Quantitative Text Analysis with R

Julian Bernauer

2019-02-06

Course Abstract

The course "Quantitative Text Analysis" provides an introduction to the retrieval, preparation, visualization and analysis of text as data using R. We draw on social science and other text examples, namely European election manifestos, books by Mark Twain, large amounts of Tweets and others. The course covers some web scraping to obtain text, preparation including the construction of word frequency matrixes or dictionaries and visualization tools beyond word clouds. For the analysis of texts, topic models such as LDA (latent Dirichlet allocation), scaling models including Wordscores and Wordfish as well as alternatives based on natural language processing tools (e.g. word embeddings) are discussed. One further theme is the cross-lingual and -contextual analysis of text. The participants also have the opportunity of helping to shape a textbook on the topic, which is contracted with SAGE and scheduled to appear in 2020.

Instructor

Julian Bernauer is a Postdoctoral Fellow at the Data and Methods Unit of the MZES (University of Mannheim). He is currently working on a research project measuring populism from political text and generally interested in Data Science. Contact: julian.bernauer@mzes.uni-mannheim.de.

Place and time

Room 211, B6, Monday 17:15-18:45, weekly February 11 - May 27 (except April 15 and 22)

Requirements

For 5 ECTS, the participants are expected to do the following for a total of ~150 hours:

- be mildly interested
- attend as many sessions as possible
- read according to good practice
- participate and critisize the chapter stubs from the book project
- present their project
- write and submit a paper with a text-as-data application (4-5k words) and replication material

Software (please install)

- R as statistical programming environment
- RStudio as editor
- quanteda package in R as workhorse for QTA
- RTools for Windows if needed
- Python to be called from R

Some useful books

- Silge, Julia and David Robinson (2017): Text Mining with R. A Tidy Approach. O'Reilly. Click here for open access.
- Ignatow, Gabe and Rada Mihalcea (2017): Text Mining. A Guidebook for the Social Sciences. SAGE.
- Wickham, Hadely and Garrett Grolemund (2017): R for Data Science. O'Reilly. Click here for open access.

Slides and code

- Are made available before sessions
- See GitHub (click here)

Sessions

February 11 - Introduction to the Course

- What's in a word: motivation
- Who are you
- Course outline
- Warm-up exercise: European party manifestos

February 18 - Context and Basic Concepts of QTA

- State of the art of QTA: computer linguistics, social science and informatics
- Digitalization and "Big Data"
- Ethics in text analysis
- Ways of approaching text as data

Literature

- Chapter Stub 1 from "QTA with R"
- Chapter 1 of Silge and Robinson (2017)
- Grimmer, Justin, and Brandon M. Stewart (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3): 267-97.
- Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley (2015) Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis* 23 (2): 254-77.

February 25 - Using R for Text Analysis

R basics for QTA

- quanteda
- Other R packages: readtext, stringr, tm, tidytext…

Introducing he text corpora used as running examples

- Political science classics: party manifestos and speeches
- Exercise: hand-coding manifestos using the CMP scheme (AfD)
- Tweets: analyzing 280 (140) characters
- Some books from Mark Twain (gutenbergr package)
- Exercise: team session looking for other applications

Literature

- Chapter Stub 2 from “QTA with R”
- Welbers, Kasper, Wouter Van Atteveldt and Kenneth Benoit (2017): Text Analysis in R, *Communication Methods and Measures* 11(4): 245-65.

Resources

- quanteda
- quanteda on stackoverflow

March 04 - Using Python from R for Text Analysis

- Python and R compared
- Some quick Python programming: Raspberry Pi
- R interface to Python: reticulate
- Alternatives for working with Python

Literature

- Chapter Stub 3 from “QTA with R”

Resources

- Raspberry Pi Python documentation
- Python resources
- reticulate

March 11 - Obtaining and Evaluating Text

- From pdf to txt
- Some webscraping - the example of the Twitter API
- Using polidoc.net
- Evaluation of text and planning the quantitative analysis

Literature

- Chapter Stub 4 from “QTA with R”
- Chapter 3 of Ignatow and Mihalcea (2017)
- Further reading: Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis (2015) *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* Chichester: Wiley.

March 18 - Cleaning and Preparing Text

- Using R, especially quanteda for text preparation
- Annoying stuff: segmenting text, unwanted content, hyphenation…
- Exercise: clean this Bundestag protocol up (DVPW Blog on Bundestag Brexit debate)

Literature

- Chapter Stub 5 from “QTA with R”
- Chapter 5 of Silge and Robinson (2017)
- Chapter 5 of Ignatow and Mihalcea (2017)

March 28 - Extracting Information from Text

- Word frequency matrix
- Patterns in text
- Sentiment analysis
- Dictionaries: example populism dictionary Rooduijn and Pauwels (2011)

Literature

- Chapter Stub 6 from "QTA with R"
- Chapters 2-4 of Silge and Robinson (2017)
- Rooduijn, Matthijs, and Teun Pauwels (2011) Measuring Populism: Comparing Two Methods of Content Analysis. *West European Politics* 34(6): 1272-83.
- Chapters 10-14 of Ignatow and Mihalcea (2017)

April 1 - Visualizing Information from Text

- Bashing some word clouds
- Better visualizations of text as data using using quanteda, ggplot(2)…

Resources

- Chapter Stub 6 from "QTA with R"
- Textual data visualization via quanteda

April 8 - Scaling Text

- Scaling text with Wordscores and Wordfish
- Using word embeddings (Mikolov et al. 2013)
- Application: Measuring populism from manifestos using word embeddings

Literature

- Chapter Stub 8 from "QTA with R"
- Lowe, Will (2008) Understanding Wordscores. *Political Analysis* 16(4): 356-71.
- Slapin, Jonathan B., and Sven-Oliver Proksch (2008) A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science 52* (3): 705-22.
- Glavas, Goran, Federico Nanni and Simone P. Ponzetto (2017) Unsupervised Cross-Lingual Scaling of Political Texts. In: Proceedings of 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, Valencia, Spain, 3-7 April 2017, 688-93.
- Mikolov, Tomas, Quoc V. Le and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. CoRR, https://arxiv.org/abs/1309.4168.

April 29 - Validity and Reliability

- A protocol for the validation of QTA
- Application: rated manifesto sentences vs. word embeddings scores (populism)
- Exercise: coding Tweets for validation
- Outlook: Crowdsourced coding

Literature

- Goet, Niels D. (2017) Measuring Polarisation with Text Analysis: Evidence from the UK House of Comons, 1811-2015. Paper Prepared for the Polarization, Institutional Design and the Future of Representative Democracy Workshop, Berlin, Harnack Haus, 5-7 October 2017.

- Lowe, Will and Kenneth Benoit (2013) Validating Estimates of Latent Traits from textual Data using Human Judgement as a Benchmark. *Political Analysis* 21(3): 298-313.
- Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov (2016) Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review* 110 (2): 278-95.

May 06 - Classifying Text

- So many algorithms: an overview (Textbooks, Allahyari et al. 2017)
- Latent Dirichlet Allocation (LDA) in more detail (Blei et. al 2003)
- Mixing classification and scaling (Baerg and Lowe 2018)

Applications

- classifying Tweets as populist or not
- Classifying Tom Sawyer and Huckleberry Finn

Literature

- Chapter Stub 7 from "QTA with R"
- Chapter 6 of Silge and Robinson (2017)
- Chapter 15 of Ignatow and Mihalcea (2017)
- Allahyari, Mehdi et al. (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. KDD Bigdas, August 2017, Halifax, Canada.
- Blei, David M., Andrew Y. Ng and Michael I. Jordan (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.
- Baerg, Nicole and Will Lowe (2018) A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods. *Political Science Research and Methods* (First View): 1-17.

May 13 - Communicating Text Analysis

- How to communicate text analysis
- Common objectives against QTA
- Exercise: writing an (illustrated) news article

Literature

- Chapter Stub 10 from "QTA with R" (9 skipped on purpose)

May 20 - Presentation of Projects

- Research question, design and data for your QTA project and paper
- Time allocated fairly among presentation givers and discussants
- First analysis welcome
- Please mention the problems

May 27 - Farewell Party

- Wrap-up
- Coffee (yes) & Cigarettes (no)