

GRAD-E1282: Quantitative text analysis

Dominic Nyhuis

1. General information

Class time	September 7, 14-18 September 21, 14-18 October 5, 14-18 October 19, 14-18 November 2, 14-18 November 16, 14-18
Venue	R2.32
Instructor	Dominic Nyhuis
Instructor's office	
Instructor's e-mail	d.nyhuis@ipw.uni-hannover.de
Instructor's phone number	+49 511 762 2951
Assistant	Name: Andrea Derichs-Carlin Email: adjunctsupport@hertie-school.org Phone: +49 30 251 219 312 Room: 2.55
Instructor's Office Hours	To get in touch, please come see me after class or write me an e-mail. If you want to talk about your term paper, please be sure to send me a draft of your paper at least one week in advance.

Instructor Information:

Dominic Nyhuis is a Post-Doctoral Researcher at the Leibniz University Hannover, Chair for Comparative Politics and the Political System of Germany. Prior to this, he was affiliated with the Universities of Frankfurt, Vienna, and Mainz. He received his PhD in Political Science from the University of Mannheim. His research focuses on comparative legislative studies, party politics, and municipal politics. Methodologically, his work relies on quantitative methods, automated web data collection and text analysis.

2. Course Contents and Learning Objectives

Course contents:

Political behavior is intimately linked with the production of text. Whether it be plenary protocols, legislative proposals, election manifestos, press releases, newspapers, or one of countless other outlets – the written form is a core characteristic of political competition. Not surprisingly, the analysis of text has always been a key interest of political scholars. However, despite the status of text analyses in political research, such efforts have traditionally been restricted to resource-intensive manual designs which are often beyond the means of young scholars. The recent explosion of machine-readable text along with easy access to massive amounts of data due to the universal adoption of the

web has engendered an unprecedented interest in the automated analysis of text. The course takes the promises that the wealth of textual data holds as a point of departure to introduce participants to the most important developments in the area of text quantification from a social science perspective.

While the main focus of the course is on automated methods, we ground our discussion in techniques for the manual analysis of text. We are currently on the verge of major methodological breakthroughs such that automated methods might soon be able to incorporate syntactical features of natural languages. For the moment, however, the available automated techniques are at times too blunt for our research questions. It is thus useful to familiarize ourselves with the standards and practices of the conventional techniques for quantifying text before turning our attention to more recent developments.

The individual sessions typically have a tripartite structure that begins with a general introduction to a specific technique along with some applied examples from the current political science literature. In a second step, the practical application of the technique is introduced using the statistical programming language R. In a third step, a set of exercises are either solved individually or in small groups to ensure that participants know how to adapt the tools to their specific research problems.

In light of the promise that automated text quantification holds, there is currently quite a bit of activity in the field. The course aims to cover the most prominent methods in depth, while hinting at ongoing and potential developments to provide participants with an idea where to go from here for a specific research question.

Main learning objectives:

- Familiarity with the most common text analysis techniques in the social sciences
- Ability to apply text analysis techniques to specific research problems
- Awareness of new developments in the field of text analysis and less conventional techniques for text quantification

Target group:

The course is open to all 2nd year Master's students with an interest in empirical political science.

Teaching style:

The session will typically begin with a short lecture that introduces the basic ideas of the topics of the session along with current examples from applied research. Next, a technical introduction is provided using R. This is followed by either individual or group exercises to apply the techniques to different research problems.

Prerequisites:

Basic familiarity with R and successful attendance of Statistics II

3. Grading and Assignments

Weekly questions You should carefully work through the basic reading assignments. Each participant is required to draft three questions for the mandatory readings and upload them to the quiz in the course Moodle no later than 72 hours before the course. Each text should be addressed by at least one question. Questions might explore theoretical, conceptual, or methodological aspects of the texts.

Please do not draft “final exam” questions such as “What is the definition of XYZ?”; “What are the three components of XYZ?”

Presentations A group of students will prepare a short presentation covering the main ideas of the optional readings each week. The presentations should last no longer than 20-25 minutes. Please send your presentation to the instructor via e-mail no later than 72 hours before the course. The presentations are assigned in the first week of the course.

Term paper A term paper is written on a topic of your choice using one of the text analysis techniques that are covered in the course. Please be sure to send me a short draft of your paper via e-mail so I can give you feedback before you start writing. The paper should have a length of about 8,000 words. Be sure to observe the following formatting guidelines: Arial with font size 12, 2.5 cm margins, line spacing 1.5. Please hand in your paper in the PDF format via e-mail. The paper is due on **December 31, 2018**. A draft version of the paper is presented in the last session of the course.

Composition of Final Grade:

Weekly questions	Deadline: Weekly	Submit by: In-class	20%
Presentation	Deadline: Assigned	Submit by: In-class	35%
Term paper	Deadline: December 31, 2018	Submit by: E-mail	45%

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two sessions. For further information please consult the Examination Rules §9.

Academic Integrity: The Hertie School of Governance is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §15.

4. General Readings

N/A

5. Session Overview

Session	Session Date	Session Title
1/2	07.09.2018	Introduction to the course and advanced topics in R
3/4	21.09.2018	Manual text quantification
5/6	05.10.2018	Quantifying texts with dictionaries
7/8	19.10.2018	Unsupervised text classification
9/10	02.11.2018	Supervised text classification
11/12	16.11.2018	Text scaling

6. Course Sessions and Readings

All readings will be accessible on the Moodle course site before the semester. In case there is a change in readings, students will be notified by email. Required readings are to be read and analysed thoroughly. Optional readings are intended to broaden your knowledge in the respective area.

Session 1/2: 07.09.2018	
Introduction to the course and advanced topics in R	
Structure of the session	Introduction to the course Overview of the main issues in quantitative text analysis Very short R refresher Advanced non-statistical topics in R Individual exercises
Required Readings	Grimmer, Justin and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. <i>Political Analysis</i> , 21(3), 267-97.
Optional Readings	N/A

Session 3/4: 21.09.2018	
Manual text quantification	
Structure of the session	The basics of manual text quantification Applied research examples Crowdcoding Prominent datasets in political science Group exercise A: Practical interactions with existing datasets Group exercise B/C: Designing coding schemes
Required Readings	Benoit, William L. 2011. Content analysis in political communication. In E. Page Budy and R. Lance Holbert (eds.): <i>Sourcebook for political communication research: Methods, measures, and analytical techniques</i> . New York: Routledge, 268-82. Krippendorff, Klaus. 2004. <i>Content analysis: An introduction to its methodology</i> . Thousand Oaks: Sage, ch. 7
Optional Readings	Reinemann, Carsten and Jürgen Wilke. 2007. It's the debates, stupid! How the introduction of televised debates changed the portrayal of chancellor candidates in the German press, 1949-2005. <i>International Journal of Press/Politics</i> , 12(4), 92-111. Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver and Slava Mikhaylov. 2016. Crowd-sourced text analysis: Reproducible and

	agile production of political data. <i>American Political Science Review</i> , 110(2), 278-95.
--	--

Session 5/6: 05.10.2018

Quantifying texts with dictionaries

Structure of the session	<p>Applied research examples using dictionaries</p> <p>Inferring indicative terms</p> <p>Text processing in R and regular expressions</p> <p>Individual exercises</p>
Required Readings	<p>von Schiller, Ruben, Thomas Zittel and Matthias Henneke. 2017. The politicization of social divisions in legislative contexts: A dictionary-based analysis of parliamentary questions in the German Bundestag. <i>Political Science Research</i>, 5(1), 1-32.</p> <p>Aaldering, Loes and Rens Vliegthart. 2016. Political leaders and the media: Can we measure political leadership images in newspapers using computer-assisted content analysis? <i>Quality and Quantity</i>, 50(5), 1871-905.</p>
Optional Readings	<p>Kleinnijenhuis, Jan, Friederike Schultz and Dirk Oegema. 2015. Frame complexity and the financial crisis: A comparison of the United States, the United Kingdom, and Germany in the period 2007-2012. <i>Journal of Communication</i>, 65(1), 1-23</p> <p>Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. <i>Political Analysis</i>, 16(4), 372-403.</p>

Session 7/8: 19.10.2018

Unsupervised text classification

Structure of the session	<p>The basics of topic models</p> <p>Applied research examples</p> <p>Two frameworks for statistical text analyses in R</p> <ul style="list-style-type: none"> - The tm package - The quanteda package <p>Basic operations in statistical text analysis</p> <ul style="list-style-type: none"> - Text preparation - Term-document matrices <p>Basic metrics in quantitative text analysis</p> <p>Applying dictionaries in practice</p> <p>Topic models in practice</p>
---------------------------------	--

	<ul style="list-style-type: none"> - Latent Dirichlet allocation (LDA) - Structural topic models (STM) <p>Individual exercises</p>
Required Readings	<p>Lucas, Christopher, Richard Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. <i>Political Analysis</i>, 23(2), 254-77.</p> <p>Welbers, Kasper, Wouter van Atteveldt and Kenneth Benoit. 2017. Text analysis in R. <i>Communication Methods and Measures</i>, 11(4), 245-65.</p>
Optional Readings	<p>Wilkerson, John and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. <i>Annual Review of Political Science</i>, 20, 529-44.</p> <p>Rothschild, Jacob E., Adam J. Howat, Richard M. Shafranek and Ethan C. Busby. Forthcoming. Pigeonholing partisans: Stereotypes of party supporters and partisan polarization. <i>Political Behavior</i>.</p>

Session 9/10: 02.11.2018

Supervised text classification

Structure of the session	<p>The basics of supervised text classification</p> <p>Applied research examples</p> <p>Supervised text classification in R</p> <p>Individual exercises</p> <p>Specialized topics in text quantification</p> <ul style="list-style-type: none"> - Text reuse - Text evolution
Required Readings	<p>D’Orazio, Vito, Steven Landis, Glenn Palmer and Philip Schrod. 2014. Separating the wheat from the chaff. Applications of automated document classification using support vector machines. <i>Political Analysis</i>, 22(2), 224-42.</p> <p>Diermeier, Daniel, Jean-Francois Godbout, Bei Yu and Stefan Kaufmann. 2011. Language and ideology in Congress. <i>British Journal of Political Science</i>, 42(1), 31-55.</p>
Optional Readings	<p>Wilkerson, John, David Smith and Nicholas Stramp. 2015. Tracing the flow of policy ideas in legislatures: A text reuse approach. <i>American Journal of Political Science</i>, 59(4), 943-56.</p> <p>Baerg, Nicole Rae and Mark Hallerberg. 2016. Explaining instability in the Stability and Growth Pact: The contribution of member state power and Euroskepticism to the Euro crisis. <i>Comparative Political Studies</i>, 49(7), 968-1009.</p>

Session 11/12: 16.11.2018

Text scaling

Structure of the session	The basics of text scaling Applied research examples Scaling texts in R Individual exercises Presentation of the term papers Conclusion
Required Readings	Laver, Michael, Kenneth Benoit and John Garry. 2003. Extracting policy positions from political texts using words as data. <i>American Political Science Review</i> , 97(2), 311-31. Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. <i>American Journal of Political Science</i> , 52(3), 705-22.
Optional Readings	Hjorth, Frederik, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. <i>Research and Politics</i> . Schwarz, Daniel, Denise Traber and Kenneth Benoit. 2017. Estimating intra-party preferences: Comparing speeches to votes. <i>Political Science Research and Methods</i> , 5(2), 379-96.