UNIVERSITÄT
LUZERN

Quantitative Text Analysis for Beginners Using R and quanteda

Tutor	Stefan Müller, University College Dublin
Organization	Digital Skills, University of Lucerne
Language	English
ECTS-Points	3
Contact	Nadia.buehler@unilu.ch
Content	Designed for 15-20 students, this course will be delivered online over three days by Stefan Müller. The course provides a thorough quantitative text analysis with practical implementation using the quanteda R package. It starts with an overview of text analysis and working with textual data, including importing texts and constructing a corpus; tokenization; creating a document-feature matrix. Day two covers textual statistics, computation of similarity and distance, and visual methods for comparing texts. On the third and final day, we will cover more advanced methods such as scaling, classification, and topic modelling. All topics will be accompanied by hands-on exercises in R, done on the students' own computers, but using starter code supplied by the instructors. **1** Preparing for the course No prior experience with quantitative text analysis is needed, although the course presumes some familiarity with the R language. Students can prepare by making sure they have prepared the following: or read the following: • Installed R (at least version 4.0) • Installed RStudio Desktop

FROHBURGSTRASSE 3
P.O.BOX 4466
6002 LUCERNE

T

- Installed the following R packages: *quanteda, quanteda.textmodels, quanteda.textstats, quanteda.textplots, tidyverse, readtext*

These are the suggested readings for the course.

- Benoit, Kenneth. 2020. "Text as Data: An Overview." In Curini, Luigi and Robert Franzese, eds. *Handbook of Research Methods in Political Science and International Relations*. Thousand Oaks: Sage. pp. 461–497. URL: https://kenbenoit.net/pdfs/CURINI_FRANZESE_Ch26.pdf
- An Introduction to RMarkdown: https://rmarkdown.rstudio.com/lesson-1.html
- 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):267–297. doi: 10.1093/pan/mps028.

2 Schedule

Day 1: Overview of Text Analysis

Friday, 1 October 2021, 09:15–12:00 & 14:00–16:45

The first day will combine a comprehensive overview of quantitative text analysis with an introduction to the structure, logic, and syntax of the quanteda text analysis package. Topics will include:

- An introduction to quantitative text analysis and its workflow;
- A comprehensive map of quantitative text analysis including the main techniques used in political and social science;
- An overview of quanteda and how to use it to create core objects for textual analysis.

	Day 2: Describing, comparing, and scaling texts *Friday, 15 October 2021, 09:15–12:00 & 14:00–16:45* The second day will use a few running examples including: • Identifying key words based on statistical association measures; • Identifying multi-word expressions via collocation analysis; • Computing and interpreting textual statistics for comparing texts; • Computing similarities and distances for identifying clusters of similar texts. ***Day 3: Advanced text analysis*** *Friday, 29 October 2021, 09:15–12:00 & 14:00–16:45* The final day will cover advanced methods including: • Feature weighting and feature selection for textual matrices; • Supervised and unsupervised document scaling for measuring ideological positions; • Supervised machine learning for document classification; and • Topic modelling.