	An MRI scan	18
4	My bone scan	15
5	The MRI scan	15
6	The bone scan	14
7	MRI scan and	12
8	And Mri scan	9
9	Scan and MRI	9

Another KWIC Example: Irish Budget Speeches



Irish Budget Speeches KIWC in quanteda

R Console

```
> data(iebudgets)
> iebudgets2010 <- subset(iebudgets, year==2010)
> kwic(iebudgets2010, "christmas", regex=TRUE)
                                                             preword       word                                     postword
     [2010_BUDGET_02_Richard_Bruton_FG.txt, 628]          and to see out this  Christmas in the hope of something
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 371]         to suggest titles for a Christmas  hit single. Fianna Fáil’s hit
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 379]       Fianna Fáil’s hit single for Christmas  will be, “I saw NAMA
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 922]       women will say goodbye after Christmas  because they must take the
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 1518]         in single golf clubs this Christmas. With a possible election next
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 1726] Community faking its message this Christmas? Is the Society of St.
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 3159]         bags. In previous years at Christmas  time people were laden down
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 346]              €204 per week or the Christmas  bonus. Of course, that is
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3239]   to social welfare payments this Christmas. The loss of the Christmas
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3244]         Christmas. The loss of the Christmas  bonus, a double payment which
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3272]     streets on Santa presents and Christmas  food. The Government’s Scrooge measures
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 5899]         their jobs, who face this Christmas  in debt, in poverty and
        [2010_BUDGET_06_Enda_Kenny_FG.txt, 2629] to implement the reduction before Christmas. I do not know whether
   [2010_BUDGET_07_Kieran_ODonnell_FG.txt, 1365]            from the change in the  Christmas period. We suggested that the
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 550]         cut of €641, including the Christmas  payment. A couple on invalidity
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 638]         are on social welfare, the Christmas  payment is gone. Earnest lectures
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 998]    of emigration. Once again this Christmas, we will witness the scenes
    [2010_BUDGET_13_Ciaran_Cuffe_Green.txt, 911]       noted recently that over the  Christmas recess work will be done
[2010_BUDGET_14_Caoimhghin_OCaolain_SF.txt, 148]             will all be over by Christmas. If it is the last
>
```