2F Quantitative Text Analysis

22 July–2 August 2013

http://www.kenbenoit.net/essex2013qta

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

July 22, 2013

Short Outline

The course surveys methods for systematically extracting quantitative information from text for social scientific purposes, starting with classical content analysis and dictionary-based methods, to classification methods, and state-of-the-art scaling methods and topic models for estimating quantities from text using statistical techniques. 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. The common focus across all methods is that they can be reduced to a three-step process: first, identifying texts and units of texts for analysis; second, extracting from the texts quantitatively measured features—such as coded content categories, word counts, word types, dictionary counts, or parts of speech—and converting these into a quantitative matrix; and third, using quantitative or statistical methods to analyse this matrix in order to generate inferences about the texts or their authors. The course systematically surveys these methods in a logical progression, with a very practical hands-on approach where each technique will be applied in lab sessions using appropriate software, on real texts.

Objectives

This course is aimed to provide a practical foundation to and a working knowledge of the main applied techniques of quantitative text analysis for social science research. The course covers many fundamental issues in quantitative text analysis such as inter-coder agreement, reliability, validation, accuracy, and precision. It also surveys the main techniques such as human coding (classical content analysis), dictionary approaches, classification methods, and scaling models. It also includes systematic consideration of published applications and examples of these methods, from a variety of disciplinary and applied fields, including political science, economics, sociology, media and communications, marketing, finance, social policy, and health policy. 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.

Prerequisites

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

- An understanding of probability and statistics at the level of an intermediate postgraduate social science course. Understanding of regression analysis is presumed. This course is not heavily mathematical or statistical but students without the prerequisite level of quantitative experience will find the second week (in particular) difficult to follow. However, it will be possible to apply all of the methods covered using the WordStat software in all but the the last two sessions of the course, even if the students fail to grasp the full statistical workings of each method.
- Willingness and ability to use the WordStat/QDAMiner software, a commercial package developed by Provalis Research. This software will be used for all but the last two lessons, although the R library (see next item) may also be used for this purpose.
- Familiarity with the R statistical package. Stata may also be used but the lab sessions will be designed to use R coupled with a customized R library designed by the instructor. This is in development and available from `http://github.com/kbenoit/quanteda`.

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.

Teaching Assistant

The teaching assistant for this course will be Dr. Paul Nulty, p.nulty@lse.ac.uk, who will lead the computer labs and contribute also to some of the lectures.

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	22 July	Introduction and Issues in text analysis	Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.
Tue	23 July	Forms of textual data, units of analysis, and definitions of features	Where to obtain textual data; formatting and working with text files; indexing and meta-data; units of analysis; and definitions of features and measures commonly extracted from texts, including stemming, stop-words, and feature weighting.
Wed	24 July	Research Strategies in Quantitative Text Analysis	This section addresses the main challenges involved in quantitative text research, reliability, measurement validity, scaling approaches and basic diagnostics, and sampling strategies.
Thu	25 July	Quantitative methods for comparing texts	Quantitative methods for comparing texts, such as characterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; dissimilarity measurse; and an in-depth discussion of text types, tokens, and equivalencies.
Fri	26 July	Quantitative content analysis	How to convert text into quantitative matrixes for analysis using human coding; reliability checks; how to develop and test a coding scheme.
Mon	29 July	Dictionary methods	How to convert text into quantitative matrixes using dictionary approaches, including guidelines for constructing, testing, and refining dictionaries. Covers commonly used dictionaries such as LIWC, RID, and the Harvard IV-4, with applications.
Tue	30 July	Document classification	Statistical methods for classifying documents into categories, the nature of category systems, and special issues arising from using words as data. The topic also introduces validation and reporting methods for classifiers and discusses where these methods are applicable.
Wed	31 July	Non-parametric scaling models for text	The "Wordscores" approach to scaling latent traits using a Naïve Bayes foundation; Correspondence Analysis applied to texts.
Thurs	1 August	Parametric scaling models for text	Poisson scaling models (aka "wordfish") of latent word and document traits, and their applications.
Fri	2 August	Working with Big Text Data: Twitter	An introduction to applying quantitative text analysis to Twitter Data, with examples.

Detailed Course Schedule

Day 1: Introduction and Issues in quantitative text analysis

This topic will introduce the goals and 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)
Grimmer and Stewart (2013)

Recommended Reading:

(example) Gebauer, Tang and Baimai (2008)
(example) Schonhardt-Bailey (2008)
Roberts (2000)
Neuendorf (2002, Chs. 1–3)

Lab session:

Exercise 1: Working with Texts in QDAMiner

Day 2: Textual Data, Units of Analysis, Definitions of Features

Textual data comes in many forms. Here we discuss those formats, and talk about text processing preparation of texts. These issues include where to obtain textual data; formatting and working with text files; indexing and meta-data; units of analysis; and definitions of features and measures commonly extracted from texts, including stemming, stop-words, and feature weighting.

Required Reading:

Krippendorff (2013, Ch. 3–4)
Jivani (2011)
http://en.wikipedia.org/wiki/Stop_words (and browse lists of stop words)
Manning, Raghavan and Schütze (2008, 117–120)

Recommended Reading:

Wikipedia entry on Character encoding, http://en.wikipedia.org/wiki/Text_encoding
Browse the different text file formats at http://www.fileinfo.com/filetypes/text
Neuendorf (2002, Chs. 4–7)

Lab session:

Exercise 2: Extracting features from texts

Day 3: Research Strategies in Quantitative Text Analysis

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. We also discuss broad categories of scaling approaches, distinguising supervised from unsupervised "learning" methods.

Required Reading:

Krippendorff (2013, Chs. 6, 10)
Däubler et al. (2012)

Recommended Reading:

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

Lab session:

Exercise 3: Descriptive summaries of texts

Day 4: Quantitative methods for comparing texts

Quantitative methods for comparing texts, such as char- acterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; dissimilarity measures; and an in-depth discussion of text types, tokens, and equivalencies.

Required Reading:

Krippendorff (2013, Chs. 4–5, 12, 13)
Choi, Cha and Tappert (2010)
Lowe et al. (2011)

Recommended Reading:

DuBay (2004)

Lab session:

Exercise 4: TBA.

Day 5: Quantitative Content Analysis

Classical 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. We will use user-friendly software (QDAMiner/WordStat) for coding, annotating, and summarizing texts.

Required Reading:

Krippendorff (2013, Review Chs. 4–5, Read Ch. 7)
Klingemann et al. (2006, skim but esp. Introduction, Appendixes I–II)
Mikhaylov, Laver and Benoit (2012)

Recommended Reading:

Benoit, Laver and Mikhaylov (2009)
Neuendorf (2002, Chs. 6–7)

Lab Session:

Exercise 5: Applying a coding scheme.

Day 6: 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. This topic covers the design model behind dictionary construction, including guidelines for testing and refining dictionaries. Hand-on work will cover commonly used dictionaries such as LIWC, RID, and the Harvard IV-4, with applications. We will also review a variety of text pre-processing issues and textual data concepts such as word types, tokens, and equivalencies, including word stemming and trimming of words based on term and/or document frequency.

Required Reading:

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

Recommended Reading:

Pennebaker and Chung (2008)

Assignment:

Exercise 6: Applying dictionary coding using QDAMiner.

Day 7: Document classification and introduction to machine learning

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. It covers in depth the nature of category systems, methods for assessing classifer performance, the effects of feature weighting on classification accuracy. The topic also covers validation and reporting methods for classifiers and discusses where these methods are applicable.

Required Reading:

Manning, Raghavan and Schütze (2008, Ch. 13)
Evans et al. (2007)
Statsoft, "Naive Bayes Classifier Introductory Overview," http://www.statsoft.com/textbook/naive-bayes-classifier/.

Recommended Reading:

An online article by Paul Graham on classifying spam e-mail. http://www.paulgraham.com/spam.html.
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.
Yu, Kaufmann and Diermeier (2008)

Assignment:

Exercise 7: Classifying movie reviews and court briefs. Uses QDAMiner/Wordstat to classify textual data from http://www.cs.cornell.edu/People/pabo/movie-review-data/ and from Evans et al. (2007).

Day 8: Non-parametric scaling models for text

This topic introduces methods for placing documents on continuous dimensions or "scales", introducing the major non-parametric methods for scaling documents and discusses the situations where scaling methods are appropriate. Building on the Naive Bayes classifier, we introduce the "Wordscores" method of Laver, Benoit and Garry (2003) and show the link between classificaiton and scaling. We also discusses the similarities and differences to other non-parametric scaling models such as correspondence analysis.

Required Reading:

Laver, Benoit and Garry (2003)
Benoit and Nulty (2013.)

Recommended Reading:

Martin and Vanberg (2007)
Benoit and Laver (2007)
Lowe (2008)

Assignment:

Exercise 8: Wordscoring political texts (using R).

Day 9: Parametric Models for Text Scaling

This session continues text scaling using unsupervised scaling methods, based on parametric approaches modelling features as Bernoulli or Poisson distributed, and contrasts these methods to other alternatives, critically examining the assumptions such models rely upon.

Required Reading:

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

Recommended Reading:

Clinton, Jackman and Rivers (2004)

Assignment:

Exercise 9: Using "Wordfish" to scale documents. (Requires R.)

Day 10: Working with Big Text Data: Twitter

Social media such as micro-blogging site Twitter provide a wealth of spontaneous, distributed, real-time text that can be used to analyze almost any topic. We introduce the growing literature applying text analysis techniques to this form of data, with examples for measuring sentiment, networks, and locational information.

Required Reading:

Ginsberg et al. (2008)
Metaxas, Mustafaraj and Gayo-Avello (2011)
Barberá (2013)

Recommended Reading:

Lampos, Preotiuc-Pietro and Cohn (2013)

Assignment:

Exercise 10: Using Twitter to model a zombie outbreak. (Requires R.) See Barbera's TwitterR website.

References

References

Banerjee, M., M. Capozzoli, L. McSweeney and D. Sinha. 1999. "Beyond Kappa: A Review of Inter-rater Agreement Measures." *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 27(1):3–23.

Barberá, Pablo. 2013. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." http://files.nyu.edu/pba220/public/birds_jan2013.pdf.

Benoit, K. and M. Laver. 2007. "Compared to What? A Comment on 'A Robust Transformation Procedure for Interpreting Political Text' by Martin and Vanberg." *Political Analysis* 16(1):101–111.

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

Benoit, Kenneth and Paul Nulty. 2013. "Classification Methods for Scaling Latent Political Traits." Presented at the Annual Meeting of the Midwest Political Science Association, April 11–14, Chicago.

Choi, Seung-Seok, Sung-Hyuk Cha and Charles C. Tappert. 2010. "A Survey of Binary Similarity and Distance Measures." *Journal of Systemics, Cybernetics and Informatics* 8(1):43–48.

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

Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2012. "Natural Sentences as Valid Units for Coded Political Texts." *British Journal of Political Science* 42(4):937–951.

DuBay, William. 2004. *The Principles of Readability*. Costa Mesa, California. http://www.impact-information.com/impactinfo/readability02.pdf: Impact Information.

Evans, Michael, Wayne McIntosh, Jimmy Lin and Cynthia Cates. 2007. "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research." *Journal of Empirical Legal Studies* 4(4, December):1007–1039.

Gebauer, Judith, Ya Tang and Chaiwat Baimai. 2008. "User requirements of mobile technology: results from a content analysis of user reviews." *Information Systems and e-Business Management* 6(4):361–384.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski and Larry Brilliant. 2008. "Detecting influenza epidemics using search engine query data." *Nature* 457(7232):1012–1014.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* (In Press).

Jivani, Anjali Ganesh. 2011. "A Comparative Study of Stemming Algorithms." *Int. J. Comp. Tech. Appl.* 2(6, Nov-Dec):1930–1938.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge and Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford: Oxford University Press.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: Sage.

Lampos, Vasileios, Daniel Preotiuc-Pietro and Trevor Cohn. 2013. A user-centric model of voting intention from Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

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

Lowe, W. 2008. "Understanding Wordscores." *Political Analysis* 16(4):356–371.

Lowe, William and Kenneth Benoit. Forthcoming. "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark." *Political Analysis* .

Lowe, William, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling Policy Preferences From Coded Political Texts." *Legislative Studies Quarterly* .

Manning, C. D., P. Raghavan and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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

Metaxas, Panagiotis T., Eni Mustafaraj and Daniel Gayo-Avello. 2011. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*.

Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20(1):78–91.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks CA: Sage.

Pennebaker, J. W. and C. K. Chung. 2008. Computerized text analysis of al-Qaeda transcripts. In *The Content Analysis Reader*, ed. K. Krippendorf and M. A. Bock. Thousand Oaks, CA: Sage.

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

Schonhardt-Bailey, Cheryl. 2008. "The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion." *British Journal of Political Science* 38:383–410.

Slapin, J. B. and S.-O. Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.

Yu, B., S. Kaufmann and D. Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology and Politics* 5(1):33–48.