University of Sheffield

(Computational) Text Analysis 1

Session 1 – Introduction & Dictionary Methods

22 November 2023

Dr. Dani Madrid-Morales

Outline

- Logistics
- Why Computational Text Analysis
- Bag of Words Approach (BoW)
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Outline

- **Logistics**
- Why Computational Text Analysis
- Bag of Words Approach (BoW)
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Introductions

- Dani Madrid-Morales
- I did my PhD in **Hong Kong**, worked at the **University of Houston**, and now at **University of Sheffield**
- I study political communication, disinformation & public diplomacy

Today's Learning Objectives

1. Be aware of the range of approaches available to researchers wanting to use **computational text analysis**;
2. Understand the principles underpinning the **bag of words (BoW)** approach;
3. Use the **quanteda package in R** to create a corpus, pre-process text data and apply a dictionary.

Today's Learning Materials

You can download the learning materials for today from
https://bit.ly/TXTatUOS2324

(the URL is CASE sensitive)

What if I get stuck…

If you have never used R, today's session will not be easy to follow.

If you have used R, but never used quanteda before, you may get stuck at some point. If that happens, here's my advice:

1. Do **not stress out**. If something does not run, ask the person sitting next to you.
2. If they can't help, you can copy and paste any errors/warnings (you'll see them popping up in red) on ChatGPT.
3. If nothing works…focus on **understanding the logic** behind the process described in the notes.

Outline

- **Logistics**
- Why Computational Text Analysis
- Bag of Words Approach (BoW)
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Outline

- ✓ Logistics
- **Why Computational Text Analysis**
- Bag of Words Approach (BoW)
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Why Computational Text Analysis?

- Social Scientists have always used **"texts as data"**.
    - Legal researchers have analysed court documents
    - Political scientists have analysed parliamentary debates
    - Media scholars have analysed newspaper articles
    - …
- There are costs (e.g., **human labor**) to large-scale text analysis.
- Computers can **lower these costs**.
    - Growth in computational power at relatively low costs.
    - Facilitated by widespread digitization of information.

Adapted from Terman (2018)

#1: Uncivility in online comments

- Question:
  - Does the frequency of **uncivil messages** significantly differ between political topics and nonpolitical topics?
- Data:
  - 17.5M comments
- Method:
  - Dictionary method

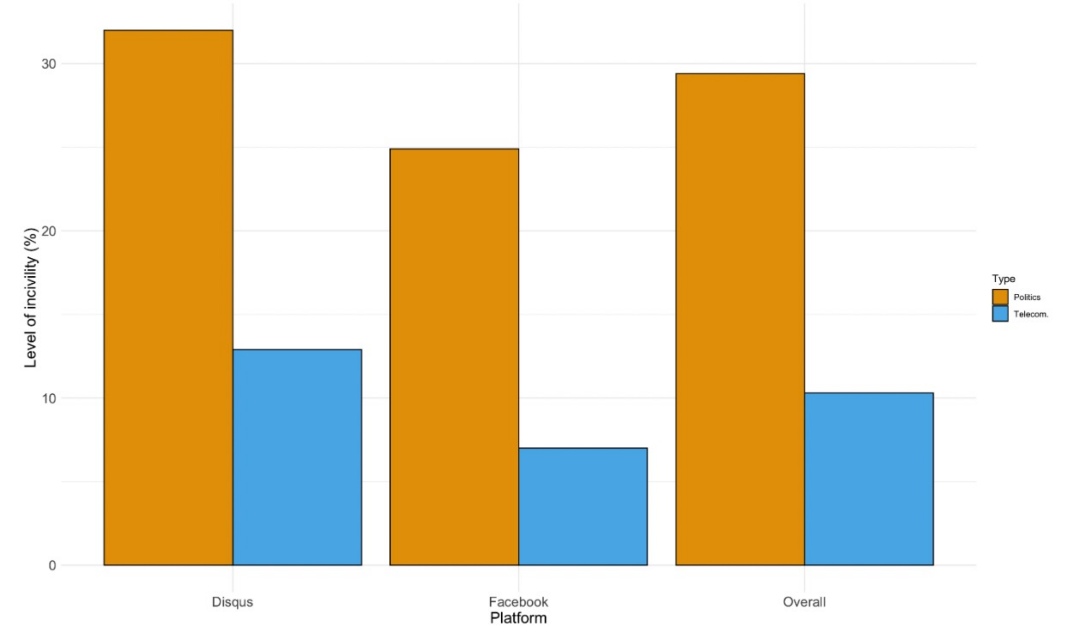


Figure 1. Average level of incivility by topics and platform.

Szabó, Kmetty & Molnár (2021)

#3: Taxpaying, political system & the media

- Question:
  - What explains **coverage of news about taxation** around the world?
- Data:
  - News articles (500.000+)
- Method:
  - Document classification (supervised)

Table 3. Binary Logistic Regression Predicting the Framing of a Taxpayer in Public Spending Terms, With the Democracy Level Measured by the Reverse-Coded Freedom House Score.

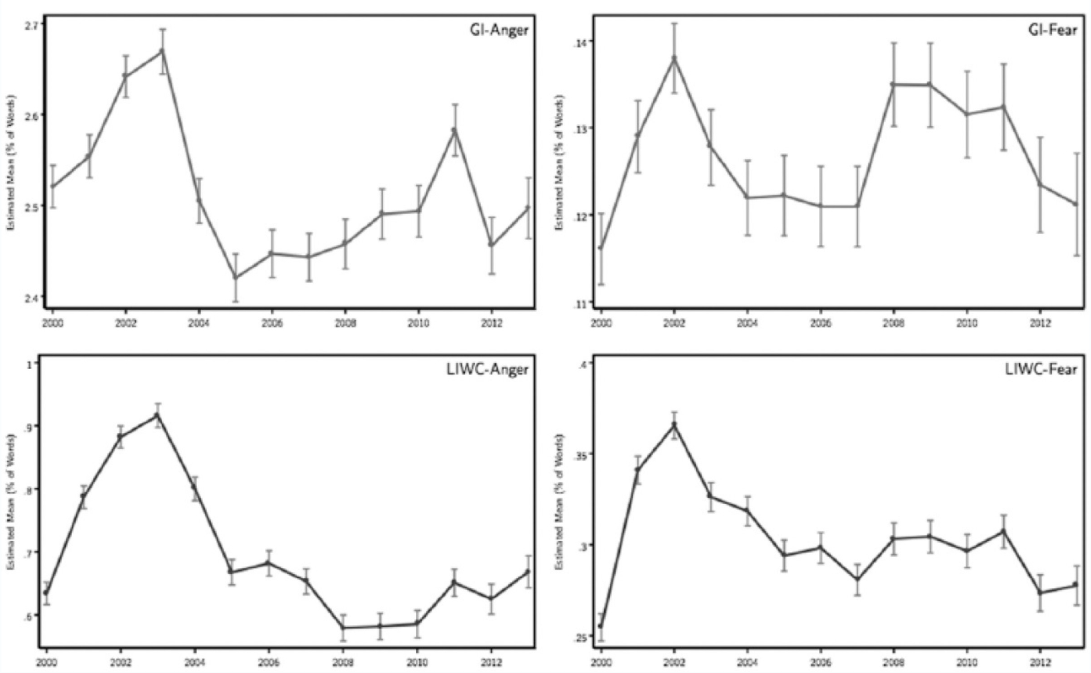
	(Model 1)		(Model 2)		(Model 3)	
	B (*SE*)	Exp (*B*)	*B* (*SE*)	Exp (*B*)	*B* (*SE*)	Exp (*B*)
Country-level variables						
Democracy level	0.130*** (0.020)	1.139	0.004 (0.023)	1.004	0.024 (0.023)	1.025
Tax reliance	0.056*** (0.005)	1.057	0.044*** (0.005)	1.045	0.050*** (0.005)	1.052
Newspaper-level variables						
State ownership			−1.039*** (0.102)	0.354	1.194*** (0.365)	3.301
News agency			−0.834*** (0.116)	0.434	−0.920*** (0.119)	0.399
Tabloid			−0.056 (0.063)	0.945	−0.066*** (0.064)	0.936
Document-level variables						
Domestic context					−0.567*** (0.063)	0.567
State ownership × Domestic context					−2.259*** (0.360)	0.104
Constant	−0.533*** (0.112)	0.587	0.686*** (0.139)	1.986	0.892*** (0.146)	2.439
N	23,343[a]		23,343[a]		23,343[a]	
Nagelkerke R^2	.029		.055		.069	
Classification accuracy	84.3		84.5		84.9	

a. Lower than the original *N* = 25,191 due to missing data on the tax reliance measure for year 2015.
***$p \leq .001$, two-tailed.

Kananovich (2018)

#2: Estimating sentiment in news stories

- Question:
  - Do **sentiments of fear an anger** fluctuate over time in the *New York Times* and *The Washington Post* newspapers?
- Data:
  - 55,000 news stories
- Method:
  - Sentiment Analysis



Soroka, Young & Balmas (2015)

What Can Computational Text Methods Do?

Haystack metaphor

Source: https://commons.wikimedia.org/wiki/File:Haystack.png

What Can Computational Text Methods Do?

Haystack metaphor ~ **Improve Reading**

✗ Interpreting meaning of a phrase [**Analyzing a straw of hay**]

- Humans: amazing! (Straussian political theory, analysis of English poetry…)
- Computers: struggle ☹

Comparing, Organizing, & Classifying Texts [**Organizing haystack**]

- Humans: terrible. Tiny active memories 😐
- Computers: amazing!

Grimmer & Stewart (2013)

Principles of Computational Text Analysis

1. All quantitative models of language are **Wrong** – but some are useful.
2. Quantitative methods for text **amplify** resources and augment humans – but they do not replace them
3. There is **no globally best method** for automated text analysis.
4. Validate! **Validate**! Validate!

Adapted from Grimmer and Stewart (2013)

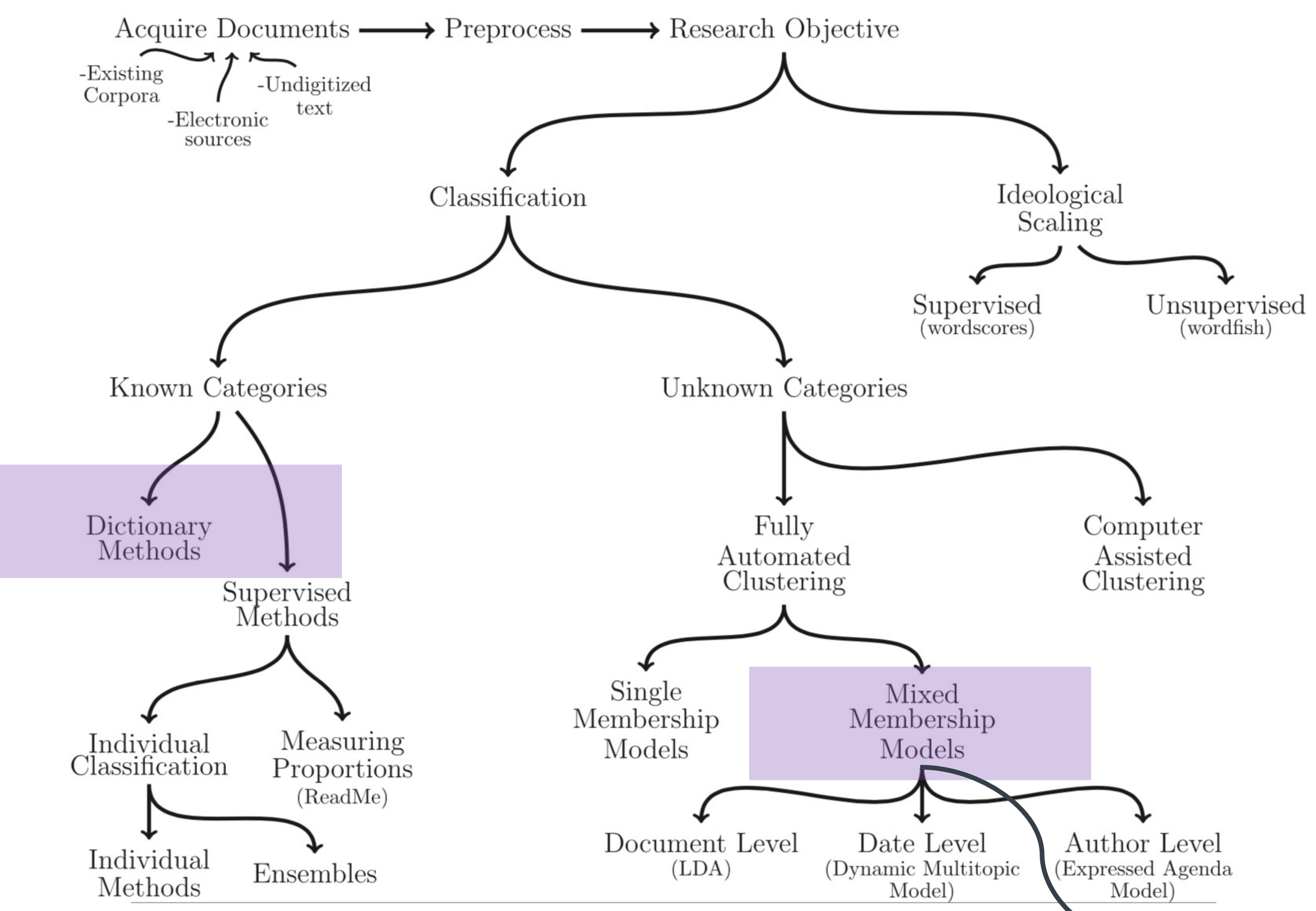


Fig. 1 An overview of text as data methods.

Structural Topic Model (stm)

Grimmer and Stewart (2013)

Outline

- ✓ Logistics
- **Why Computational Text Analysis**
- Bag of Words Approach (BoW)
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ▪ **Bag of Words Approach (BoW)**
- ▪ R demo: Introduction to quanteda
- ▪ R demo: Text pre-processing
- ▪ R demo: Dictionary methods
- ▪ Next steps

Assumptions of QTA

- Texts can represent a measurable latent or manifest characteristic of interest to researchers such as…
  - An **attribute** of the author (e.g., an author's ideology)
  - A **topic or theme**
  - A **sentiment or emotion**
  - The salience of a **political issue**
  - …

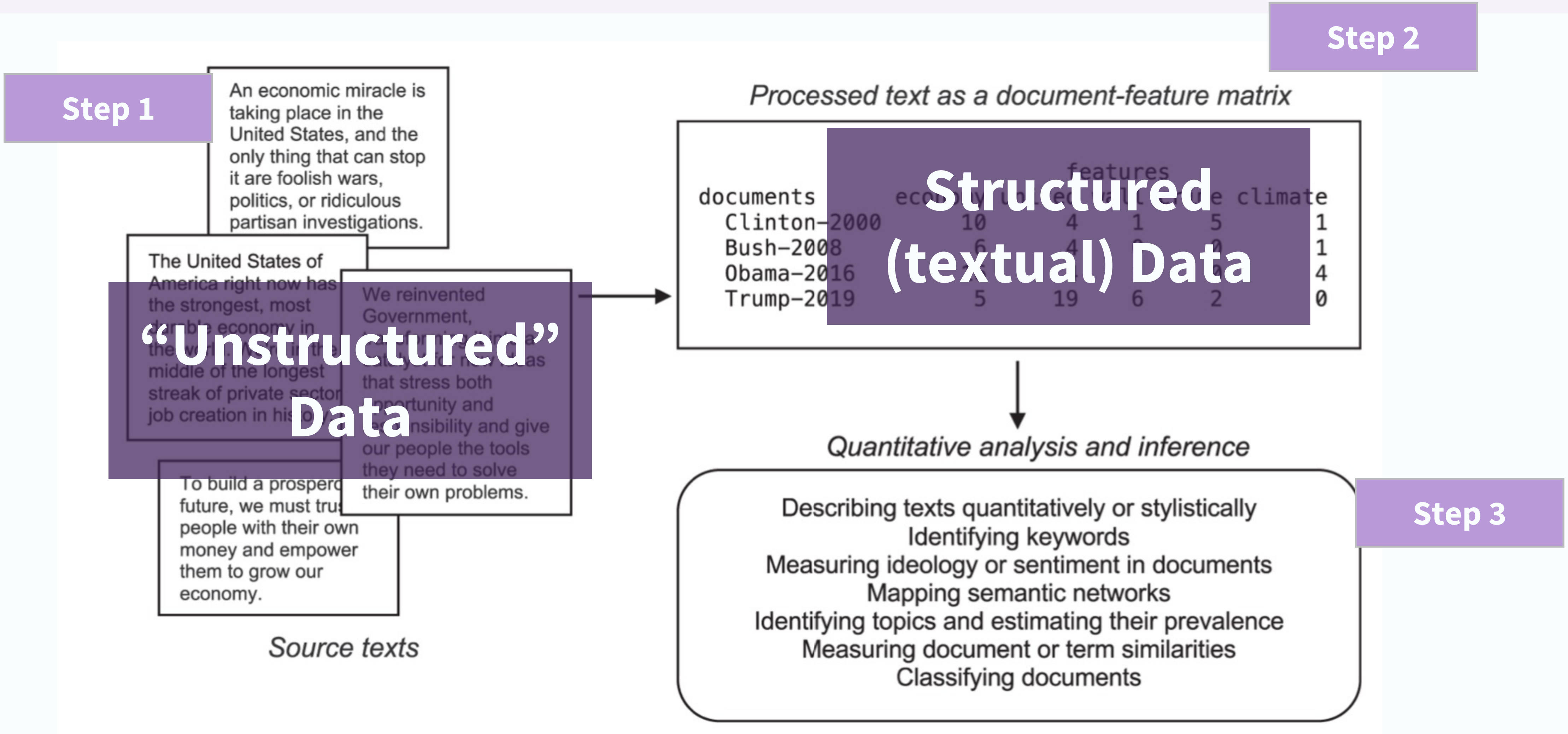
Adapted from Barberá & Benoit (2018)

Assumptions of QTA

- Texts can be represented through extracting their features
  - the most common is the "**bag of words**" assumption
  - other approaches based on "**string of words**" are becoming prevalent

Barberá (2018)

Text → DTM/DFM → Analysis



Benoit (2020)

Assumptions of QTA

- Texts can be represented through extracting their features
  - most common is the "**bag of words**" assumption
  - other approaches based on "**string of words**" are becoming prevalent
- A **document-feature matrix (DFM)** can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Barberá (2018)

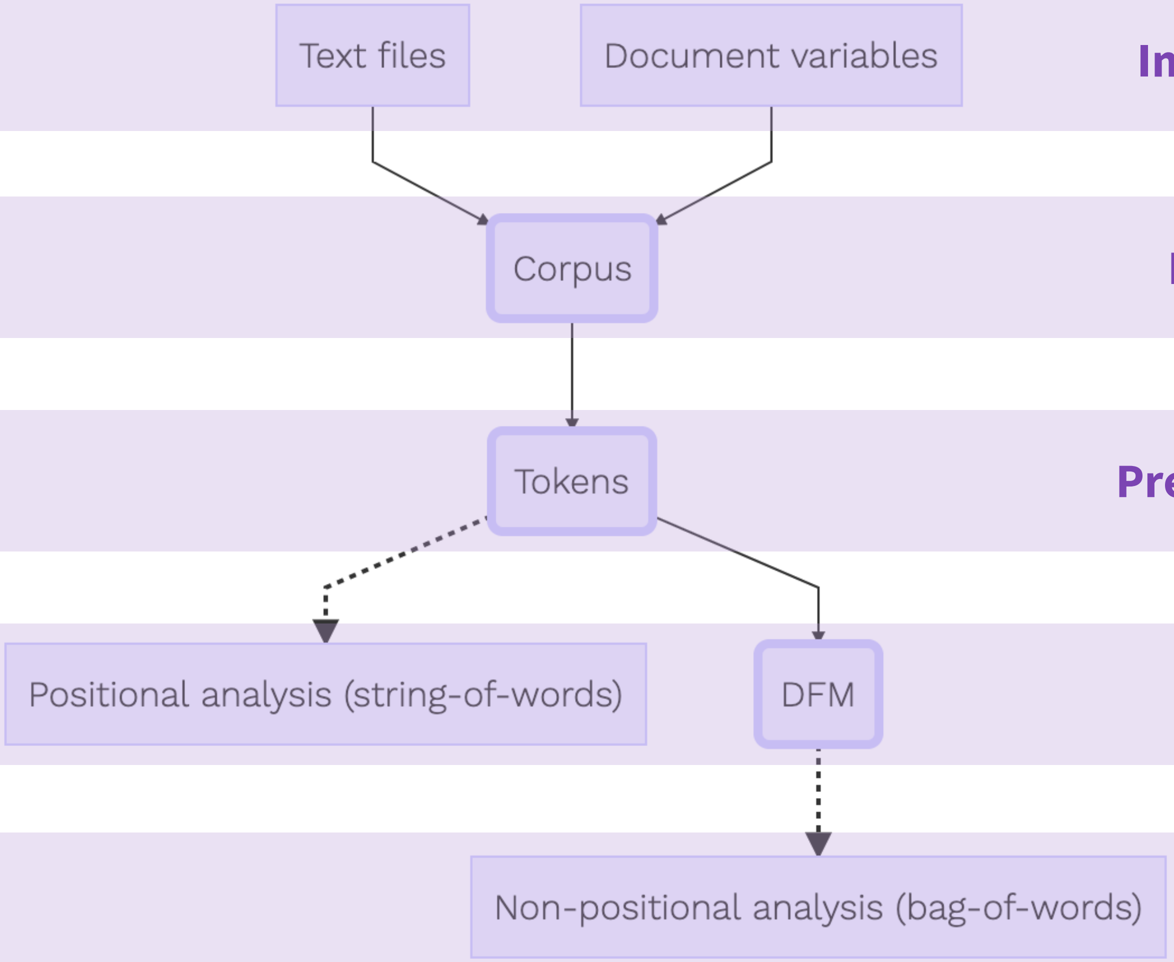
Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- **Bag of Words Approach (BoW)**
- R demo: Introduction to quanteda
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- **R demo: Introduction to quanteda**
- R demo: Text pre-processing
- R demo: Dictionary methods
- Next steps

Standard QTA procedure in quanteda



Importing documents and metadata (or docvars)

Binding a collection of documents into a corpus

Pre-processing text data and selection of features

Determining a suitable approach: BoG or SoW

Analysis of texts using BoW based approaches

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- **R demo: Text pre-processing**
- R demo: Dictionary methods
- Next steps

Key features 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 (segmentation or aggregation)

Barberá (2017)

Key features of quantitative text analysis

4. Defining and refining 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

Barberá (2017)

Preprocessing for QTA

- This is one (of many) recipes for preprocessing. The end goal is to **retain useful information only**.
  1. Remove capitalization, punctuation
  2. Discard Word Order (Bag of Words Assumption)
  3. Discard stop words
  4. (Create Equivalence Class: Stem, Lemmatize, or synonym)
  5. Discard less useful features~ depends on application
  6. Other reduction, specialization
- Output: Count vector, each element counts **occurrence of tokens**

Grimmer (2018)

Document-term matrix (or DTM)

	Word 1	Word 2	Word 3	Word 4	Word 5	⋮	*M* Words
Document 1	1	3	2	0	0	…	
Document 2	0	0	1	1	0	…	
Document 3	1	1	0	2	3	…	
Document 4	3	1	0	0	0	…	
Document 5	0	1	0	3	1	…	
…							
Document *n*	0	1	1	0	1	…	

$$\times = \begin{pmatrix} 2 & 1 & 0 & \cdots & 2 \\ 1 & 0 & 1 & \cdots & 3 \\ 3 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & \cdots & 3 \end{pmatrix}$$

From words to numbers

1. Preprocess text (raw data)

Tweet 1 **“@MyPolitician thank you and congratulations, you’re the best #elections”**

Tweet 2 **“@MyPolitician You’re an enemy, I would never vote for you”**

From words to numbers

2. Preprocess text: lowercase

Tweet 1	**“@MyPolitician thank you and congratulations, you’re the best #elections”**
	“@mypolitician thank you and congratulations, you’re the best #elections”
Tweet 2	**“@MyPolitician You’re an enemy, I would never vote for you”**
	“@mypolitician you’re an enemy, i would never vote for you”

From words to numbers

3. Preprocess text: lowercase, remove stop words, remove punctuation

Common English stop words

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

N = 175

From words to numbers

3. Preprocess text: lowercase, remove stop words, remove punctuation

Tweet 1	**“@MyPolitician thank you and congratulations, you’re the best #elections”**
	“@mypolitician thank congratulations best #elections”
Tweet 2	**“@MyPolitician You’re an enemy, I would never vote for you”**
	“@mypolitician enemy never vote”

From words to numbers

4. Preprocess text: lowercase, remove stop words, remove punctuation, stem, tokenize

Tweet 1	**“@MyPolitician thank you and congratulations, you’re the best #elections”**
	“@mypolitician thank congratul best #elections”
Tweet 2	**“@MyPolitician You’re an idiot, I would never vote for you”**
	“@mypolitician enemy never vote”

From words to numbers – DFM (Document Feature Matrix)

	@mypolitician	thank	congratul	never	#elections	enemy	best	vote
Document 1	1	1	1	0	1	0	1	0
Document 2	1	0	0	1	0	1	0	1

From words to numbers

1. Bag-of-words assumption
2. Pre-processing text
   - Capitalization, cleaning digits/URLs, removing stop words and sparse words…
   - Stemming
   - [Part-of-speech tagging]
3. Document-term matrix
   - **W**: matrix of *N* documents by *M* unique words
   - *Wim* = number of times *m*-th words appears in *i* -th document.
   - Usually large matrix, but sparse (so it fits well in memory)

Barberá (2016)

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- **R demo: Text pre-processing**
- R demo: Dictionary methods
- Next steps

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- ✓ R demo: Text pre-processing
- **R demo: Dictionary methods**
- Next steps

Dictionaries

- Dictionaries are **lists of words** belonging to a category.
- Dictionaries have two components:
  - key ~ the label for the **equivalence class** for the concept or canonical term
  - values ~ (multiple) terms or patterns that are declared equivalent occurrences of the key class

Adapted from Terman (2018)

Dictionaries – Autocratic/Democratic Policies

autocratic			**key**
authoritarian	criminal*	delinquen*	**values**
destructive	instability	lawless*	
offence*	prison*	riot*	
sedition	thug*	unlawful	

democratic			**key**
democracy	consensus	deliberate*	**values**
dialogue	election	equal*	
freedom	justice	multilateral	
parliament*	peace*	redistribution	

Maerz (2019)

Dictionaries

- Dictionaries are **lists of words** belonging to a category.
- Dictionaries have two components:
  - key ~ the label for the **equivalence class** for the concept or canonical term
  - values ~ (multiple) terms or patterns that are declared equivalent occurrences of the key class
- Rather than count ALL words that occur in a text we count **pre-defined words** associated with specific meanings.

Adapted from Terman (2018)

Dictionary structures

- Keys can be **labels** or **weights** or **scores**
  - Binary:
    - {Positive, Negative}
    - {Positive = +1, Negative = -1}
  - Numerical: {−2,−1,1,2}

Terman (2018)

Sentiment Dictionaries

Bing Liu Sentiment Lexicon

word	label
zombie	negative
zippy	positive
zest	positive
zenith	positive
zealously	negative
zealot	negative
zeal	positive
zaps	negative
zapped	negative
zap	negative

AFINN-111 Dictionary

word	value
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
adduction	-2
abhor	-3
abhorred	-3
abhorrent	-3
abhors	-3
abilities	2

Loughran-McDonald Lexicon

word	value
compelling	constraining
compensatory	litigious
complain	negative
compliment	positive
confuses	uncertainty
extant	superfluous
Failed	negative
forego	negative
honors	positive
hurt	negative

Dictionary structures

- Keys can be **labels** or **weights** or **scores**
  - Binary:
    - {Positive, Negative}
    - {Positive = +1, Negative = -1}
  - Numerical: {−2,−1,1,2}
- Non-sentiment dictionaries:
  - Words about sports, food, places...

Terman (2018)

Thematic Dictionaries

Populism (Rooduijn & Pauwels, 2011)

word	label
elit*	populism
consensus*	populism
corrupt*	populism
betray*	populism
establishm*	populism
scandal*	populism
truth*	populism
ruling*	populism
referend*	populism
shame*	populism

Laver-Garry policy positions in the UK

word	value
media	culture
opera*	culture
museum*	culture
emission*	environment
recycl*	environment
warming	environment
assault	law & order
court	law & order
lawless*	law & order
police	law & order

Newsmap geographical dictionary (Watanabe)

word	key level 1	key level 2
China	CN	East Asia
Beijing	CN	East Asia
Japan	JP	East Asia
Tokyo	JP	East Asia
Thailand	TH	Southeast Asia
Bangkok	TH	Southeast Asia
Indonesia	ID	Southeast Asia
Jakarta	ID	Southeast Asia
India	IN	South Asia
Mumbai	IN	South Asia
New Delhi	IN	South Asia

Dictionary methods

- Classifying documents when **categories are known** using dictionaries:

1. Lists of words that correspond to each category:
   - Positive or negative (for sentiment)
   - Sad, happy, angry, anxious (for emotions)
   - Insight, causation, discrepancy, tentative (for cognitive processes)
   - Sexism, homophobia, xenophobia, racism (for hate speech)

Adapted from Barberá (2016)

Dictionary methods

2. Count **number of times** they appear in each document
3. Normalize by document length (optional)
4. Validate, **validate**, validate.
   - Check sensitivity of results to exclusion of specific words
   - Code a few documents manually and see if dictionary prediction aligns with human coding of document

Adapted from Barberá (2016)

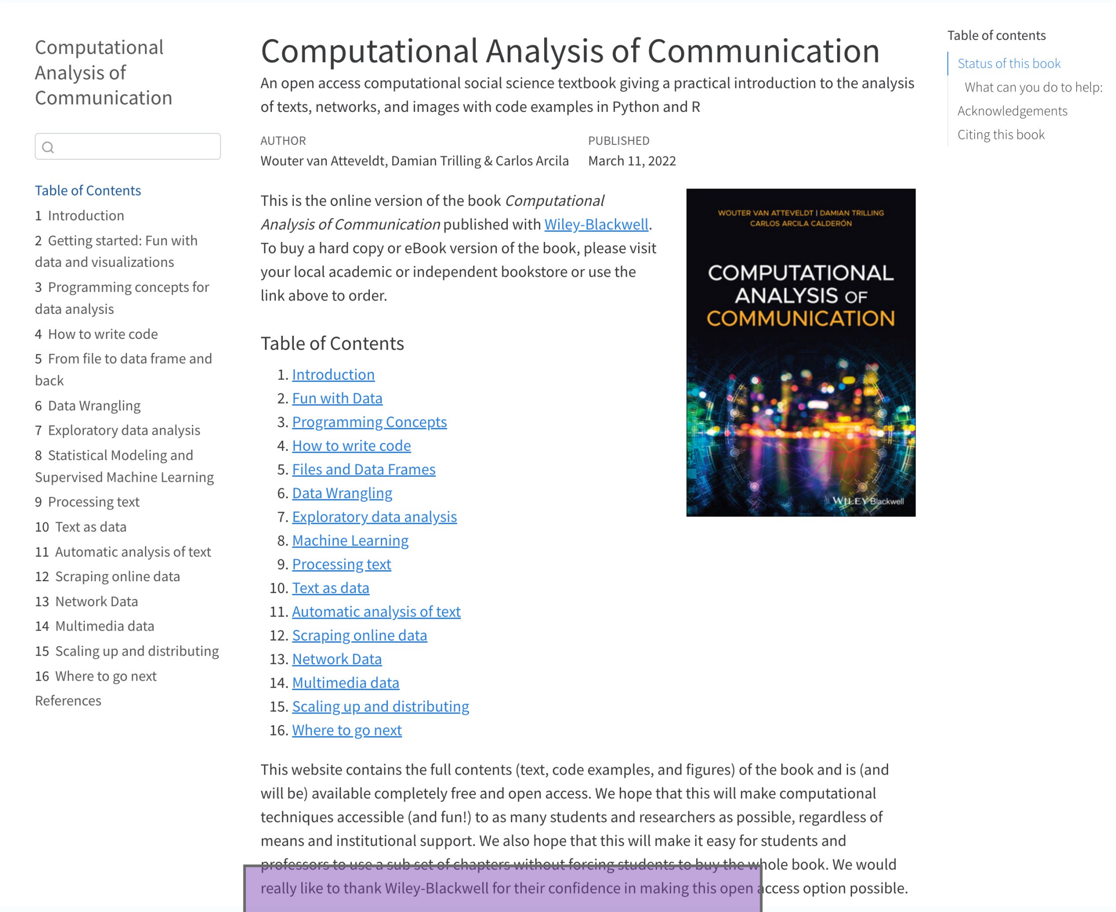
Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- ✓ R demo: Text pre-processing
- **R demo: Dictionary methods**
- Next steps

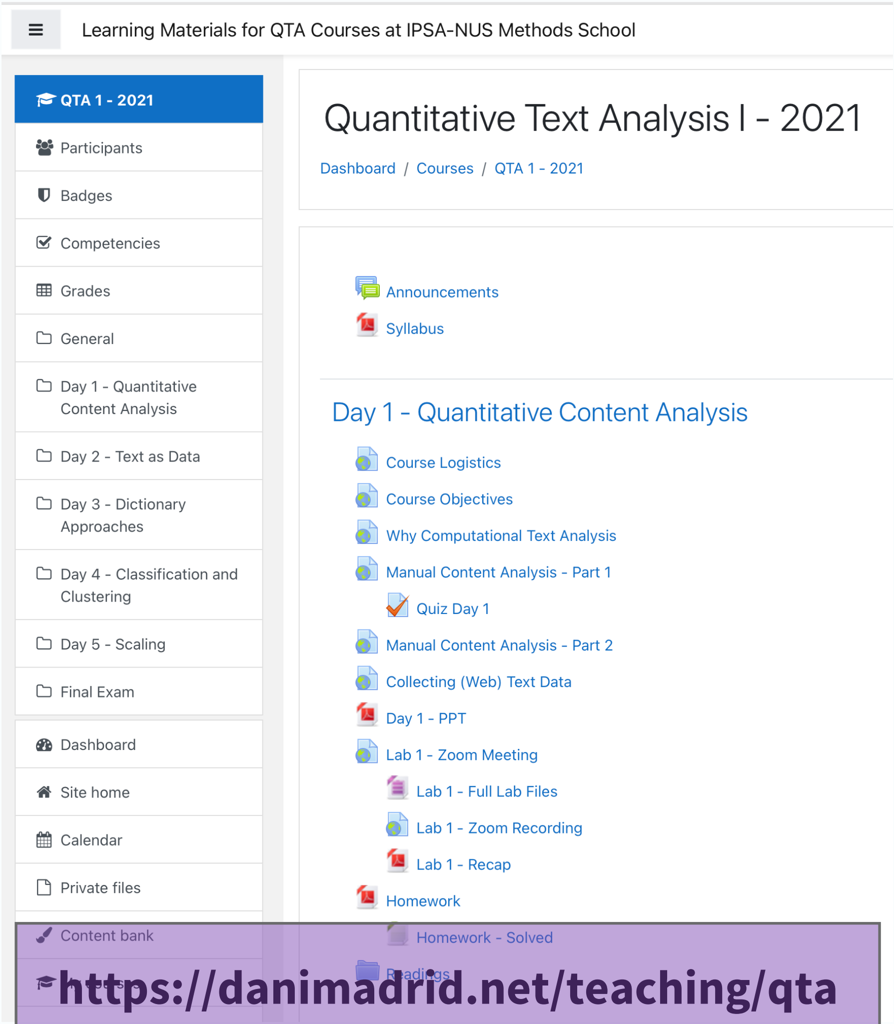
Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- ✓ R demo: Text pre-processing
- ✓ R demo: Dictionary methods
- **Next steps**

Want to learn more?



https://cssbook.net/



https://danimadrid.net/teaching/qta

Outline

- ✓ Logistics
- ✓ Why Computational Text Analysis
- ✓ Bag of Words Approach (BoW)
- ✓ R demo: Introduction to quanteda
- ✓ R demo: Text pre-processing
- ✓ R demo: Dictionary methods
- ✓ Next steps

Sources

Barberá, P. (2017). *POIR 613 (Fall 2017)*. https://github.com/pablobarbera/POIR613-2017

Barberá, P. (2019). *Lecture materials for MY459, LT 2019*. https://github.com/pablobarbera/lectures-1

Benoit, K. (2014). *The Quantitative Analysis of Textual Data (NYU Fall 2014)*. https://kenbenoit.net/nyu2014qta/

Benoit, K. (2018). *Quantitative Text Analysis (TCD 2018)*. https://kenbenoit.net/quantitative-text-analysis-tcd-2018/

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Terman, R. (2018). *PS239T: Introduction To Computational Tools And Techniques For Social Research*. https://github.com/rochelleterman/PS239T

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. The ANNALS of the American Academy of Political and Social Science, 659(1), 108–121. https://doi.org/10.1177/0002716215569217

Szabó, G., Kmetty, Z., & Molnár, E. K. (2021). Politics and Incivility in the Online Comments: What is Beyond the Norm-Violation Approach? *International Journal of Communication*, 15, 1659–1684, https://ijoc.org/index.php/ijoc/article/view/16411

University of Sheffield

(Computational) Text Analysis 1

Session 1 – Introduction & Dictionary Methods

22 November 2023

Dr. Dani Madrid-Morales