

Campus Luzern - Cooperation between Lucerne's universities

«Introduction to Text Analysis with R»

Dr. Andrea De Angelis

Tutor	Dr. Andrea De Angelis is senior researcher and lecturer at the Department of Political Science, University of Lucerne. He teaches a master seminar on Quantitative Research Design and his research interests include Political Behavior, Political Sociology, Political Communication and Computational Social Science methods.
Organization	Campus Luzern
Dates	13th December 2019, 9.00-17.00
Room	Will be announced
Language	English
Contact	For registration and questions regarding the content please contact graduate.resources@campus-luzern.ch
Contents	<p>Text Analytics is one of the fastest-growing areas of social science methodology. It is estimated that about 80% of world's data is stored as unstructured text format. Indeed, text corpora can be formed with documents from the most various sources, from customers' reviews to party manifestos, from political speeches to webpages and social media posts. Considering text as data that can be analyzed, it becomes possible to extract meaningful quantities of interest for testing new theories, or for better informing market decisions. For these reasons, applications of text analytics have quickly become standard analytical tools for social researchers and data scientists. This workshop provides a gentle introduction to text analytics based on the Bag-of-Word approach. The R programming language is used throughout the sessions and participants will familiarize with established text-mining R packages (i.e. tm, quantda, stm). This one-day workshop is divided in four blocks:</p> <ol style="list-style-type: none"> 1. General introduction to natural language processing (semantic and bag-of-word approach, building a corpus, text preprocessing, the document-feature matrix). 2. Basic forms of textual data visualization (lexical dispersion and frequency plots) 3. Text modelling (dictionaries, text scaling, statistical topic modeling, structural topic modeling).

	4. Advanced topics: elements of scraping (html, regex, working with APIs); mention of Beyond Bag-of-Words: POS tagging and Word Embeddings (Word2vec)
Prerequisites/ Materials	Basic knowledge of R programming (as for “Introduction to R for Data Analysis”) is assumed. Some prior data analysis experience is advised but not required. Workshop participants have to bring their own laptop with R and RStudio previously installed and functioning (instructions will be circulated). Students can reach the instructor with their own data for exercising during the workshop. Please contact the instructor in advance to communicate special requests and/or specific interests. Learning material (including slides, texts and documentation, exercises, cheatsheets, toy datasets) will be circulated by the instructor.
Reading list	<p>Munzert, S. et al. (2014). Automated Data Collection with R. Chichester, U.K.: Wiley.</p> <p>Jockers, M.L. (2014). Text Analysis with R for Students of Literature. new York: Springer.</p> <p>Benoit, K. et al. (2018). Quanteda: An R Package for Quantitative Analysis of Textual Data. Journal of Open Source Software, vol. 30(3):774. Link: https://joss.theoj.org/papers/40b988ba4827f8fdc07a29351c2f74b8</p> <p>Quanteda Quick-start Guide. Link: https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html.</p>
Charge	This offer of the Lucerne Campus is aimed at researchers, post-docs, doctoral students and employees interested in a dissertation of all Lucerne universities and the Graduate School@WBK and is free of charge for these persons.