Quantitative text analysis: overview and fundamentals

Blake Miller

MY 459: Quantitative Text Analysis

January 20, 2020

Course website: lse-my459.github.io

Text as data

Text as data

Source: The Comparative Manifesto Project, https://manifesto-project.wzb.eu

Text as data

AGAPETI PAPÆ I
EPISTOLÆ.

EPISTOLA JUSTINIANI
AD AGAPETUM.

More majorum suorum apud pontificem Romanum recens electum fidei suæ professionem edit, eamdem quam supra ad Joannem papam II miserat.

In nomine Domini nostri Jesu Christi Dei imperator Cæsar Flavius Justinianus, Alemanicus, Gothicus, Francicus, Germanicus, Anticus, Alanicus, Vandalicus, Africanus, Pius, Felix, Inclytus, Victor, ac Triumphator semper Augustus, Agapeto sanctissimo archiepiscopo almæ urbis Romæ et patriarchæ.

Ante tempus in hac regia urbe nostra quorumdam de causa fidei exstitit morbosa contentio; quam nos congrue respuentes interposito edicto repressimus. Et quia studii nostri est emergentes hujus-

Reddentes honorem apostolicæ sedi et vestræ sanctitati, quod semper nobis in voto fuit, et est, ut decet patrem, honorantes vestram beatitudinem, C omnia, quæ ad Ecclesiarum statum pertinent, festinamus ad notitiam deferre vestræ sanctitatis: quoniam semper magnum nobis fuit studium unitatem vestræ apostolicæ sedis, et statum sanctarum Dei Ecclesiarum custodire, quæ hactenus obtinet, et incommote permanet, nulla intercedente contrarietate. Petimus ergo vestrum paternum affectum, ut vestris ad nos destinatis litteris, et ad sanctissimum episcopum hujus almæ urbis et patriarcham vestrum fratrem, quoniam et ipse per eosdem scripsit ad vestram sanctitatem, festinans in omnibus consequi sedem apostolicam beatitudinis vestræ, manifestum nobis faciatis, quod omnes qui prædictam fidem recte

Text as data

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
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

Targets

- Learning objectives
  - fundamentals of text analysis
  - availability and consequences of *choices*
  - practical ability to work with texts in `R`
  - issues of text for social science
- Prerequisites
  - linear algebra and quantitative methods (MY452 or equivalent regression analysis course)
  - familiarity with R and RStudio
  - (optional) ability to process text files in a programming language such as Python

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

About me

- Assistant Professor in Computational Social Science at the Methodology Department, LSE
- Previously at **Dartmouth College**
- PhD in Political Science and Scientific Computing, **University of Michigan**
- **My research:**
  - Chinese politics, and authoritarian politics
  - Information control (censorship, propaganda, etc.), information and mobilization of violence
  - Supervised machine learning for texts
- **Contact:**
  - `b.a.miller@lse.ac.uk`
  - `www.blakeapm.com`
  - Office hours: Mondays 10:00-12:00 (COL 7.14) during the term (book through Student Hub)

Your turn!

1. Name?
2. MSc/PhD Programme?
3. Previous experience with R / GitHub / text analysis?

Course resources

- Course website: lse-my459.github.io
  - Class description
  - Course schedule
  - Slides from class
  - Readings list
  - Links to exercises and datasets
  - Submission links for homeworks
- Moodle page
  - Supporting materials
- Readings
  - Mainly articles
  - Complement content covered in lectures and seminars

Course schedule

- **Lectures**: Mondays 12:00-14:00 in CBG.2.06
- **Classes** only for weeks 2, 4, 7, 9, 11:
  1. Tuesdays 9:00-11:00 FAW.4.01
  2. Tuesdays 12:00-14:00 FAW.4.01
- No lectures or classes during Reading Week (week 6)

Week	Topic	Week	Topic
1	Overview and Fundamentals	7	Supervised Scaling Models for Texts
2	Descriptive statistical methods for text analysis	8	Unsupervised Models for Scaling Texts
3	Quantitative methods for comparing texts	9	Similarity and clustering methods
4	Automated dictionary methods	10	Topic models
5	Machine Learning for Texts	11	Working with Social Media
6	*Reading Week*		

Evaluation

- **Formative coursework**:
  - Five problem sets, building upon content of lab sessions
  - 60% of course grade
  - Submitted via GitHub classroom (please create an account before first lab session)
- **Project**:
  - Original analysis of texts using methods covered in class
  - It can replicate or extend a published work
  - 3,000 words (5,000 for MY559), due at the beginning of ST (May 4th, 5pm)
  - 40% of course grade

Assessment criteria

- 70–100: Very Good to Excellent (Distinction).
  - Perceptive, focused use of a good depth of material with a critical edge. Original ideas or structure of argument.
- 60–69: Good (Merit)
  - Perceptive understanding of the issues plus a coherent well-read and stylish treatment though lacking originality
- 50–59: Satisfactory (Pass)
  - A "correct" answer based largely on lecture material. Little detail or originality but presented in adequate framework. Small factual errors allowed.
- 30–49: Unsatisfactory (Fail)
- 0–29: Unsatisfactory (Bad fail)
  - Based entirely on lecture material but unstructured and with increasing error component. Concepts are disordered or flawed. Poor presentation. Errors of concept and scope or poor in knowledge, structure and expression.

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

Why quantitative text analysis?

Justin Grimmer's haystack metaphor: QTA improves reading

- Analyzing a straw of hay: understanding the meaning of a sentence
  - Humans are great! But computer struggle
- Organizing the haystack: describing, classifying, scaling texts
  - Humans struggle. But computers are great!
  - (What this course is about)

Principles of quantitative text analysis (Grimmer & Stewart, 2013)

1. All quantitative models are wrong – but some are useful
2. Quantitative methods for text amplify resources and augment humans
3. There is no globally best method for automated text analysis
4. Validate, validate, validate

Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
   - An attribute of the author
   - A sentiment or emotion
   - Salience of a political issue
2. Texts can be represented through extracting their *features*
   - most common is the bag of words assumption
   - many other possible definitions of "features" (e.g. word embeddings)
3. A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

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

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 documentary 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

Overview of text as data methods

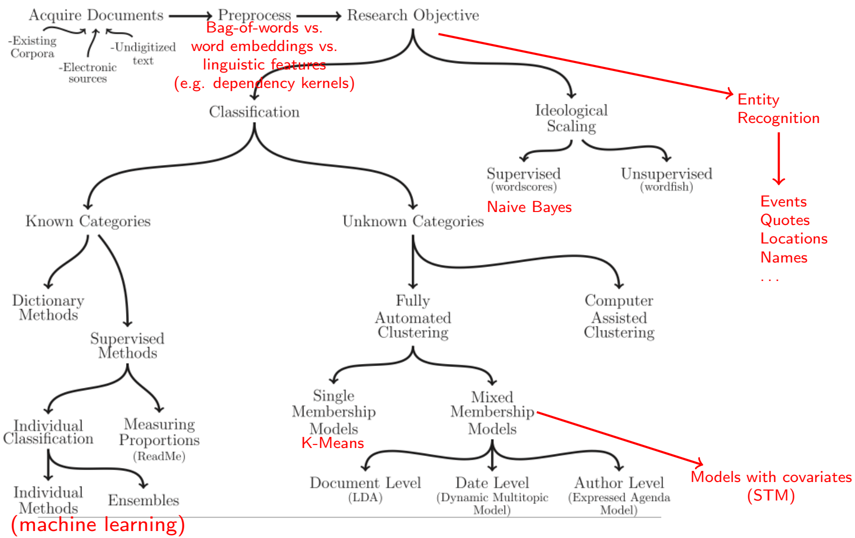
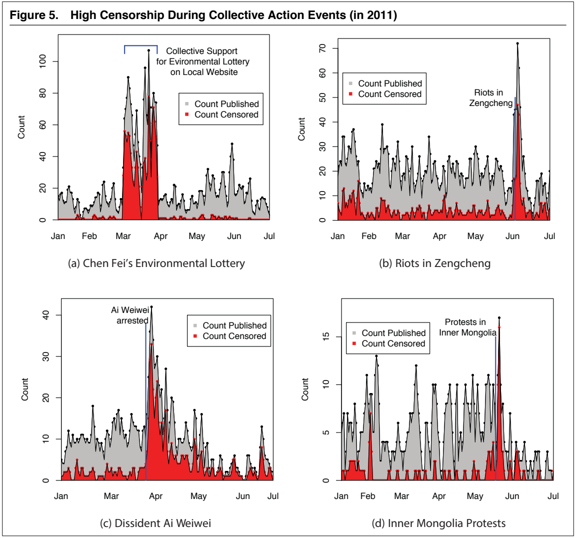


Fig. 1 in Grimmer and Stewart (2013)

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

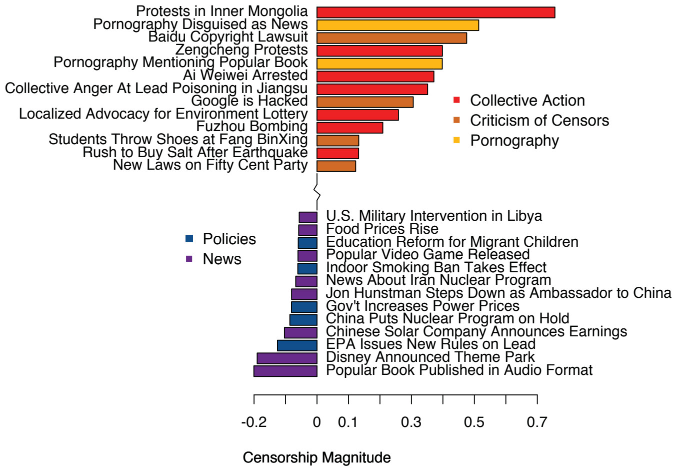
Descriptive text analysis



King, Pan, & Roberts (2013)

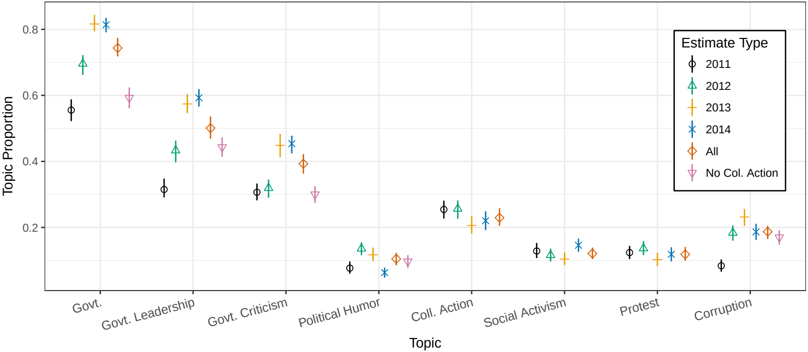
Descriptive text analysis

Figure 4. Events with Highest and Lowest Censorship Magnitude



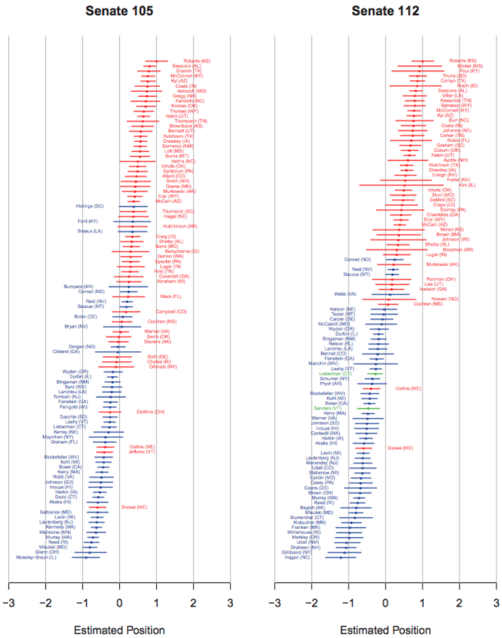
King, Pan, & Roberts (2013)

Document classification into known categories



Miller, *working paper*, 2020.

Ideological scaling (Lauderdale & Herzog, PA 2016)

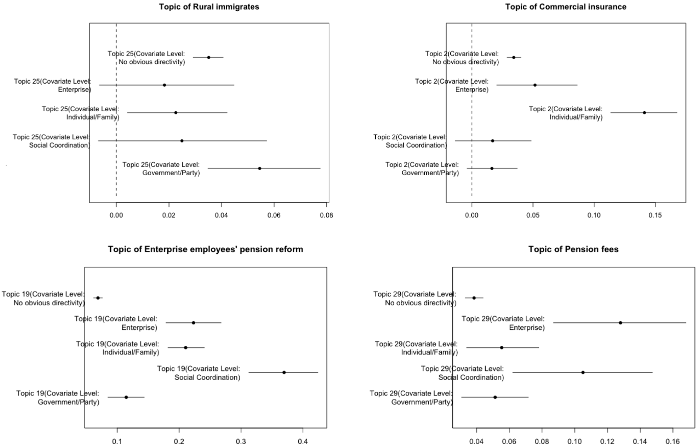


Document classification into unknown categories

Wang, Yan, *Governmentality and Counter-conduct in Authoritarian Regimes, Dissertation*, 2020.

- Data: State media reports about pension reform in China.
- Hand coded covariates: *Who should be responsible for this reform/policy?*
- Automated text analysis to discover unknown categories; hand-coded covariates to analyze state discourse on reforms.
- Paired with qualitative discourse analysis of how the state explained its reforms to the public using propaganda.

Document classification into unknown categories



Wang, Yan, *Governmentality and Counter-conduct in Authoritarian Regimes, Dissertation*, 2020.

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

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

document each of the units of the corpus

types for our purposes, a unique word

tokens any word – so token count is total words

e.g.
```
A corpus is a set of documents.
This is the second document in the corpus.
```

is a corpus with 2 documents, where each document is a sentence. The first document has 6 types and 7 tokens. The second has 7 types and 8 tokens. (We ignore punctuation for now.)

Some more key basic concepts

stems: words with suffixes removed (using set of rules)

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

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

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

"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

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

Basic QTA adopts a bag-of-words approach

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

Bag-of-words approach

From words to numbers:

1. Preprocess text: lowercase, remove stopwords and punctuation, stem, tokenize into unigrams and bigrams (bag-of-words assumption)

   "A corpus is a set of documents."
   "This is the second document in the corpus." "a corpus is a set of documents."
   "this is the second document in the corpus." "~~a~~ corpus ~~is~~ ~~a~~ set ~~of~~ documents~~.~~"
   "~~this~~ ~~is~~ ~~the~~ second document ~~in~~ ~~the~~ corpus~~.~~" "corpus set document~~s~~"

   "second document corpus" [corpus, set, document, corpus set, set document]

   [second, document, corpus, second document, document corpus]
2. Document-feature matrix:
   - **W**: matrix of N documents by M unique n-grams
   - w_{im} = number of times m-th n-gram appears in i-th document.

corpus set document corpus set … M n-grams

Bag-of-words approach

QTA often disregards grammar and word order and uses word frequencies as features.

Why? What are the main advantages and limitations of this assumption?

Word frequencies and their properties

Bag-of-words approach disregards grammar and word order and uses word frequencies as features. Why?

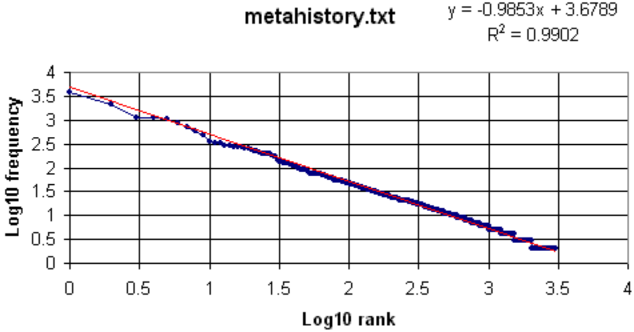
- *Context is often uninformative*, conditional on presence of words:
  - Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Single words tend to be the most informative, as co-occurrences of multiple words (*n*-grams) are 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

- Basic idea: word frequency follows a power distribution; "of" and "the" make up 10% of all occurrences and "aardvark" hardly ever occurs.
- 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.
- Fun fact: this law also holds for measures such as the population of global cities.

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.



Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

Strategies for selecting units of textual analysis

What can the document be?

- Words
- *n*-word sequences
- Sentences
- Pages
- Paragraphs
- Natural units (a speech, a poem, a manifesto)
- Aggregation of units (e.g. all speeches by party and year)
- Key: depends on the research design
- Frequent trade-off between cost and accuracy

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

Outline

- Motivation for this course
- Logistics
- Foundations
- Examples
- Key terms in quantitative text analysis
- Justifying a term/feature frequency approach
- Selecting texts / defining documents
- Selecting features

Defining Features

- characters
- 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.
  Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
  (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１８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
- word embeddings (more on this later in the course)

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

Parts of speech (cont.)

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

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director
Nov. 29. 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"         "./."
```

Parts of speech (cont.)

Example: Creating an index of editorialization of journalists' and media outlets' political news coverage.

Proportion of tweets that: (1) mention a major party or candidate, (2) include at least one adjective.

Table 2.4 Determinants of editorialisation and popularity of news accounts on twitter (OLS regressions)

	DV = Editorialisation		DV = Popularity	
	Model 1	Model 2	Model 3	Model 4
Type: journalist	5.10^{***}	4.32^{***}	2.70^{***}	2.49^{***}
	(1.13)	(1.26)	(0.22)	(0.30)
Tweets about Europe (%)	-0.03^{+}	-0.03^{+}	0.01^{***}	0.01^{***}
	(0.02)	(0.02)	(0.002)	(0.002)
Editorialisation Index			0.02^{***}	0.02^{***}
			(0.004)	(0.004)
(Intercept)	7.58^{**}	7.94^{**}	-4.03^{***}	-3.92^{***}
	(2.59)	(2.47)	(0.40)	(0.41)
Country fixed effects	YES	YES	YES	YES
Outlet fixed effects	YES	YES	YES	YES
R^2	0.12	0.12	0.71	0.71
Adj. R^2	0.08	0.08	0.70	0.70
Num. obs.	2662	2662	2662	2662
RMSE	7.63	7.63	1.08	1.08

Barberá, Vaccari, Valeriani (2016) [control variables ommitted]

Strategies for feature selection

How to choose which features to include?

- All? Computationally inefficient, and rare words are generally uninformative

Potential criteria to select features ("trim" the DFM):

- 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, also known as *thesaurus* (more on this later)

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

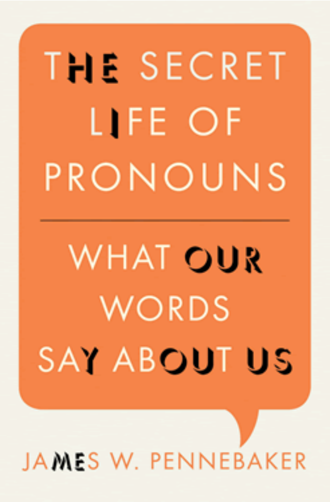
A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, 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, aren't, 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, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, 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, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let's, 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, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero

Stopwords

Are there cases in which we would want to keep stopwords? Or should we always exclude them from our analysis?

Stopwords sometimes can be informative!



But sometimes we want to add/remove our own new stopwords
(e.g. female pronouns, legislative terms, directional terms)

Stemming words

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

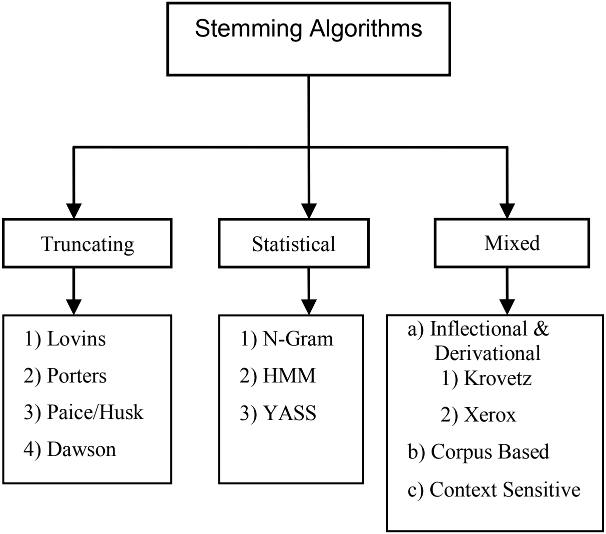
stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

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

Why? Reduce feature space by collapsing different words into a stem (e.g. "happier" and "happily" convey same meaning as "happy")

Varieties of stemming algorithms



Issues with stemming approaches

- The most common is probably 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 approaches designed to overcome these limitations
- Key for you is to be careful through inspection of morphological variants and their stemmed versions
- Sometimes not appropriate! e.g. Schofield and Minmo (2016) find that "stemmers produce no meaningful improvement in likelihood and coherence (of topic models) and in fact can degrade topic stability"

Where to obtain textual data?

Some tips...

- Existing datasets, e.g.
  - UCD's EuroParl project
  - Hansard Archive of parliamentary debates in UK
  - Media archives (newspaper articles, TV transcripts...) at LexisNexis, ProQuest, Factiva...
  - Academic articles (JSTOR Data for Research)
  - Open-ended responses to survey questions
- Collect your own data:
  - From social media (Twitter, FB) and blogs
  - Scraping other websites
- Digitize your own text data using OCR (optical character recognition) software
  - Options: Tesseract (open-source), Abbyy FineReader

Where to obtain textual data?

What type of textual data have you worked with?
What data would you be interested in collecting?

Wrapping up...

Big questions we answered today:

- Quantitative Text Analysis: why?
- Key terms: document, corpus, feature, document feature matrix, type, token
- How to select the unit of analysis (i.e. documents)?
- How to select features? Bag-of-words, stemming, stopwords, part-of-speech tagging

Before next class

- Do readings for today and next class
- Create a GitHub account