Hertie School

Master of International Affairs/Master of Public Policy
Fall Semester 2020

Course Syllabus, Version 29.06.2020

GRAD-E1282: Text as Data: Quantitative Text Analysis for Political Science and Public Policy

Concentration: Policy Analysis

Dr. William Lowe

1. General information

Class time	Thursday, 16-18h
Course Format	This course uses a "flipped classroom" format and combines 50 minutes of pre-recorded material (audio or video) with a 50-minute interactive seminar. Students will use the pre-recorded material to prepare for the seminar. The seminar is taught onsite at the Hertie School, or online via the platform Clickmeeting, depending upon your location. For those attending the online seminar, Clickmeeting allows for interactive, participatory seminar style teaching.
Instructor	Dr. William E. M. Lowe
Instructor's office	3.14
Instructor's e-mail	lowe@hertie-school.org
Assistant	Email: adjunctsupport@hertie-school.org
Instructor's Office Hours	Students can make individual or group appointments by email.

Link to Module Handbook **MIA** and **MPP**
Link to **Study, Examination and Admission Rules**

Instructor Information:

Dr. William Lowe is a political methodologist with interests in text analysis, causal inference, and machine learning. He has a B.A. in Philosophy and an M.Sc. and Ph.D in Cognitive Science and Natural Language Processing, but joined the Government Department at Harvard as a postdoc and has worked in Political Science almost ever since.

Course Contents and Learning Objectives

Course contents:
This course introduces students to contemporary approaches to analyzing text as data as it is applied to policy questions.

Main learning objectives:
Familiarity with statistical and practical issues arising from textual data; the ability to fit and interpret the basic classes of text analysis models, and how to use them to complement a wider research project.

Target group:
2nd year MIA and MPP students.
The target audience for this course is interested in expanding their methodological skills to the quantitative analysis of text. There is no restriction on the substantive domain of application and students are encouraged, though not required, to come with their own substantive applications. The course mixes methodological instruction and discussion with regular practical applications. These will take the form of exercises and short reports using R.

Teaching style:
The course will consist of short pre-recorded lectures from the instructor and live discussion of these conducted either in person, remotely, or a mix of both, depending upon the location of students. There are regular exercises to be submitted the week after they are set, and there will small group projects that are presented at the end of the course.

Prerequisites:
Statistically, students should be familiar with fitting and interpreting linear models and with the basics of logistic regression. Previous exposure with any kind of measurement model or (index construction process) will be helpful, e.g. factor analysis, or IRT, but this material will be presented as needed.

Practically, students should be competent, though need not be expert, at manipulating vectors and data.frames in R. Text data is unavoidably unwieldy and much of any text analysis is spent manipulating data, which the course will provide practice for but not teach from scratch. Experience with R graphics will also be an advantage, though is not required. The Data Science Lab's help desk can suggest materials to fulfil the data manipulation prerequisites.

Diversity Statement:
Happily, contemporary R as a programming community and Text as Data as a methodological community, specifically as represented by its annual conferences and indirectly by this course, are in general much more concerned about this issues than many subfields, so we expect that those of you entering this domain for the first time will find it a welcoming experience. (If you don't, we want to hear about it). In particular, the course instructor tries to provide a diverse and inclusive engagement with the subject material and will expect you to do the same.

Grading and Assignments

Composition of Final Grade:

Assignment 1: data analysis exercise	Deadline: five times, (roughly every other week)	Submit via Moodle	Each 10% (so 50% total)
Assignment 2: group work	Deadline: last week of class	Submit via Moodle	40%
Participation grade			10%

Assignment Details

Assignment 1
Data analysis exercises are guided practical exercises based around a data set or text collection, interspersed with conceptual questions about the tools being used, interpretation of results, etc. The

exercises are designed to improve practical skills and test knowledge from lectures and reading. Grading is based on success in the practical components and high quality answers. Answers to conceptual questions are intended to require *at most* several paragraphs of text. As the course progresses we will require higher levels of presentation to each exercise as they converge on the form of the data-driven report that Assignment 2 requires as its written component.

Assignment 2

Group projects address an empirical question using text models or measures and involve 2-3 people. Groups present their work in the form of a 5 minute (yes 5 minute) presentation followed by a question and answer period, and submit a detailed two part report of their work. The first page of the report explains the question, key results, graphs, and conclusions to a non-technical decision-maker emphasizing uncertainty, implications, and limitations of the work and should stand alone.
The subsequent pages report the work in detail. Project grade is shared equally between the participants, and the instructor will not enforce collaborative arrangements or adjust grades per individual except in extraordinary circumstances, so choose your colleagues wisely.

Participation grade

The participation grade is based on the assumption that students take part, not as passive consumers of knowledge, but as active participants in the exchange, production, and critique of ideas—their own ideas and the ideas of others. Therefore, students should come to class not only having read and viewed the materials assigned for that day but also prepared to discuss the readings of the day and to contribute thoughtfully to the conversation. Participation is graded subtractively; students receive the full grade except to the extent they fail to take adequate part in the class. Participation is marked by its active nature, its consistency, and its quality, but note that it is both unnecessary and also unwise, to monopolize conversation in order to maximize participation grade. Participation that makes it harder for other class members to engage in discussion will lead to a lower grade, regardless of the quality of interventions.

Late submission of assignments: For each day the assignment is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

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 out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School 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 §16.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office *in advance* of the deadline.

2. General Readings

Readings will be provided in the form of articles, preprints, and occasionally online resources as the course progresses. There is regrettably not (yet) a single textbook that adequately treats the topics of this course.

3. Session Overview

Session	Session Date	Session Title
1	10.09.2020	Text as data
2	17.09.2020	Text as Data as Measurement
3	24.09.2020	Dictionaries (1. construction)
4	01.10.2020	Dictionaries (2. evaluation and analysis)
5	08.10.2020	Topic models (1. construction)
6	15.10.2020	Topic models (2. extensions and limitations)
Mid-term Exam Week: 19.10 - 23.10.2020 – no class		
7	29.10.2020	Space and similarity
8	05.11.2020	Sentiment
9	12.11.2020	Scaling (1)
10	19.11.2020	Scaling (2)
11	26.11.2020	Text Classification
12	03.12.2020	Causal Inference with Text
Final Exam Week: 14.12 - 18.12.2020 – no class		

4. Course Sessions and Readings

Many, but not all readings will be accessible on the Moodle course site before semester start. In the case that there is a change or addition in readings, new readings will be provided on Moodle as necessary.

Required readings are to be read and analysed thoroughly. Optional readings are intended to broaden your knowledge in the respective area and it is highly recommended to at least skim them.

Session 1: Text as data	
Learning Objective	Understand the approach and tension/complementarity with alternatives such as Discourse Analysis

Required Readings	TBA
Optional Readings	

Session 2: Text as Data as Measurement	
Learning Objective	Understand how classification and measurement models from statistics relate to text analysis
Required Readings	TBA
Optional Readings	

Session 3: Dictionaries (1. construction)	
Learning Objective	Understand the generative process assumed by dictionary-based content analysis and how this motivates the process of dictionary construction and application
Required Readings	TBA
Optional Readings	

Session 4: Dictionaries (2. evaluation and analysis)	
Learning Objective	How to use the interpret, evaluate, and improve a dictionary-based analysis
Required Readings	TBA
Optional Readings	

Session 5: Topic models (1. construction)	
Learning Objective	Understand how topic models work and how they relate to content analysis dictionaries
Required Readings	TBA
Optional Readings	

Session 6: Topic models (2. extensions and limitations)	
Learning Objective	Appreciate the strengths and limitations of topic modelling; evaluate and criticize fitted models; understand the structural topic model framework

Required Readings	TBA
Optional Readings	

Mid-term Exam Week: 19 – 23.10.2020 – no class

Session 7: Space and similarity	
Learning Objective	Understand how word embeddings, document similarity, and other 'semantic space' approaches to text analysis work and their strengths and limitations.
Required Readings	TBA
Optional Readings	

Session 8: Sentiment	
Learning Objective	Understand sentiment analysis and its applications, as a dictionary-based method or a classification model
Required Readings	TBA
Optional Readings	

Session 9: Scaling (1)	
Learning Objective	Understand the measurement basis of unidimensional scaling models from text, their strengths and limitations
Required Readings	TBA
Optional Readings	

Session 10: Scaling (2)	
Learning Objective	Understand extensions of association and correspondence analysis to multiple dimensions and their graphical interpretation
Required Readings	TBA
Optional Readings	

Session 11: Text Classification	
Learning Objective	Understand the document classification task, evaluate classification models, evaluate, and deal with errors
Required Readings	TBA

Optional Readings	

Session 12: Causal inference with text	
Learning Objective	Understand how to treat quantities from a text analysis as treatments, as confounders, and as outcomes.
Required Readings	TBA
Optional Readings	

Final Exam Week: 14 - 18.12.2020 – no class