Methods Short Course, University of Texas at Austin

Content Analysis in the Social Sciences
From Manual to Automated Approaches

February, 2017

Instructor: Prof. Stuart Soroka (Michael W. Traugott Collegiate Professor of Communication Studies and Political Science, University of Michigan, ssoroka@umich.edu, snsoroka.com)

Outline: The increasing ready availability of digital texts has led to an explosion of interest in content analytic methods. This has been true for those interested in the content of news programming, in traditional newspapers, television transcripts, and online news. It is true for entertainment media researchers, now able to analyze years of scripts from movies and film, and from video games. Social media analysts have ready access to miles of data, particularly from Twitter. And that same content has been of interest to scholars interested in predicting election outcomes, macro-economic trends, and stock prices; and for those interested in the nature of legislative debate and political party positioning.

Content analysis has been around for a long time, however; and many of the core ideas developed 40 years ago are equally important today. The objective of this short course is to offer a theoretically-informed introduction to large-scale content-analytic methods for the social sciences (albeit with an emphasis on political communication). We draw on past work on human-coded methods; and focus on the application recently-developed automated approaches focused on word frequencies and co-occurrences, dictionary-based approaches, and machine learning.

Classes combine lectures, discussion, and computer-based work. Students can watch analyses conducted by the instructor, but you are also encouraged to bring laptops to run analyses yourselves as well. All the scripts used on-screen will be provided to students. Analysis will be done in R, so you will need R (r.com), RStudio (rstudio.com), and Lexicoder (lexicoder.com) installed on your computers. This software is all free, and multi-platform. Given limited class time, I will not be provided technical help in class – you should make sure that R is properly installed, and functioning, before class. If you are not already familiar with R, then I strongly recommend a basic primer (such as Quick-R at statmethods.net). Lexicoder and other content-analytic tools will also require that you have an updated version of Java installed (java.com).

By the end of the class, you will be familiar with a range of automated content analytic approaches, and, if you've brought a computer with the necessary software, you'll have conducted automated analyses of large-scale corpuses in R.

Readings: Given limited time in class, I strongly recommend just a little background reading before class. These readings will make a very big difference:

1. Klaus Krippendorff. 1989. "Content analysis." Pp. 403-7 in the *International Encyclopedia of Communications*, Erik Barnouw et al., eds., Oxford: Oxford University Press.
2. H. Andrew Schwartz and Lyle H. Unger. 2015. "Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods." *American Academic of Political and Social Science* 659: 78-94.
3. Justin Grimmer and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*.
4. Kenneth Benoit and Alexander Herzog. 2015. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words." In *Analytics, Policy and Governance*, eds. Jennifer Bachner, Kathyrn Wagner Hill, and Benjamin Ginsberg.
5. Lori Young and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts," *Political Communication* 29: 205-231.

All readings can be provided electronically to students.

Schedule: The course is set up as four 1.5-hour blocks, two on each day (with some breaks), as follows:

Session 1. Introduction, and Building a Corpus. This session introduces some central ideas and objectives in content analysis, and then focuses on issues of sampling and pre-processing, as they relate to both human and automated approaches.

Session 2. Exploring Frequent, Discriminating, and Co-occurring Words. This session offers a first foray in computer-automated content analysis in R, focused on approaches that look at individual words. We look at word counts, word clouds, and some basic clustering methods that explore connections between words.

Session 3. Reliability and Validity in Human Coding and Dictionary-Based Approaches. This session focuses on approaches that rely on human and automated – both dictionary-based and machine-learning approaches – to capturing frames, or topics, or the tone of text.

Session 4. Supervised Learning, Unsupervised Learning, and Topic Modeling. This session offers and introduction to supervised and unsupervised learning approaches, and a preliminary application of LDA topic modeling.