university of groningen

Quantitative Text Analysis – Essex Summer School

Introduction to text as data

dr. Martijn Schoonvelde

University of Groningen

Today's plan

- Getting to know each other
- Setting up the course
- What is quantitative text analysis? Principles of developing a corpus
- Practice working (text) data in RStudio

Who am I?

Assistant professor in European Politics & Society at University of Groningen

- PhD from Stony Brook University. Previously worked at Exeter, EUI, UCD, Vrije Universiteit Amsterdam
- Research focuses on patterns of rhetorical responsiveness and democratic accountability in Europe
- Also interested in political methodology, specifically text as data
- Email: martijn.schoonvelde@rug.nl

TA: James Rice

PhD candidate in the Department of Government at the University of Essex, working at the interface of computational social science, political economy, and finance and accounting.

- MSc degrees in Philosophy from LSE and the University of Edinburgh and a BS in Economics
- Research focuses on the impact of climate misinformation on financial markets and political values
- Email: james.k.rice@essex.ac.uk

Contact

- Ask questions, come talk to us – happy to help / set up a Zoom meeting
- Use the Slack workspace for this module to communicate with each other: `essqta25.slack.com`
  - We will regularly check the workspace and engage in conversations
- All materials (slides / code scripts / etc) available at `https://github.com/hjmschoonvelde/ess_qta_2025`

Who are you?

- Tell us a bit about yourself. Why did you choose this course? What would make this a successful course for you?

Making online meetings a success

- We'll take regular breaks (after each hour)
- We will combine lectures with hands-on coding sessions (making some use of breakout groups but not too much)
- Ask questions! Engage! Either through Slack on during our meetings.

Quantitative text analysis

The New York Times

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Sept. 5, 2018

13133

Leer en español · 阅读简体中文版 · 閱讀繁體中文版 · 한국어로 읽기 · 日本語で読む

Written by **"Anonymous"**

Quantitative text analysis

The New York Times

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Sept. 5, 2018

13133

Leer en español · 阅读简体中文版 · 閱讀繁體中文版 · 한국어로 읽기 · 日本語で読む

"We may no longer have Senator McCain. But we will always have his example – a lodestar for restoring honor to public life and our national dialogue."

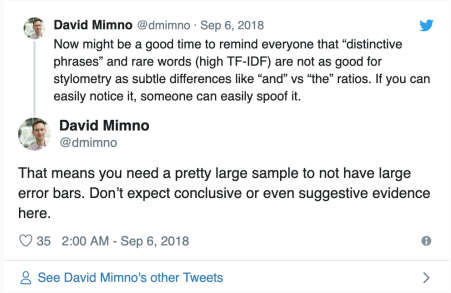
“Lodestar”

‘a person or thing that serves as an inspiration or guide’

Quantitative text analysis

Different approach that does not a priori rely on most noticeable words (aka features in text as data parlance)

1. Collect textual data from a range of potential authors
2. Compare their relative use of words across the entire vocabulary
3. Calculate the probability that the unknown document was written by each one of the authors based on the words it contains
4. Inspect which features are most predictive (which may or may not be the most noticeable words)



“Anonymous”

In October 2020, “Anonymous” revealed himself to be Miles Taylor, a former senior Trump administration official in the DHS.

What is quantitative text analysis

An approach to learning from text that relies on quantification of its textual contents.

- Different from, for example, discourse analysis, which is generally more interested in interpretation, in reading between the lines (Benoit, 2020)

We can distinguish between manual approaches and computational approaches to QTA

- . . . or a combination of both – e.g., computational grounded theory (Nelson, 2020); hybrid content analysis (Baden *et al.*, 2020)

Our focus in this class is on learning about such computational approaches (we'll encounter dictionary, supervised methods, semi-supervised, and unsupervised methods)

What is quantitative text analysis

Computational quantitative text analysis is not one-size-fits all, but highly task-dependent. Generally, an application follows three steps;

1. Identify texts and units of analysis
   - Develop a corpus
2. Extract quantitatively measured features from these texts and convert them to a quantitative feature matrix
   - Decide on the most informative way to represent the text for the research question at hand
3. Analyse this matrix with statistical methods to draw inferences about these texts. Or use this matrix as an input for downstream tasks (in the case of supervised machine learning)

Why quantitative text analysis?

- As humans we produce huge amounts of text, much of which is stored online
  - Speeches, books, interviews, blog posts, manifestos, social media posts, institutional documents, etc.
- These texts are often rich in information, but we'll have to separate the signal from the noise
  - Data is becoming cheaper over time, but the cost of thought is at least as high as before (Grimmer, Roberts, Stewart, 2022)
- Doing so requires a new set of tools and methods, which quantitative text analysis provide

Goals of quantitative text analysis

Goals are much in line with more general social science research objectives

- Exploration – discover a question of interest, generate hypotheses, and formulate a conceptualization
- Measurement – use text as an expression of a latent concept of interest
- Inference – using text to make causal or descriptive statements about a social phenomenon

This course

- Introduction of (computational) quantitative text analysis methods using R
- We'll cover the bigger picture of doing research using text – use this course to figure out what interests you and what you want to pursue further
- Ask questions – and help each other out

This course

- Lots of cool developments across disciplines!
  - In computer science and computational linguistics (natural language processing)
  - But also in communication science and psychology, economics and history (digital humanities)
- Lots of cool developments outside of academia
  - LLMs. GenAI

This course

- Day 1: What is QTA? Developing a corpus
- Day 2: Core assumptions in QTA. Text Representations, Preprocessing and feature selection.
- Day 3: Advanced text representations. Word embeddings
- Day 4: What can we do with dictionaries and how can we validate them? Sensitivity and specificity.
- Day 5: Human coding (or machine coding) and document classification using supervised machine learning. Evaluating a classifier.
- Day 6: Supervised, semi-supervised and unsupervised approaches to place text on an underlying dimension.
- Day 7: Understanding topic models. Discussing their pros and cons.
- Day 8: Multilingualism. Automated speech recognition. Images as data.
- Day 9: Deep learning. Transfer Learning.
- Day 10: LLMs. Concluding remarks.

A Note on QTA and NLP

- The way we conduct Quantitative Text Analysis (QTA) is changing rapidly with advancements in technology. Increasingly, text analysis leverages pre-trained large language models. These models are adapted for specific tasks using transfer learning.
- We interact with these models through natural language, making them more accessible and intuitive to use.
- Yet principles of research design remain critically important. It's essential to:
  1. Clearly define research questions and hypotheses
  2. Ensure the quality of the data
  3. Appropriately choose and justify the NLP methods used
  4. Rigorously validate and test the models to avoid biases and ensure generalizability

Requirements: grit

Requirements: fun

Course objectives

- Learn how computational text analysis methods are used in social science
- Practice various ways of analyzing text using R
- Critically evaluate existing text as data research
- Get to know fellow aspiring text analysts
    - Personal note: It's really fun to see the work that alumni from this course (now in its 4rd year) are doing and presenting at conferences

Why R?

- Encompasses all steps of the research process (from collecting data to analysis and visualization)
- Tremendously helpful user community
- Lots of developments, new packages
  - We'll mostly rely on quanteda (Benoit *et al* 2018), tidyverse (Wickham *et al.*, (2019), ggplot2 (Wickham, 2016), and stringr (Wickham, 2019)
- Other languages such as Python have a head-start in natural language processing – but developers are building wrappers to access their functionality in R, such as, spacyR (Benoit & Matsuo, 2020) for sentence parsing, grafzahl (Chan, 2023) and text (Kjell *et al.*, 2023) for transformers, rollama (Gruber & Weber, 2024), quanteda.llm (Maerz & Benoit, 2025), ellmer (Wickham *et al.*, 2025) for locally deployed LLMs.


R Studio®

GitHub copilot

- Once you get the hang of coding in R, I much recommend GitHub copilot as a coding assistant
- Code completion tool that supports a wide range of programming languages
- Excellent integration with RStudio
- See Cooper *et al.* (2024) for a discussion about the pros and cons.

Developing a corpus

Principles of selection (Grimmer, Roberts and Stewart, 2022)

1. The usefulness of a corpus depends on the question the researcher wants to answer and the population they want to study
   - The preponderance of textual data doesn't mean all text is useful. E.g., X data may be less useful for studying public opinion (increasingly so)
2. There is no values-free construction of a corpus. Selecting which documents to include has ethical ramifications
   - Just because it is found data (Salganik, 2018) doesn't mean you can just run with it

Developing a corpus

Be mindful of four types of possible bias (Grimmer, Roberts and Stewart, 2022)

1. Resource bias – texts often better reflect populations with more resources to produce, record and store documents
2. Incentive bias – strategic behavior can drive the production and retention of documents
   - E.g., skydive pictures on Instagram; hot takes on *X*
3. Medium bias – medium in which text is produced may constrain its content
   - E.g., more polite political discussions on Twitter after doubling of character limit (Jaidka *et al.*, 2019). See also algorithmic drift (Salganik, 2018)
4. Retrieval bias – our methods to sample documents may have biases baked into them
   - E.g., searching for newspaper articles about 'the economy' using 'economy' as a keyword

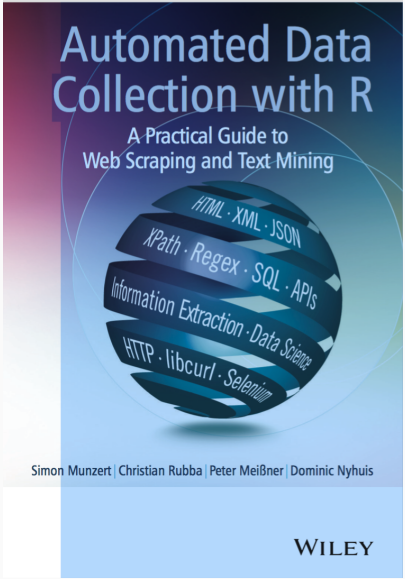
Pre-existing corpora

- Repositories such as Lexis Nexis or Factiva (newspaper data)
- Existing text datasets. For example:
  - EUSpeech (Schumacher *et al.*): Parlspeech (Rauh *et al.*); ParlEE (Silvester *et al.*);
  - UN General Debate Corpus (Jankin *et al.*)
  - DICEU (Wratil & Hobolt)
  - ECB speeches: `https://www.ecb.europa.eu/press/key/html/downloads.en.html`
  - Party manifestos: `https://manifesto-project.wzb.eu/`
  - A general repository of political datasets: `https://github.com/erikgahner/PolData`
- Replication data folders

Getting data from the web

- APIs (Application Program Interface) – make data sitting on a website available to you
  - Various R packages to access such APIs: **GuardianR, WikidataR, openai**
- Web scraping / screen scraping
  - **rvest**
- Data in APIs or scraped from webpages often stored in JSON, HTML and XML format – expect lots of data wrangling
  - **rjson, jsontools**

Getting data from the web



Representations of text

There are many different ways to represent text: **bag of words, word embeddings, sentence embeddings, document embeddings, dependency trees**, to name just a few

- These representations vary in their complexity and in the information they contain.

There is no one right way to represent text for all research questions (Grimmer, Roberts, Stewart, 2022) – it really depends on the question

- What is the quantity of interest that you are trying to measure? How will it manifest itself in the text?

Tomorrow

- Read the assigned papers
- Make sure you are up to speed with using RStudio
- Look at the following snippet of text and list all the ways (you can think of) that it needs to be cleaned:

```
<p>Ladies and gentlemen,</p><p>It is an honour to be here today to introduce the
theme of 'recession and recovery'. If you will permit,   I would like to suggest that
this afternoon we focus more on recovery than on recession. I think we know enough
about the recession side   of the story.</p><p>It started with the fall of Lehman
Brothers on 15 September 2008.. I happened to be here, at the Blouin Creative
Leadership Summit, only ten days later. Everyone was talking about the collapse of
Lehman. They were shocked and alarmed.   But even then we could hardly imagine that
its impact would be so dramatic, so historic.</p><p>As we now know, this event
triggered a global financial and economic crisis. Governments were forced to give
cash injections running into billions to prevent an economic and financial meltdown.
When credit dried up and demand fell, businesses struggled  to keep their heads above
water, and many went under. Ordinary people's jobs, homes and pensions were at
risk.</p><p>
```