2F Quantitative Text Analysis

21 July–1 August 2014

http://www.kenbenoit.net/essex2014qta

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

July 3, 2014

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

The course is also designed to cover many fundamental issues in quantitative text analysis such as inter-coder agreement, reliability, validation, accuracy, and precision. It focuses on methods of converting texts into quantitative matrixes of features, and then analysing those features using statistical methods. The course briefly covers the qualitative technique of human coding and annotation (classical content analysis), but the main focus is on more automated approaches. These automated approaches include dictionary construction and application, classification and machine learning, scaling models, and topic models. For each topic, we will systematically cover 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. This year we will be working primarily in R, using the `quanteda` package.

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	21 July	Quantitative text analysis overview and fundamentals	Course goals; logistics; software overview; conceptual foundations; quantitative text analysis as a field; objectives; examples.
Tue	22 July	Working with texts, defining documents and features, weighting	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; identifying collocations.
Wed	23 July	Descriptive statistical methods for textual analysis	Quantitative methods for describing texts, such as characterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; and an in-depth discussion of text types, tokens, and equivalencies.
Thu	24 July	Quantitative methods for comparing texts	Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models.
Fri	25 July	Automated 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.
Mon	28 July	Document classifiers	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.
Tue	29 July	Unsupervised models for scaling texts	The "Wordscores" approach to scaling latent traits using a Naïve Bayes foundation; Correspondence Analysis applied to texts.
Wed	30 July	Supervised models for scaling texts	Poisson scaling models (aka "wordfish") of latent word and document traits, and their applications.
Thurs	31 August	Clustering methods and topic models	An introduction to hierarchical clustering for textual data, including parametric topic models such as Latent Dirichlet Allocation (LDA)
Fri	1 August	Mining Social Media: An application to textual analysis of Twitter data.	Methods for extracting text and meta-data from Twitter feeds and applying sentiment analysis to these feeds.

Detailed Course Schedule

Day 1: Quantitative text analysis overview and fundamentals

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. We will also work through some published examples.

Required Reading:

Krippendorff (2013, Ch. 1–2)
Grimmer and Stewart (2013)

Recommended Reading:

Roberts (2000)
Neuendorf (2002, Chs. 1–3)

Lab session:

Exercise 1: Working with Texts in quanteda

Day 2: Textual Data, Units of Analysis, Definitions of Documents and 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. 5, 7)
Jivani (2011)
http://en.wikipedia.org/wiki/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)
Krippendorff (2013, Ch. 6) CHECK

Lab session:

Exercise 2: Extracting features from texts

Day 3: Descriptive statistical methods for textual analysis

Here we focus on quantitative methods for describing texts, focusing on summary measures that highlight particular characteristics of documents and allowing these to be compared. These methods include characterizing texts through concordances, co-occurrences, and keywords in context; complexity and readability measures; and an in-depth discussion of text types, tokens, and equivalencies.

Required Reading:

Krippendorff (2013, Chs. 9–10)
Dunning (1993)
Däubler et al. (2012)

Recommended Reading:

DuBay (2004)

Lab session:

Exercise 3: Descriptive summaries of texts

Day 4: Quantitative methods for comparing texts

Quantitative methods for comparing texts, through concordances and keyword identification, dissimilarity measures, association models, and vector-space models.

Required Reading:

Krippendorff (2013, Ch. 10)
Choi, Cha and Tappert (2010)
Lowe et al. (2011)
Manning, Raghavan and Schütze (2008, Section 6.3)

Recommended Reading:

DuBay (2004)

Lab session:

Exercise 4: TBA.

Day 5: Automated dictionary methods

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)
Rooduijn and Pauwels (2011)

Recommended Reading:

Pennebaker and Chung (2008)
Loughran and McDonald (2011)

Assignment:

Exercise 5: Applying dictionary coding using QDAMiner.

Day 6: Document classifiers

Classification methods permit the automatic classification of texts in a test set following machine learning from a training set. We will introduce machine learning methods for classifying documents, including one of the most popular classifiers, the Naive Bayes model, as well as k-nearest neighbour and Support Vector Machines (SVMs). The topic also introduces 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 7: Unsupervised Models for Scaling Texts

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 classification 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 (2008)
Lowe (2008)

Assignment:

Exercise 8: Wordscoring political texts (using R).

Day 8: Supervised Models for Scaling Texts

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 (2013)

Recommended Reading:

Clinton, Jackman and Rivers (2004)

Assignment:

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

Day 9: Clustering methods and topic models

An introduction to hierarchical clustering for textual data, including parametric topic models such as Latent Dirichlet Allocation (LDA).

Required Reading:

Blei (2012)
Blei, Ng and Jordan (2003)
Manning, Raghavan and Schütze (2008, Ch. 16–17)
Beil, Ester and Xu (2002)

Recommended Reading:

Chang et al. (2009)

Assignment:

Exercise 9: Using LDA to estimate document topics.

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 analyze sentiment.

References

References

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.

Beil, F, M Ester and X Xu. 2002. Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD* . . . .

Benoit, K. and M. Laver. 2008. "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 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.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4, April):77.

Blei, D.M., A.Y. Ng and M.I. Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

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.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence." *Computational linguistics* .

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.

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* 21(3):267–297.

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

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.

Loughran, Tim and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66(1, February):35–65.

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

Lowe, William and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Lowe, William, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling Policy Preferences From Coded Political Texts." *Legislative Studies Quarterly* 26(1, Feb):123–155.

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)*.

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.

Rooduijn, Matthijs and Teun Pauwels. 2011. “Measuring Populism: Comparing Two Methods of Content Analysis.” *West European Politics* 34(6, November):1272–1283.

Slapin, Jonathan B. and Sven-Oliver 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.