BAMBERG
GRADUATE SCHOOL
OF SOCIAL SCIENCES

Text as Data: Social Science Inference from Language

Instructors: Brandon Stewart, Princeton University

Time: Monday, July 9, 2018: **09:00 – 17:00**

Tuesday, July 10, 2018: **09:00 – 17:00**

Wednesday, July 11, 2018: **09:00 – 12:30**

Place: Feldkirchenstr. 21, RZ/00.04

Registration: Please send a mail to Julian Hohner (julian.hohner@uni-bamberg.de) to register.

Registration Deadline: **Monday, July 4, 2018 – Maximum of 20 participants**

Description

Never before in human history has so much information been so easy to access. The promise of this wealth of information is immense, but because of its pure volume it is difficult to summarize and interpret. However, a burgeoning array of algorithms and statistical methods are beginning to make analysis of this information possible. These new forms of data and new statistical techniques provide opportunities to observe behaviour that was previously unobservable, to measure quantities of interest that were previously unmeasurable, and to test hypotheses that were previously impossible to test.

In this short course, we will explore a social science logic for how text can be included in every stage of the research process. Our goal is to describe the prevalence of a social behaviour or phenomenon and make inferences about its origins. This goal is qualitatively different from the goals that have been often used to evaluate text analytic methods, which often focus on performing a specific task. The focus on inference will push us to reconsider when and how some methods are useful, suggest new ways to evaluate methods, and will present new open questions in the use of text as data.

Unlike other courses which are built around individual techniques or software packages, this class will be built around three key tasks in the research process: discovery, measurement and causal inference. We will explore each in turn and explore how different methods facilitate each task.

Each session will be a combination of lecture and short hands-on exercises in R.

Prerequisites

No formal prerequisites. The coding sections will be more enjoyable if you have a basic fluency with R, but as long as you are able to open a console and follow along, you should be fine.

For an Introduction into the topic, please feel free to take a look into the courses “Introduction to R”, “Intermediate R” and for a deeper insight “Introduction to Machine Learning” and “Cluster Analysis in R" on DataCamp. Once you registered yourself for the workshop, we will invite you into a DataCamp Classroom in order to support your preparation for the course.

DAY 1 – Basics

09:15 - 10.45 Text as Data in Social Science: In this session we lay the framework for social science inference with text including the core principles that will guide our inferences.

11:00 – 12:30 Representing Text as Data: The basics of how to ingest text and represent it numerically using bag of words model.

12:30 – 13:30 Break

13:30 – 15:00 Discovery (Part 1): Techniques for enabling discovery in text including clustering and methods for identifying separating words.

15:15 – 16:45 Discovery (Part 2): Mixed-Membership topic models and additional approaches to discovery in text

DAY 2 – Measurement and Inference

09:15 - 10.45 Measurement (Part 1): Approaches to measurement using topic models including repurposed discovery methods and supervised learning.

11:00 – 12:30 Measurement (Part 2): Strategies based on the supervision of convenience, scaling, textual complexity, text reuse and other forms of measurement.

12:30 – 13:30 Break

13:30 – 15:00 Causal Inference: A framework for causal inference using text as outcome and text as treatment.

15:15 – 16:45 What We Didn't Cover: In 2.5 days, we can only cover so much. In this section, I will provide a brief glimpse into some of the methods there was not time for and provide pointers to where you can learn more.

DAY 3 – Discussion and Feedback

09:15 - 10.45 Student Presentations and Feedback: Students will present ongoing projects and we will discuss as a group. We can also use this opportunity to take questions.

11:00 – 12:30 An additional opportunity for student presentation and discussion.

Recommended readings

Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents" *Political Analysis*. 21, 3 267-297.