Workshop

Introduction to Quantitative Text Analysis

23-26 April 2019

Sponsoring Professors: Elias Dinas & Ellen Immergut
Instructor: Iñaki Sagarzazu

For inquiries, please contact: Miriam Boussairi

If interested in the workshop, please register Online by Thursday, 14th March.

Kindly note that due to a limited number of places, priority will be given to current EUI researchers.

Texts have become an increasingly popular source of data for research in many fields in the social and behavioral sciences, marketing, and medical sciences. The increased availability of text (in their many shapes and forms) together with advances in computational power have made a significant contribution to the big data revolution.

This workshop provides participants with a set of quantitative methods and computational techniques for acquiring, measuring, and modeling text as data. It introduces the participant to the most important concepts for analyzing textual content and presents them with different techniques – from manual to dictionary-based content analysis – within a unified measurement framework that clarifies the different substantive assumptions motivating each model. The workshop also provides participants with hands-on exercise in fitting, interpreting, and criticizing each kind of model using real-world data and practical solutions to the problem of efficiently 'harvesting' and storing data from publicly available sources.

The workshop emphasizes the question of how to integrate the results of textual analyses with other quantitative and qualitative methods and discusses how to make best use of text analysis in various research designs.

Detailed Description
The workshop begins by placing quantitative text analysis in context and demonstrating its connections to other qualitative and quantitative social science research methods. It also introduces the key mathematical ideas that guide all the subsequent models of text and takes a fresh look at the most common form of social science text analysis, i.e., manual content analysis.

After a critical examination of manual content analysis, we consider some recent, closely related methodological developments, such as semi-automated, dictionary-based content analysis and statistical topic models. We show how to build and deploy these models to quantify the content of political speeches, manifestos, legal materials, etc. We then turn to the problem of assigning large numbers of documents into categories, e.g., to determine media agendas by assigning many years of newspaper stories into different categories of news.

The workshop also offers practical solutions to problems arising from acquiring, pre-processing, and storing large numbers of texts, e.g., from government or NGO websites. These topics, often described as 'spidering' and the 'scraping' of data, are directly linked to the projects and research interests of the participants. Depending on these projects and interests, the course also addresses issues arising from non-English language materials as well as scripts, word segmentation, etc.

Detailed schedule

Date	Time	Topic	Room
Tuesday 23rd April	09.00 – 13.00	Intro to QTA ; Text as Data	Sem. Room 2
	15.00 – 17.00	Extra lab time	
Wednesday 24th April	09.00 – 13.00	Dictionary based approaches	Sem. Room 2
	15.00 – 17.00	Extra lab time	
Thursday 25th April	09.00 – 13.00	Supervised Scaling and Topic Coding	Sem. Room 4
	15.00 – 17.00	Extra lab time	
Friday 26th April	09.00 – 13.00	Unsupervised Scaling and Topic Coding	Sem. Room 2

The instructor

Iñaki Sagarzazu is Assistant Professor of Political Sciences at Texas Tech University. He is a graduate of the Simón Bolívar University as well as the University of Houston, where he earned his PhD. Prior to joining Texas Tech University, he held positions at the University of Glasgow and the

University of Oxford's Nuffield Center for Experimental Social Sciences. His research focuses on comparative politics and quantitative methods, with a special focus on statistical content analysis with applications to political communication and institutions. He has published in such leading academic journals American Journal of Political Science, British Journal of Political Science, Journal of Politics, and Political Science Research & Methods. In addition to teaching mathematics, programming, and multivariate statistics for social scientists at the University of Houston, he has previously taught courses on text analysis at Texas Tech University, the Institute of International Relations (Brazil), the ECPR Winter and Summer Schools in Methods and Techniques at the University of Bamberg and the Central University Budapest, and the IPSA Summer School at the University of São Paulo, at the National University Singapore, and at FLACSO Mexico City.

"Professor Sagarzazu has an excellent lecture style. I also like how approachable he is, always ready to answer questions and always willing to help." — participant from Singapore

"Professor Sagarzazu has a very accessible and lively way of presenting the material; he critically discusses limitations and allows time for questions. He creates a nice atmosphere and he has an intuitive way of teaching that makes difficult things seem easy. Thank you for a really good course!" — participant from Bamberg, Germany

"Overall a valuable course, learning an entirely new concept in a short amount of time. Professor was always available to assist students and went out of his way to help." — participant from Texas Tech

Prerequisites

There are no formal prerequisites. It would be beneficial if participants had some experience with the statistical software R and were familiar with basic statistical concepts. However, participants unfamiliar with these concepts and tools will be able to effectively participate in the course.

Requirements

Participants are expected to bring a WiFi-enabled laptop computer. Ahead of the course they will be provided with instructions on what free software to install and how to install it.