Automated Text Analysis in Political Science

- POLS 5198 – Automated Text Analysis in Political Science
- Lecturer: Martijn Schoonvelde
- Email: `mschoonvelde@gmail.com`
- Credits: 2.0
- Program: 1 Year MA Political Science, 2 Year MA Political Science
- Spring Term 2018–2019
- Dates: 6–17 May 2019
- Course requirements: familiarity with R
- Office hours: upon appointment, either through email or in person

Course introduction

Automated text analysis has become very popular across the social sciences over the last few years. With the massive availability of text data on the web, social and political scientists increasingly recognize automated text analysis (or "text as data") as a useful approach to analyzing social and political behavior. This course – in which we use R – introduces students to a variety of its methods and tools to learn about, among other things, content, ideology and sentiment in text. The course – which combines lectures and coding sessions – will be hands-on, with an emphasis on dealing with practical issues in each step of the research process (ranging from collecting and pre-processing text data to validating and visualizing output of the analysis). Students who have finished this course are well-positioned to apply automated text analysis methods in their own work, and will be able to critically evaluate existing work.

NB: This course assumes familiarity with R. Students who have not used R before will need to get themselves up to speed before the start of the course, for example by working their way through a free online R resource, like `https://www.datacamp.com/courses/free-introduction-to-r` or the resources that are listed on `https://www.rstudio.com/online-learning/#R`. Students working on their own laptops will need to have R and RStudio installed.

Learning Outcomes

This course introduces students to various approaches of automated text analysis in social science research, emphasizing hands on analysis of real (political) texts. Students will learn how to extract useful quantities of interest from text, evaluate the outcomes and write up the results of an analysis that uses automated text analysis. Furthermore, students will be able to critically evaluate (social science) research that uses automated text analysis methods.

Assessment

Students are assessed on the basis of 4 components:

1. **Attendance and participation in class (10% towards the final grade)**

- This course will involve quite some reading, some of which is technical. I expect that you come to class prepared, having read all required papers, and ready to discuss your questions, criticisms and thoughts. Furthermore, since this is a short class I expect you to attend all sessions. Missing class is only acceptable for urgent reasons and students will need to communicate this with me in advance.

2. **Two coding assignments (15% each, 30% total towards the final grade)**

   - The coding assignments (one in week 1 and one in week 2) are designed to experience the workflow of a text research project. The first assignment will concern getting from text to data that can be analyzed. The second assignment will involve applying some of the methods we discuss in class to this data. Both assignments rely on the EUSpeech dataset.

3. **Presentation of a research design (15% towards the final grade)**

   - On the last day of the course all students will give a brief presentation of a research design they have developed to address a topic they want to study using (one of) the methods discussed in class. This presentation should at least contain a research question, a discussion of the text sources, as well as the (expected) steps to address this question using the methods discussed in class. **NB**: Depending on time and enrollment, students will also act as discussant of the research design of another student with the goal of providing constructive comments to improve their work.

4. **Research note (45% towards the final grade)**

   - All students will hand in a research note of about 2000 words (excluding references and appendices) in which they briefly but clearly write up the results of a small research project based on the design they presented in class. Students are free to collect their own data or use existing data (like the EUSpeech dataset or a replication file from a published research paper). Creativity is encouraged.

     This note should contain the following elements:

     (a) Introduction & research question (±300 words): introduction to the topic.

     (b) Data & methods (±400 words): description of the data sources as well as the methods employed.

     (c) Analysis: (±1000 words): a discussion (with figures and tables) of the results of the analysis.

     (d) Conclusion (±300 words): a brief evaluation of the results and steps to push the research forward.

     Since time is short this is not likely to be a very polished research project (nor is this expected). Rather the research note is a transparent write-up of the work the student put in to address a research question of their interest using text.

Table 1: Grade Breakdown

A	94–100
A-	87.00–93.99
B+	80.00–86.99
B	73.00–79.99
B-	66.00–72.99
C+	59.00–65.99
F	0–58.99

Grading

Grading on a 100 point scale is reported in Table 1.

Course outline

** This outline serves a general plan for the course; deviations (announced) may be necessary.*

May 6: 15:30 - 17:10:

- Introduction to text as data. Introduction to EUSpeech, a dataset which will use for running examples: `https://dataverse.harvard.edu/dataverse/euspeech`
  - **Required reading**:
    - Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. and Pinker, S., (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331(6014)*, 176–182.
    - Schumacher, G., Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). EUSpeech: a new dataset of EU elite speeches. In: *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, 75–80.
    - Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annual Review of Political Science 2*0: 529– 544.

May 7: 15:30 - 17:10:

- A survey of automated text analysis in political science. Supervised and unsupervised methods. Validation, validation, validation. Text Analysis in R.
  - **Required reading**:
    - Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21(3)*, 267–297.
    - Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11(4)*, 245–265.
    - Benoit, K., Watanabe, K., Wang, H, Nulty, P., Obeng, A., Müller, & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3(30)*, 774.

May 8: 15:30 - 19:00:

- Pre-processing data. Going from text to data, including a few notes of caution. Discussion of the research design and research note.
  - **Required reading**:
    - 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(2)*, 168–189.
    - Schoonvelde, M., Schumacher, G. and Bakker, B.N., (2019). Friends with text as data benefits: assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology, 7(1)*, 124–143.

May 9: 15:30 - 17:10:

- Describing and comparing texts: readability, distinctive features, text similarity
  - **Required reading**:

    - Chapters 3 and 4 of Silge, J., & Robinson, D. (2018). Text Mining with R: A Tidy Approach. O'Reilly Media, Inc. Available at `https://www.tidytextmining.com`
    - Cross, J. & Hermansson, H., (2017). Legislative amendments and informal politics in the European Union: A text reuse approach. *European Union Politics, 18(4)*: 581–602.
    - Bischof, D. & Senninger, R., (2018). Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research, 57(2)*: 473–495.

May 10: 15:30 - 17:10:

- Using and developing dictionaries to measure sentiment, morality, populism, personality, and other things we are interested in.
  - **Required reading**:
    - Pennebaker J. & King, L. (1999) Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology, 77(6)*, 1296–1312.
    - Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29(2)*, 205–231.
    - Kraft, P. (2018). Measuring morality in political attitude expression. *Journal of Politics, 80(3)*: 1028–1033.
    - Hawkins, K. & Castanho Silva, B. (2018). Text Analysis: Big Data Approaches. In: *The Ideational Approach to Populism: Theory, Method & Analysis*, edited by Kirk A. Hawkins, Ryan Carlin, Levente Littvay, and Cristóbal Rovira Kaltwasser. London: Routledge.
    - Ramey, A. J., Klingler, J. D., & Hollibaugh, G. E. (2019). Measuring elite personality using speech. *Political Science Research and Methods, 7(1)*,163–184.
  - **19:00, Coding Assignment 1 due**

May 13: 15:30 - 17:10:

- Supervised and unsupervised methods to locate text on an underlying (political) dimension. How do they work? And how should we interpret them?
  - **Required reading**:
    - Slapin J. & Proksch S. (2008). A scaling model for estimating time-serial positions from texts. *American Journal of Political Science 52*, 705–722.
    - Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M. E., & Kurrild-Klitgaard, P. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. Research & Politics, 2(2).
    - Daniel Schwarz, Denise Traber, & Kenneth Benoit (2017). Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods 5(2):* 379–396.

May 14: 15:30 - 17:10:

- Topic models, unsupervised models for summarizing what a text is about. How do they work? And how should we interpret them?
  - **Required reading**:
    - Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4)*, 77–84.
    - Roberts, M et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58(4)*, 1064–1082.
    - Boussalis, C. & Coan, T. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*: 89–100.

May 15: 15:30 - 19:00:

- New developments in data: (i) crowd-sourcing data (ii) images as data, (iii) automated speech recognition, (iv) machine translation.
  - **Required reading**:
    - Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review, 110(2)*, 278–295.
    - Proksch, S.O., Wratil, C. and Wäckerle, J., (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, 1–21
    - De Vries, E., Schoonvelde, M. & Schumacher, G., (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis, 26(4)*, 417–430.
    - Torres, M. (2019). Give me the full picture: Using computer vision to understand visual frames and political communication. *Working paper.*
  - **13:00: Coding Assignment 2 due**

May 16: 15:30 - 17:10:

- New developments in modeling: (i) word embeddings, (ii) ltta.
- Flash talks
- Loose ends
  - Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š. & Sedlmair, M., (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12(2-3)*, 140–157.
  - Kleinberg, B., Mozes, M., & van der Vegt, I. (2018). Identifying the sentiment styles of YouTube's vloggers, EMNLP 2018.

May 17: 15:30 - 17:10:

- Research design presentations.

May 26, 17:00: **Final Assignment Due: Research Note**