

**Text Mining in R for Beginners (short course)**

Instructor: Miklós Sebők, PhD

Senior Research Fellow, Centre for Social Sciences

Venue: Corvinus University of Budapest

Date: December 5-6, 2019

E-mail: sebok.miklos@tk.mta.hu

The aim of this short course is to provide an introduction into the theory and practice of text mining in the social sciences. The course schedule comprises three 90-minute sessions which cover the basics of the theory of text as data as well as a minimalist selection of text mining techniques. The focus is on the latter: we employ the most prevalent text mining packages in R in order to solve a limited set of challenges. Course prerequisites include a working level of English and all participants should bring a laptop with the latest versions of R and R studio installed.

**SCHEDULE**

Session I

(4:00 pm - 5:30 pm, December 5, Thursday)

* Setting up R and R studio
* My first text mining project in R
* The concept of text as data
* Basic methods of text mining

Session II

(9:00 am - 10:30 am, December 6, Friday)

* Preprocessing
* The document-term matrix
* Descriptive statistics
* Dictionary-based methods
* Sentiment analysis

Session III

(10:45 am - 12:15 pm, December 6, Friday)

* Unsupervised learning
* Supervised learning

**HOMEWORK AND GRADING**

Participants are expected to perform at least four basic text analysis tasks on either a corpus provided by the instructor or compiled/procured by the participant. The tasks may include preprocessing steps, the creation of a document-term matrix, descriptive statistics, sentiment analysis, dictionary-based methods, unsupervised learning, supervised learning or any other methods covered during the course. The materials should be submitted via email in a zip/rar file containing the corpus and the code in R.

The grading schedule is as follows: High Pass, Pass, Fail.

The deadline for submission is December 17, 2019, 4:00 pm CET

Grades and brief instructor comments will be available by January 10, 2020 or before.

**LITERATURE**

|  |
| --- |
| * AGGARWAL, C. C. & ZHAI, C. 2012. *Mining text data*, Springer Science & Business Media. * BENOIT, K., WATANABE, K., NULTY, P., OBENG, A., WANG, H., LAUDERDALE, B. & LOWE, W. 2017. quanteda: Quantitative analysis of textual data. * BRADY, H. E. 2018. The challenge of big data and data science. *Annual Review of Political Science,* 22. * DENNY, M. J. & SPIRLING, A. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis,* 26**,** 168-189. * FEINERER, I. 2018. Introduction to the tm Package Text Mining in R. * GRIMMER, J. & STEWART, B. M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis,* 21**,** 267-297. * KWARTLER, T. 2017. *Text mining in practice with R*, John Wiley & Sons. * LAVER, M., BENOIT, K. & GARRY, J. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review,* 97**,** 311-331. * SILGE, J. & ROBINSON, D. 2017. *Text mining with R: A tidy approach*, " O'Reilly Media, Inc. * SLAPIN, J. B. & PROKSCH, S. O. 2008. A scaling model for estimating time‐series party positions from texts. *American Journal of Political Science,* 52**,** 705-722. * WELBERS, K., VAN ATTEVELDT, W. & BENOIT, K. 2017. Text analysis in R. *Communication Methods and Measures,* 11**,** 245-265. * WILKERSON, J. & CASAS, A. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science,* 20**,** 529-544. * YOUNG, L. & SOROKA, S. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication,* 29**,** 205-231. |