Computational Tools for Text Mining, Processing and Analysis

May 25th 2017, 9:00-17:00

Organizers: Dror Walter, Sijia Yang, Wouter van Atteveldt

Manual content analysis has been one of the most distinctive and influential techniques in communication research for more than half a century. With the rise of digital and social media, recent years have seen a sharp growth in the sheer amount and types of textual data communication scholars often wish to explore, as well as changes to required skillsets needed for acquiring, storing, and processing data. Due to these changes researchers in communication often find manual content analysis methods inadequate for their needs. As a result, computational approaches to text mining are becoming gradually more valuable and even necessary for contemporary communication scholars. The pre-conference workshop "Computational tools for text mining, processing and analysis" aims to engage with these computational methods.

This pre-conference offers five talks given by experts working at the frontier of computational textual analysis. The program covers both introductory materials aimed at providing less experienced scholars with practical tools for analysis, as well as in-depth critical discussions on advanced issues including assumptions, properties, inferences, triangulation with other methods, and theory development. At the concluding panel, the invited speakers, panelists and the audience will engage in a discussion about the future of computational textual analysis in communication research and social science in general. Confirmed panelists include Dr. Joseph Cappella (the Gerald R. Miller Professor of Communication at the University of Pennsylvania) and Dr. Dhavan Shah (the Louis A. & Mary E. Maier-Bascom Professor at the University of Wisconsin-Madison).

It is our hope that participants will leave this full-day workshop not only with ready-to-use tools for their day-to-day research but also with a more comprehensive understanding of these methods' assumptions, properties, theories and debates. The goal is to promote not only the usage, but a responsible usage, of computational methods for textual analysis.

Tentative Schedule:

9:00-9:15	Introduction and overview
9:15-10:15	Dr. Hai Liang: Scraping and preprocessing of social media data
10:15-10:30	Break
10:30-11:30	Dr. Molly Roberts: Structural Topic Modelling
11:30-11:45	Break
11:45-12:45	Dr. Andrew Schwartz: Machine learning on social media textual data for predicting psychological and health outcomes
12:45-1:45	Lunch break
1:45-2:45	Dr. Daniel Angus: Emerging methods for text visualization
2:45-3:00	Break
3:00-4:00	Dr. Justin Grimmer: Statistical Models for Computational textual analysis and applications
4:00-4:15	Break
4:15-5:00	Summary and roundtable: the future of computational methods in communication research

As registered participants, you get the opportunity to bring methodological challenges from your own research to our speakers. We will collect questions beforehand and share them with the speakers so that they can be best prepared. After each talk, at least 15 minutes will be devoted to facilitating discussions between speaker and participants.

Speakers:

Dr. Hai Liang: Assistant Professor in the School of Journalism and Communication, the Chinese University of Hong Kong; Experienced in teaching application of computational tools for the analysis of social media data. in His talk will cover the following:

- Tools and methods for social media data gathering and pre-processing
- The limitations and future directions of social media data gathering

Dr. Margaret Roberts: Assistant Professor in the Department of Political Science at the University of California, San Diego; the author of R's STM package for structural topic modeling (STM). Her talk will cover the following:

- What is topic modeling and what is Latent Dirchlet Allocation (LDA)?
- What distinguishes STM from LDA?
- Why STM is particularly useful for social science applications? How does STM help to estimate the relationships between topic solutions and covariates, either experimentally manipulated or observationally measured?

Dr. Andrew Schwartz: Assistant Professor in the Department of Computer Science at Stony Brook University; Lead Research Scientist for the World Well-Being Project. His talk will cover the following:

- How to use machine learning techniques to conduct large and scalable language analyses for psychological and health discovery?
- How to build prediction models from large-scale social media corpus for population-level health outcomes?
- What is the open-vocabulary approach to analyzing social media data and what insights can it reveal?

Dr. Daniel Angus: Lecturer in Computational Social Science at the University of Queensland; Received his PhD in computer science from Swinburne University of Technology; pioneered the development of the *Discursis* computer-based visual text analytic too. in His talk will cover the following:

- Current and emerging methods for text visualization
- Applying data visualization techniques to various types of corpora

Dr. Justin Grimmer: Associate Professor in Stanford University's Department of Political Science and (by courtesy) an Associate Professor in the Department of Computer Science

- How to conduct causal inference with textual data?
- A new methodology for using high-dimensional text as treatments and estimating their effects

This preconference was made possible thanks to the ICA Computational Methods Interest Group. We are especially grateful for financial support from the Annenberg School for Communication at the University of Pennsylvania and the Underwood Institute.

For more details visit: https://www.icahdq.org/mpage/PC26