Class 9: Quantitative Text Analysis: Data Preparation

MAST5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

8 February 2023

Outline of Today's Class

Text Analysis Methods

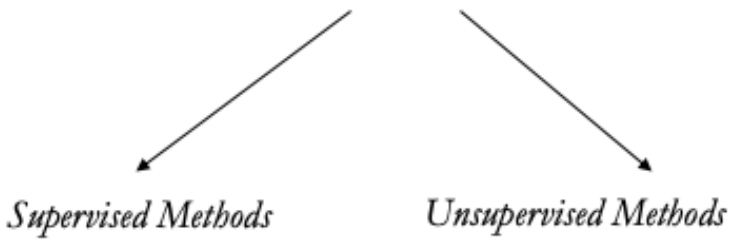
Text Analysis Methods

What for?

- *General*: Measuring text or author's/context characteristics
- *Quantitative*: Organising and describing ('structuring') large amounts of text data

Text Analysis Methods

Fully Qualitative	*Content Analysis*	*Fully Quantitative*
Few texts	Medium-sized sets of texts	Large sets of texts
Pragmatics/Connotative - Words + co-text + context	Mix - Systematic codebook but built & executed by humans	Literal/Denotative - Bag of words/word frequencies
Transcripts/Themes	Numerical Summaries / Stats	Numerical Summaries / Stats



Content Analysis

Example: Comparative Manifesto Project

- 1000 party manifestoes from 1945 until today
- 50 countries
- 7 macro-classes:
  1. External Relations
  2. Freedom & Democracy
  3. Political System
  4. Economy
  5. Welfare and Quality of Life
  6. Fabric of Society
  7. Social Groups
- Between 5-10 sub-categories each
- Manifestoes split in 'quasi-sentences': critiqued for unreliability
- `https://manifesto-project.wzb.eu/down/papers/handbook_2014_version_5.pdf`

Content Analysis

Example: Comparative Manifesto Project

DOMAIN 4: Economy

401 Free Market Economy

Favourable mentions of the free market and free market capitalism as an economic model. May include favourable references to:

- Laissez-faire economy;
- Superiority of individual enterprise over state and control systems;
- Private property rights;
- Personal enterprise and initiative;
- Need for unhampered individual enterprises.

402 Incentives: Positive

Favourable mentions of supply side oriented economic policies (assistance to businesses rather than consumers). May include:

- Financial and other incentives such as subsidies, tax breaks etc.;
- Wage and tax policies to induce enterprise;
- Encouragement to start enterprises.

403 Market Regulation

Support for policies designed to create a fair and open economic market. May include:

- Calls for increased consumer protection;
- Increasing economic competition by preventing monopolies and other actions disrupting the functioning of the market;
- Defence of small businesses against disruptive powers of big businesses;
- Social market economy.

404 Economic Planning

Favourable mentions of long-standing economic planning by the government. May be:

- Policy plans, strategies, policy patterns etc.;
- Of a consultative or indicative nature.

405 Corporatism/ Mixed Economy

Favourable mentions of cooperation of government, employers, and trade unions simultaneously. The collaboration of employers and employee organisations in overall economic planning supervised

Content Analysis

Example: Comparative Manifesto Project

The extract shown above from the electoral program of the Democratic Party looks like following after the coding:

quasi_sentence	category	description
President Obama has already signed into law $2 trillion in spending reductions as part of a balanced plan to reduce our deficits by over $4 trillion over the next decade	414	Economic Orthodoxy
while taking immediate steps to strengthen the economy now.	408	Economic Goals
This approach includes tough spending cuts that will bring annual domestic spending to its lowest level as a share of the economy in 50 years,	414	Economic Orthodoxy
while still allowing us to make investments that benefit the middle class now	704	Middle Class and Professional Groups
and reduce our deficit over a decade.	414	Economic Orthodoxy

Quantitative Text Analysis Methods

By research objective

Classification		**Scaling**	
Supervised	*Unsupervised (unknown categories)*	*Supervised*	*Unsupervised (unknown scale)*
Naive Bayes	Topic Models	Dictionaries Wordscores	Wordfish

Quantitative Text Analysis Methods

Example 1: Supervised Classification [1]

Table 5: Sample of Words Most Discriminative of the Conservative and Liberal Positions on Affirmative Action Among Amicus Curiae and Principal Litigants in the *Bollinger* Cases

Term[a]	*Avg. Freq. per Lib. Brief*	*Avg. Freq per Cons. Brief*	*Chi²*	*Interpretive Code Examples*[b]
Conservative Words				
PREFER*	2.83	41.79	39.18	Proceduralist; Race/Gender Neutral Justice
BENIGN	0.07	1.17	36.14	Intent vs. Consequences; Constraint
DISCRIM*	14.86	25.04	24.13	Proceduralist; Race/Gender Neutral Justice
PURPORT*	0.44	1.88	24.13	Skepticism
CLASSIF*	2.1	11.54	22.39	Proceduralist; Race/Gender Neutral Justice
NARROW-TAILORING	0.05	0.96	19.73	Proceduralist; Strict Scrutiny
REJECT*	2.75	7.79	19.15	Oppositional Posture
JUSTIF*	2.39	12.79	18.91	Proceduralist; Constraint
FORBID*	0.38	1.63	18.91	Proceduralist; Constraint; Race/Gender Neutral Justice
PROHIBITS	0.13	0.71	18.08	Proceduralist; Constraint
RATIONALE	0.66	5.92	17.58	Proceduralist; Legalistic
AMORPHOUS	0.25	1.29	14.62	Proceduralist; Skepticism
RACE-BASED	1.08	10.46	10.59	Proceduralist; Pejorative counterpart to liberal RACE-CONSCIOUS
Liberal Words				
LEADERS	2.70	0.13	31.03	Impact; Development
WORLD	3.00	0.42	18.74	Impact; Global
NATION*	21.0	7.04	17.90	Impact; Communitarian
IMPACT*	4.13	1.04	17.49	Impact
EFFECTIVE	2.78	0.75	16.54	Impact; Effectiveness
SOCIAL	6.84	1.71	16.05	Impact; Communitarian
COMMUNIT*	8.75	1.75	15.35	Impact; Communitarian
BUSINESS*	4.56	0.58	10.28	Impact; Efficiency; Distributive Justice
DESEGREGATION	2.34	0.17	10.24	Remedial Justice
GROW*	2.38	0.33	10.24	Change; Development
WORKFORCE	1.64	0.00	9.81	Impact; Distributive Justice; Development
RACE-CONSCIOUS	7.14	1.50	7.80	Proceduralist; Euphemistic counterpart to conservative RACE-BASED

[a]For the sake of parsimony, asterisks are used to denote lemmatized terms where morphologically-related variants are all highly discriminative. For example, "preference," "preferences," and "preferred" all had high chi² values and so we lemmatized them with "prefer.*"

[1] Evans, M., McIntosh, W., Lin, J., Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4), 1007-1039.

Quantitative Text Analysis Methods

Example 2: Unsupervised Classification [2]

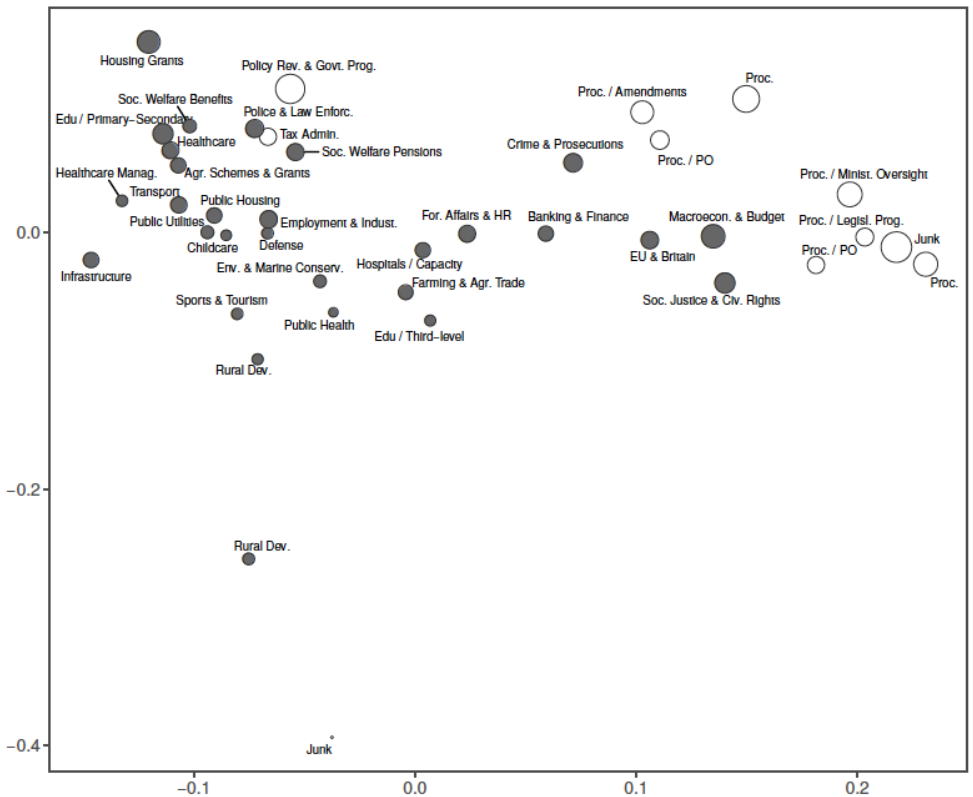


Fig. 1: *Topic similarities for the 2015 time slice.* This figure illustrates topic similarity by displaying Jensen-Shannon distances which are projected onto a 2D space with the use of classical multidimensional scaling (MDS). Topics are represented by circles, whose size corresponds to the prevalence of the topic throughout the corpus. Topics that use similar words are closer together and vice-versa. Solid circles represent substantive topics. Hollow circles represent topics that are either "junk" or procedural in nature.

[2] Boussalis, C., McElroy,G., Sorace, M. (2022) The Conditional Nature of the Descriptive-Substantive Link in the Representation of Women *Working Paper*

Quantitative Text Analysis Methods

Example 4: Unsupervised Scaling [3]

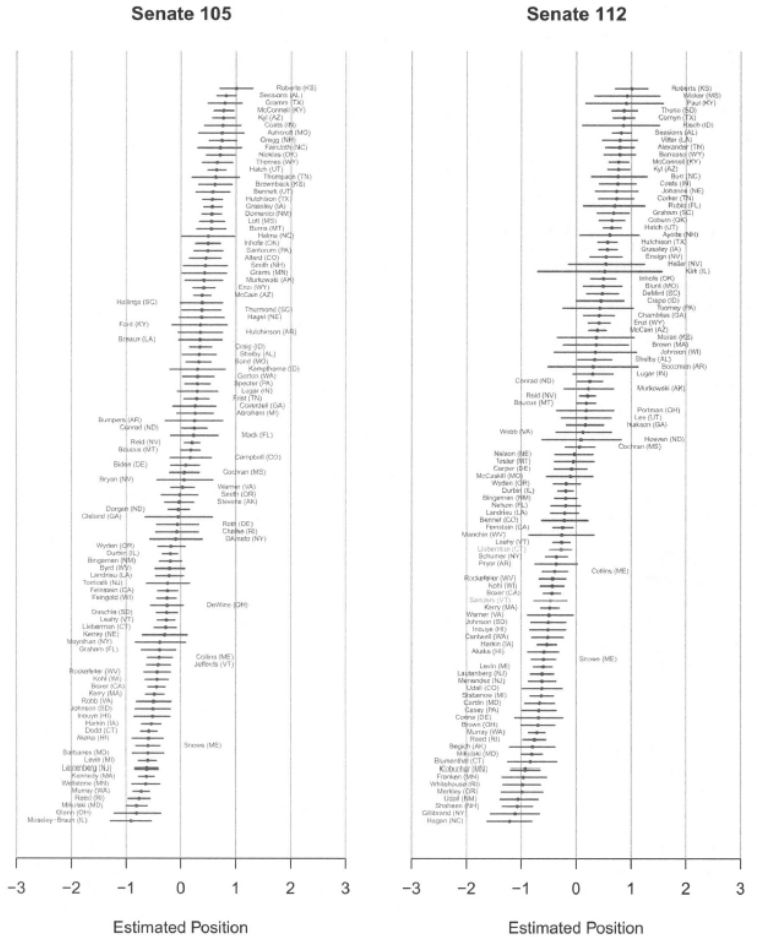


Fig. 4 Wordshoal estimates for the 105th and 112th U.S. Senates. Republican senators names are to the right of the estimates, Democrats and Independents are to the left.

[3] Lauderdale, B. E., Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, 374-394.

Key Terminology & Steps of Quantitative Text Analysis

Core Terminology

- **Corpus**: set of documents, structured collection of texts
- **Document**: textual unit of analysis (word, sentence, paragraph, speech, collection of speeches ...)
- **Type**: unique term/word
- **Token**: specific occurrence of a type
- **Stem**: word minus its suffixes
  - e.g. winning = win, wins = win, BUT won = won, winner = winner
- **Lemma**: root of the word, canonical word form
  - e.g. winning, wins, winner, won: all become WIN
- **Keyword**: word of special interest, due to meaning or rate of occurrence

Core Terminology

- **Stop words**: words identified as unnecessary/not meaningful for analysis

```
 [1] "i"         "me"        "my"        "myself"    "we"        "our"       "ours"      "ourselves"
 [9] "you"       "your"      "yours"     "yourself"  "yourselves" "he"       "him"       "his"
[17] "himself"   "she"       "her"       "hers"      "herself"   "it"        "its"       "itself"
[25] "they"      "them"      "their"     "theirs"    "themselves" "what"     "which"     "who"
[33] "whom"      "this"      "that"      "these"     "those"     "am"        "is"        "are"
[41] "was"       "were"      "be"        "been"      "being"     "have"      "has"       "had"
[49] "having"    "do"        "does"      "did"       "doing"     "would"     "should"    "could"
[57] "ought"     "i'm"       "you're"    "he's"      "she's"     "it's"      "we're"     "they're"
[65] "i've"      "you've"    "we've"     "they've"   "i'd"       "you'd"     "he'd"      "she'd"
[73] "we'd"      "they'd"    "i'll"      "you'll"    "he'll"     "she'll"    "we'll"     "they'll"
[81] "isn't"     "aren't"    "wasn't"    "weren't"   "hasn't"    "haven't"   "hadn't"    "doesn't"
[89] "don't"     "didn't"    "won't"     "wouldn't"  "shan't"    "shouldn't" "can't"     "cannot"
[97] "couldn't"  "mustn't"   "let's"     "that's"    "who's"     "what's"    "here's"    "there's"
[105] "when's"   "where's"   "why's"     "how's"     "a"         "an"        "the"       "and"
[113] "but"      "if"        "or"        "because"   "as"        "until"     "while"     "of"
[121] "at"       "by"        "for"       "with"      "about"     "against"   "between"   "into"
[129] "through"  "during"    "before"    "after"     "above"     "below"     "to"        "from"
[137] "up"       "down"      "in"        "out"       "on"        "off"       "over"      "under"
[145] "again"    "further"   "then"      "once"      "here"      "there"     "when"      "where"
[153] "why"      "how"       "all"       "any"       "both"      "each"      "few"       "more"
[161] "most"     "other"     "some"      "such"      "no"        "nor"       "not"       "only"
[169] "own"      "same"      "so"        "than"      "too"       "very"      "will"
```

Stopwords: Why Not?

Steps of Quantitative Text Analysis

1. Select & acquire texts for analysis (previous classes)
2. **Text pre-processing**
   - Create Document-Feature Matrix (DFM)
   - Trim/Clean the DFM
3. Measure underlying characteristics of interest with statistical methods (next 2 classes)

The Document-Feature Matrix (DFM)

Matrix of documents (rows) by type (column)
Word frequencies (or relative frequencies) in the cells
Simplest form of text representation in numbers

	@joebiden	made	better	ask	believ	kamala	abil	help	make	us	countri	import	elect
Obama_tweets.csv.1	1	1	2	2	2	1	1	2	1	1	1	1	1
Obama_tweets.csv.2	1	0	1	0	0	0	0	0	0	0	0	0	1
Obama_tweets.csv.3	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.4	0	0	1	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.5	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.6	1	0	0	0	0	0	0	0	1	0	0	0	1
Obama_tweets.csv.7	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.8	0	0	0	0	0	0	0	0	1	0	0	0	0
Obama_tweets.csv.9	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.10	1	0	0	0	0	0	0	0	0	0	0	0	2
Obama_tweets.csv.11	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.12	1	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.13	0	0	0	0	0	0	0	0	0	0	1	0	0
Obama_tweets.csv.14	0	0	0	0	0	0	0	0	0	2	0	0	0
Obama_tweets.csv.15	0	0	0	0	0	0	0	0	1	0	0	1	0
Obama_tweets.csv.16	0	0	0	0	0	0	0	0	0	0	0	0	0
Obama_tweets.csv.17	0	0	0	0	0	0	0	0	3	0	0	0	1

The 'Bag of Words' Approach

- The DFM disregards syntax and grammar - hence it is like a 'bag of words'
- The analysis building blocks in this model are word (type) frequencies
- Unrealistic - but every model is an approximation: central issue is whether it is *useful* for our specific research

The 'Bag of Words' Approach

Justification

- In most cases individual words hold most of the information (e.g. ideology, affect/sentiment ...), co-text and context are often unnecessary/redundant
- Natural languages follow **Zipf's Law**: words mentioned more frequently reflect most important concerns/meaning to be conveyed

Zipf's Law and Human Language [4]

- Words occur in natural languages following a mathematical law, whereby word frequency is inversely proportional to its rank.
  - E.g. 2nd most common word occurs 1/2 as often as the first; 3rd most common occurs 1/3 as often & so on ...
- This is a universal pattern in human languages.
- The reason for this statistical regularity in language use has to do with memory/cognition limits and the need for communicative optimisation: word re-use is fundamental for successful communication among humans.

[4] Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5), 1112-1130.

Pre-Processing Text Data in R

Text Pre-Processing

Steps

In order to make text analysis computations more efficient, it is essential to generate and to 'trim' a DFM:

1. Convert text data into a corpus
2. Summarise and explore your corpus
3. Convert the corpus object into a dfm
4. Trim (or 'clean') the dfm
   - stop words removal
   - lower case
   - eliminate punctuation and special symbols (unless useful for analysis)
   - Optional pre-processing steps (consider if appropriate):
     - stemming
     - weighting: tf-idf / relative frequency

Stemming

- Reduce inflected words to their word stem:
  - E.g. worker, working, works = work
- Stemmer = algorithm that accomplishes word stemming
  - removing common suffixes/affixes; using lookup table of word roots

Trimming DFMs: The tf-idf weighting strategy

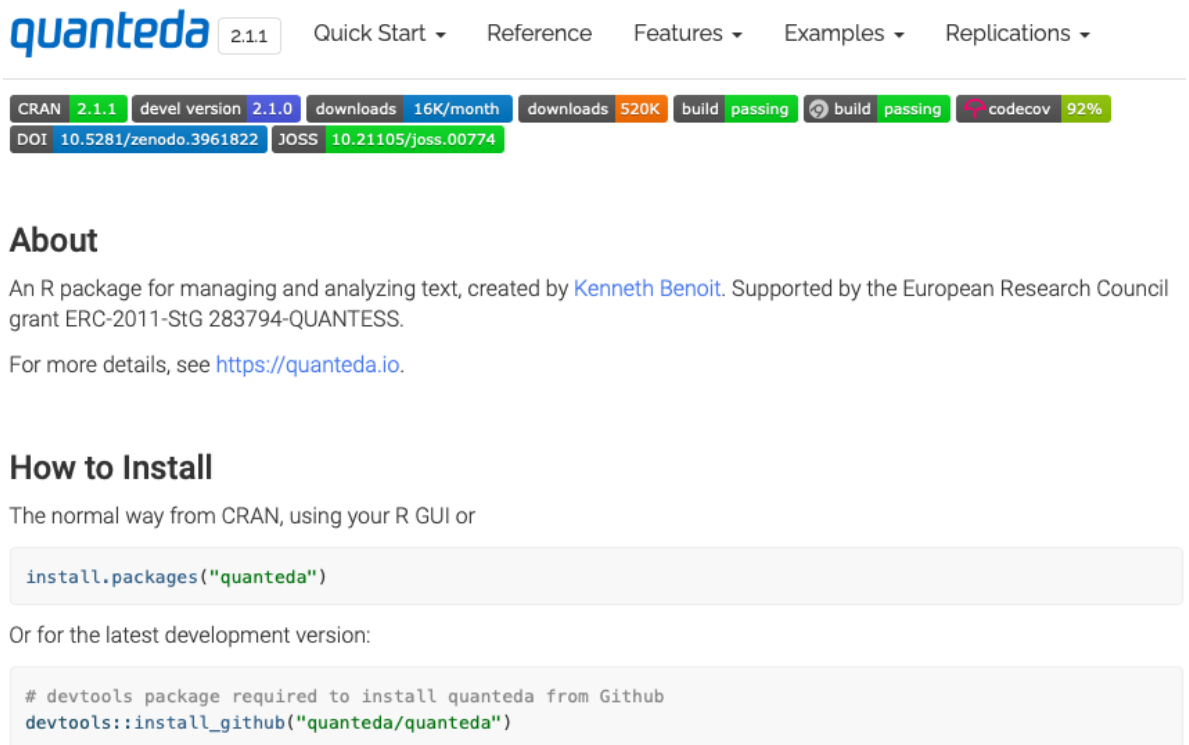
- To further reduce the feature matrix, it is advised to eliminate very low frequency words as well as highly common words (unlikely to be discriminant)
- The tf-idf weight accomplishes just that:
  - It computes the frequency of a term in a document as well as its frequency across the total number of documents.
  - It takes the inverse document frequency of the term and multiplies it to its frequency in the document.

tf-idf: example

- Document A is 1000 words long and it contains 20 instances of the term "transport". 40 documents in the corpus (total: 100 documents) contain the word "transport".
  - The term frequency (tf) is $20/1000 = 0.02$
  - The *inverse* document frequency is $\ln(100/40) = 0.9$
  - The tf-idf is 0.02*0.9 = 0.018
  - If the term only appeared in 15 documents, the tf-idf would be 0.02*1.89= 0.038, nearly 2.5 times higher
- The weight is higher when the term is more discriminating: e.g. when it appears in fewer documents and/or has high frequency in a particular document.
- tf-idf therefore 'filters out' common terms

Quantitative Text Analysis in R: quanteda

- Explore `https://quanteda.io`, lots of useful resources and examples!
- The reference section offers detailed information on each text analysis function.



R code demonstration

- R Code
- Use your scraped tweets!