1H Computer-Aided Text Analysis

Essex Summer School Course Details

9–20 July 2012

Kenneth Benoit
Methodology Institute
London School of Economics and Political Science
kbenoit@lse.ac.uk

June 27, 2012

Short Outline

The course is intended to survey and characterize methods for systematically extracting information from text for social scientific purposes, starting with classical content analysis methods and proceeding forward to state of the art scaling methods for estimating quantities from text using statistical methods. The course lays a theoretical foundation for text analysis but mainly takes a very practical and applied approach, so that students learn how to apply these methods in actual research. It takes as a starting point more traditional methods of content analysis, but is aimed at the most recent advances in quantitative content analysis that treat words as data to be analysed using statistical tools. The course surveys several of these methods but also applies the statistical framework to more traditional non-automated coding schemes such as the Comparative Manifesto Project and the Policy Agendas Project. It is also designed to cover many fundamental issues such as inter-coder agreement, reliability, validation, accuracy, and precision. Lessons will consist of a mixture of theoretical grounding in content analysis approaches and techniques, with hands-on analysis of real texts using content analytic and statistical software.

Prior Knowledge

Ideally, students in this course will have prior knowledge in the following areas:

- A basic understanding of probability and statistics at the level of an introductory postgraduate social science course. Understanding of regression analysis is presumed;
- The ability to learn to use text analysis software (on a demonstration basis) such as QDAMinder/Wordstat (or if preferred, MaxQDA). Most of the applications and exercised will be performed in QDAMiner/Wordstat. No prior knowledge is assumed, but students should be willing and able to learn their basic use (with guided tutorials). As these are both user-friendly packages and both are available in limited demonstration versions that can be freely downloaded, this should not be too difficult.
- The ability to manipulate text files using a text editor. (It does not matter which text editor you use, but you should use plain text editor – e.g. TextEdit, Notepad, Emacs, BBEdit, etc -âĂŤ and not a word processor such as Microsoft Word.)
- Familiarity with a statistical package such as Stata or (ideally) R. In a pinch, a spreadsheet could be used but a statistical package is greatly preferred, and instructional examples will

use Stata and R. Note that these packages will be required only for the last two sessions (the advanced quantitative topics).

Detailed Outline

Meetings

Classes will meet for ten sessions. Approximately 2/3 of the time will be devoted to lectures, and the other half will consist of "lab" sessions where we will work through exercises in class.

Computer Software

Computer-based exercises will feature prominently in the course, especially in the lab sessions. The use of all software tools will be explained in the sessions, including how to download and install them.

Grading

Grading will be based on a combination of four take-home exercises assigned during the 10-day course, as well as a take-home final exam.

Recommended Texts

There is no really good single textbook that exists to cover computerized or quantitative text analysis. While not ideally fitting our core purpose, Krippendorf's classic *Content Analysis* — just updated — is the next best thing. The staple book-length reading is therefore:

- Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 3rd edition.

Another good general reference to content analysis that you might find useful as a supplement is:

- Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA.

Other readings will consist of articles, reproduced in the coursepack (and if possible, available as downloadable pdf files from the course web page).

Short Course Schedule

Day	Date	Topic(s)	Details
Mon	9 July	Introduction and Issues in text analysis	Course goals; logistics; software overview; conceptual foundations; content analysis; objectives; examples.
Tue	10 July	Textual Data, Sampling, and Working with a Text Corpus	Where to obtain textual data; formatting and working with text files; indexing and meta-data; sampling concerns with textual data.
Wed	11 July	Descriptive inference from text	Methods of summarizing texts and features of texts in order to characterize their properties. It covers many basic quantitative textual measures.
Thu	12 July	Research Design issues in textual studies	Reliability and validity and their role in designing and evaluating content-analysis based research; measures of reliability.
Fri	13 July	Thematic analysis, key words in context	Computer-assisted methods for developing themes from texts, examining key words in context, applying codes to texts.
Mon	16 July	Classical quantitative content analysis	Manual unitization and coding approaches, including the CMP, Policy Agendas Project, and self-constructed themes. The exercise will consist of an on-line quantitative coding experiment.
Tue	17 July	Automated dictionary-based approaches	Dictionary construction, and methods for automatically indexing texts for compiling scales of substantive quantities of interest.
Wed	18 July	Word-scoring for automatic dictionary construction	Automatic "word indexing" and scoring using "Wordscores"; scaling models using algorithmic and probability-based scales.
Thurs	19 July	Classifiers: Introduction to the Naive Bayes Classifier	Extends "wordscores" into classification and scaling.
Fri	20 July	Document Scaling: Parametric models	Continues text scaling using completely automated methods based on parametric (Poisson) scaling.

Detailed Course Schedule

Day 1: Introduction to Quantitative Text Analysis

This topic will introduce the goals of the course, the logistics of the course, provide an overview of the topics to be covered, and preview the software to be used. It will also introduce traditional (non-computer assisted) content analysis and distinguish this from computer-assisted methods and quantitative text analysis. We will cover the conceptual foundations of content analysis and quantitative content analysis, discuss the objectives, the approach to knowledge, and the particular view of texts when performing quantitative analysis. Two examples will be discussed (based on the Gebauer et. al. and Schonhardt-Bailey readings).

Required Reading:

Krippendorff (2013, Ch. 1–2)
Roberts (2000)

Recommended Reading:

(example) Gebauer et al. (2007)
(example) Schonhardt-Bailey (2008)
Roberts (1989)
Neuendorf (2002, Chs. 1–3)

Lab session:

Exercise 1: Working with Text

Day 2: Textual Data, Sampling, and Working with Texts

Topics to be covered include the organization of textual data and how to work with text files. It will also cover concerns with sampling texts in research designs and how to choosing and observing units.

Required Reading:

Krippendorff (2013, Chs. 3–4)

Recommended Reading:

Neuendorf (2002, Ch. 4–7)
Benoit et al. (2009)

Lab session:

Exercise 2: Working with Text II

Day 3: Descriptive Inference from Text

This topic covers methods of summarizing texts and features of texts in order to characterize their properties. It covers many basic quantitative textual measures.

Required Reading:

Krippendorff (2013, Chs. 6, 10)

Recommended Reading:

Neuendorf (2002, Ch. 4–7)
Benoit et al. (2009)

Lab session:

Exercise 3: Descriptive summaries of texts

Day 4: Research Design issues in textual studies

Here we focus on two key research design issues central to any systematic text-based analysis: reliability and validity, goals which tend to tradeoff with one another. This topic thoroughly discusses both concepts and discusses their role in designing and evaluating content-analysis based research. This section also covers several key measures of reliability and agreement from a mathematical standpoint.

Required Reading:

Krippendorff (2013, Chs. 4–5, 12, 13)
Daübler et. al. (2012)

Recommended Reading:

Klingemann et al. (2006, Appendixes I–II)
Banerjee et al. (1999)

Lab session:

Exercise 4: Anatomy of a coding scheme.

Day 5: Thematic analysis

Thematic analysis as developed here will involve systematic exploration of the texts to extract themes, locate key words and colloations, explore how key words are used in context, and mark up sections of text after having identified themes.

Required Reading:

Krippendorff (2013, Review Chs. 9, 11)
Klingemann et al. (2006, skim but esp. Introduction, Appendixes I–II)

Recommended Reading:

Neuendorf (2002, Chs. 6–7)

Recommended Reading:

Exercise 5: Thematic analysis, coding, and KWIC using QDAMiner.

Day 6: Classical Quantitative Content Analysis

Classic (quantitative) content analysis involves the development of coding schemes, the conversion of texts into discrete units and the assignment of codes to each unit based on the coding scheme. This topic covers manual unitization and coding approaches, including the construction of coding frames and different schemes for unitizing texts. It examines two widely used schemes in political science: the Comparative Manifesto Project and the Policy Agendas Project. User-friendly software packages (e.g. MaxQDA) for applying coding frames will be used for this topic.

Required Reading:

Krippendorff (2013, Review Chs. 4–5, Read Ch. 7)
Klingemann et al. (2006, skim but esp. Introduction, Appendixes I–II)

Recommended Reading:

Neuendorf (2002, Chs. 6–7)

Recommended Reading:

Exercise 6: Applying a coding scheme.

Day 7: Automated dictionary-based approaches

Automatic dictionary-based methods involve association of pre-defined word lists with particular quantitative values assigned by the researcher for some characteristic of interest. Here will we will learn methods for constructing dictionaries as well as several methods for using computerized tools to apply the dictionaries to texts. We will also covers a variety of statistical issues surrounding text types, tokens, and equivalencies, including stemming, lemmatization, and trimming of texts based on word frequencies and *tf-idf*.

Required Reading:

Neuendorf (2002, Chs. 6)
Laver and Garry (2000)

Recommended Reading:

Mikhaylov et al. (2010)

Assignment:

Exercise 7: Applying dictionary coding using QDAMiner.

Day 8: Text Scaling Models – from Dictionaries to "Word-scoring"

This topic introduces methods for placing documents on continuous dimensions or 'scales'. This topic introduces the major methods for scaling documents and discusses their similarities and differences to other scaling models such as factor analysis and ideal point analysis, and discusses the situations where scaling methods are appropriate. The focus here is on the algorithmic method knows as "Wordscores" that applies a probability model of words given texts that can be used to estimate their characteristics along a latent dimension.

Required Reading:

Laver, Benoit and Garry (2003)
Lowe (2008)

Recommended Reading:

Martin and Vanberg (2007)
Benoit and Laver (2007)

Assignment:

Exercise 8: Wordscoring political texts. (Requires Stata or R)

Day 9: Classifers: An Introduction to the Naive Bayes Classifier

Classification methods permit the automatic classification of texts in a test set following machine learning from a training set. This topic introduces classifiers and explains one of the most popular classifiers, the Naive Bayes model. We will show how this method is a generalization of Wordscores and can be used to scale documents in the same method as Wordscores.

Required Reading:

Manning, Raghavan, and Schütze (2009, Chapter 13)
Statsoft, "Naive Bayes Classifier Introductory Overview," http://www.statsoft.com/textbook/naive-bayes-classifier/.
Bionicspirit.com, 9 Feb 2012, "How to Build a Naive Bayes Classifier," http://bionicspirit.com/blog/2012/02/09/howto-build-naive-bayes-classifier.html.

Recommended Reading:

An online article by Paul Graham on classifying spam e-mail. http://www.paulgraham.com/spam.html.

Assignment:

Exercise 9: Classifying movie reviews. Uses QDAMiner/Wordstat to classify textual data from http://www.cs.cornell.edu/People/pabo/movie-review-data/.

Day 10: Parametric Models for Text Scaling

This session continues text scaling using completely automated methods, based on parametric (Poisson) scaling, and contrasts these methods to other alternatives, critically examining the assumptions such models rely upon.

Required Reading:

Slapin and Proksch (2008)
Lowe and Benoit (2011)

Recommended Reading:

Clinton et al. (2004)

Assignment:

Exercise 10: Using "Wordfish" to scale documents. (Requires R.)

References

Alexa, M. and Zuell, C. (2000b). Text analysis software: commonalities, differences and limitations: the results of a review. Quantity and Quality, 34:299–321.

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 27(1):3–23.

Bara, J., Weale, A., and Biquelet, A. (2007). Analysing parliamentary debate with computer assistance. Swiss Political Science Review, 13(4):577–605.

Benoit, K. and Laver, M. (2003). Extracting policy positions from political texts using phrases as data: A research note. Paper presented the 2003 annual meeting of the Midwest Political Science Association, Palmer House Hilton and Towers, Chicago, IL, 3–6 April.

Benoit, K., Laver, M., and Mikhaylov, S. (2009). "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2, April): 495-513

Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call voting: A unified approach. American Journal of Political Science, 98(2):355–370.

Daübler, Thomas, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012 (forthcoming). "Natural Sentences as Valid Units for Coded Political Texts." *British Journal of Political Science*.

Gebauer, J., Tang, Y., and Baimai, C. (2007). User requirements of mobile technology: Results from a content analysis of user reviews. Information Systems and E-Business Management.

Hilliard, D., Purpura, S. J., and Wilkerson, S. (2006). Computer assisted topic classification for mixed methods social science research. Journal of Information Technology and Politics, 4(4).

Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., and McDonald, M. (2006). Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003. Oxford University Press, Oxford.

Krippendorff, K. (2013). Content Analysis: An Introduction to Its Methodology. Sage, Thousand Oaks, CA, 3rd edition.

Laver, M., Benoit, K., and Garry, J. (2003). Estimating the policy positions of political actors using words as data. *American Political Science Review*, 97(2):311–331.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. American Journal of Political Science, 44(3):619–634.

Lowe, W. (2008). Understanding Wordscores. Political Analysis, 16(4).

William Lowe and Kenneth Benoit. "Estimating Uncertainty in Quantitative Text Analysis." Paper prepared for the 2011 Midwest Political Science Association. Version: 30 March 2011.

Martin, L. W. and Vanberg, G. (2007). A robust transformation procedure for interpreting political text. Political Analysis, 16(1):93–100.

McIntosh, W., Evans, M., Lin, J., and Cates, C. (2007). Recounting the courts? applying automated content analysis to enhance empirical legal research. Journal of Empirical Legal Studies, 4(4):1041–1057.

2012. Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. “Coder Reliability and Misclassification in the Human Coding of Party Manifestos.” *Political Analysis* 20(1): 78âĂŞ91.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich SchüÌĹtze. 2009. *An Introduction to Information Retrieval.* Cambridge Univerisity Press. Online at *http://nlp.stanford.edu/IR-book/*.

Monroe, B. and Maeda, K. (2004). Talk's cheap: Text-based estimation of rhetorical ideal-points. POLMETH Working Paper.

Neuendorf, K. A. (2002). The Content Analysis Guidebook. Sage, Thousand Oaks, CA.

Pennebaker, J. W. and Chung, C. K. (2008). Computerized text analysis of al-Qaeda transcripts. In Krippendorf, K. and Bock, M. A., editors, The Content Analysis Reader. Sage.

Proksch, S.-O. and Slapin, J. (2008). Position-taking in european parliament speeches. Paper presented at the Annual Meeting of the Midwest Political Science Association, March 2008.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. Quality and Quantity, 34(3):259–274.

Schonhardt-Bailey, C. (2005). Measuring ideas more effectively: An analysis of Bush and Kerry's national security speeches. PS: Political Science and Politics, 38.

Schonhardt-Bailey, C. (2008). The congressional debate on partial-birth abortion: Constitutional gravitas and moral passion. British Journal of Political Science, 38:383–410.

Slapin, J. and Proksch, S.-O. (2008). A scaling model for estimating time series policy positions from texts. American Journal of Political Science, 52(8).

Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying party affiliation from political speech. Journal of Information Technology and Politics, 5(1):33–48.