

Introduction to Quantitative Text Analysis using Quanteda, January 31, 2018

In this one-day workshop you will learn how to import, transform and analyze textual data using an R package [Quanteda](#). With flexible and intuitive commands, it allows social scientists to investigate a range of textual data from different angles with a minimal amount of programming. While the "bag-of-words" approach has been dominant in quantitative text analysis, this workshop introduces a set of functions for a "string-of-words" approach, in which multi-word expressions (e.g., proper names and phrases) are handled more appropriately. The techniques to be covered in this workshop are particularly useful in analysis of party manifestos, newspaper articles and micro-blog posts.

What is Quanteda?

Quanteda is an R package for quantitative text analysis developed by a team of political scientists and computer scientists. Since the first CRAN release in 2015, it has undergone a series of major upgrades to improve its ability to perform analysis of large data. Many of the underlying functions in the version 1.0 are written in C++ for high computing and memory efficiency. The package can also process all Unicode-formatted languages and is not limited to English texts.

Target group

This workshop is designed for researchers in social sciences or humanities who wish to perform quantitative analysis of textual data. Although this workshop is organized as part of the Manifesto Corpus Conference, the workshop materials also cover different types of texts

such as newspaper articles and micro-blog posts.

Participants of this workshop are expected to have basic knowledge of data manipulation in R, but not experience in quantitative text analysis or computer programming. If you are not familiar with R, see "Before the Workshop".

Goal of the Workshop

The goal of the workshop is to understand typical workflows and operations in quantitative text analysis in social scientific research using Quanteda. We will cover how to import different types textual data and transform texts into shapes that are suitable for statistical analysis. We will also apply relative frequency analysis, collocation analysis, document scaling (Wordscores/Wordfish) and topic classification (topic models).

Schedule

9.00 a.m.	Recap: Basic objects in R (vector, list, data.frame and matrix)
9.30 a.m.	Tools to import texts into R (readtext)
10.30 a.m.	<i>Coffee Break</i>
10.45 a.m.	Basic operations on Quanteda's objects (kwic, corpus, tokens, dfm)
11.45 a.m.	Combine basic operations for different types of analysis
12.30 p.m.	<i>Lunch Break</i>
1:30 p.m.	Basic statistical analysis of texts (textstat_collocation, textstat_keyness)
2.30 p.m.	Perform supervised and unsupervised machine learning (textmodels_*)
3:30 p.m.	Other text analysis tools (lda, random forest)
4:30 p.m.	End of Quanteda Workshop

Before the Workshop

If you have no experience in R, it is highly recommended to go through online materials:

- [Code School Try R](#)
- [LSE Introduction to R](#)

Please also make sure that you have the following programs installed in your laptop:

- R (version 3.4 or newer)
- RStudio

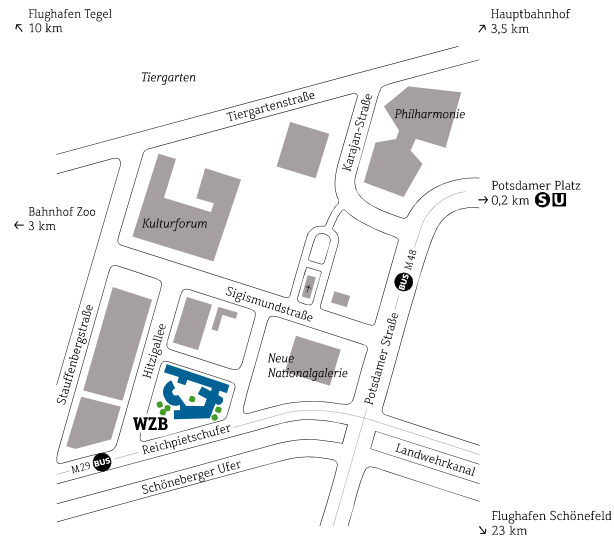
Instructors

The instructors of this workshop are Kohei Watanabe and Stefan Müller. Kohei is a post-doctoral researcher at Methodology Department of the London School of Economics and Political Science, who specializes in longitudinal analysis of news and develops core functions of Quanteda in R and C++. Stefan is a doctoral candidate in political science at Trinity College Dublin, who uses quantitative text analysis to combine the research areas of election pledge fulfilment, salience, and party positions.

If you have questions about the workshop, please do not hesitate to contact us:

k.watanabe1@lse.ac.uk (Kohei)

mullers@tcd.ie (Stefan)



WZB

Wissenschaftszentrum Berlin
für Sozialforschung

The Workshop is organized by

The DFG Project
Manifesto Research on Political Representation (MARPOR)
WZB Berlin Social Science Center

Workshop Instructors: Kohai Watanabe (London School of Economics and Political Science), Stefan Müller (Trinity College Dublin)

Workshop Organizers: Nicolas Merz

Workshop Assistance: Carolin Herrmann

Project Director: Bernhard Weßels

Registration: If you would like to participate and can commit to attending the full workshop, please send an email to manifesto-conference@wzb.eu with the following details until 31 December:

- Your name
- Research Unit ("Forschungseinheit") and Research Area ("Schwerpunkt") at the WZB / institutional affiliation
- Please state whether you are employed by the WZB/ a scholarship holder/ from another institution.
- Please state what kind of text documents you are interested in (eg. electoral programs, social media, traditional media,...)

Wissenschaftszentrum Berlin
für Sozialforschung

WZB Berlin Social Science Center

Reichpietschufer 50
10785 Berlin

Telefon +49 (30) 25491-0
Telefax +49 (30) 25491-684

wzb@wzb.eu
www.wzb.eu

Workshop
Introduction to Quantitative
Text Analysis using
Quanteda

January 31, 2018
WZB Berlin Social Science Center
Room B001