The Quantitative Analysis of Textual Data

Autumn 2014

http://www.kenbenoit.net/nyu2014qta
Meets: Tuesdays (see dates) 10:00–11:50, Room 217

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

Version: September 11, 2014

Short Outline

The course surveys methods for systematically extracting quantitative information from political 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 practical, hands-on approach where each technique will be applied using appropriate software to 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 but only for the purposes of creating a validation set for 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 but focusing on political science. 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

Students in this course will have prior knowledge in the following areas:

- An intermediate to advanced understanding of probability and statistics.
- Familiarity with the R statistical package. All methods will be implemented in R, using primarily the `quanteda` R package available from `http://github.com/kbenoit/quanteda`.
- A desire to learn methods on the cutting edge in several disciplines.

Detailed Outline

Meetings

Classes will meet for eight sessions. Lessons will consist of two-hour lectures followed by supervised problem sets to be completed outside of class. These will involve computer exercises applied to texts supplied by the instructor.

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, although I am currently writing one (*The Quantitative Analysis of Textual Data*). While not ideally fitting our core purpose, Krippendorf's classic *Content Analysis* — just updated — is a good primer for manual methods of content analysis and coverage of some of the same fundamentals faced in quantitative text analysis.

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

Other readings will consist of articles (which I will make available as pdf files).

Short Course Schedule

Day	Date	Topic(s)	Details
Tues	16 Sept	Course overview and introduction to software	We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the approaches to be covered. We will also discuss and demonstrate the software.
Tues	23 Sept	Quantitative text analysis overview and fundamentals, defining documents and textual features	Conceptual foundations; 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.
Tues	30 Sept	Descriptive statistical methods for textual analysis	Quantitative methods for describing texts, such as characterizing texts through concordances, co-occurrences, and keywords in context; identifying collocations; complexity and readability measures; and an in-depth discussion of text types, tokens, and equivalencies.
Tues	14 Oct	Quantitative methods for comparing texts	Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models; "keyness" association with labels or classes.
Tues	21 Oct	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.
Tues	28 Oct	Document classifiers and supervised scaling models.	Statistical methods for classifying documents into categories, the nature of category systems, and special issues arising from using words as data; the "Wordscores" approach to scaling latent traits using a Naïve Bayes foundation. The topic also introduces validation and reporting methods for classifiers and discusses where these methods are applicable.
Tues	4 Nov	Unsupervised models for scaling texts	Correspondence analysis; Poisson scaling models (aka "wordfish") of latent word and document traits, and their applications.
Tues	18 Nov	Clustering methods and topic models	Topic extraction clustering for textual data, including nonparametric models based on principal components methods, and the parametric Latent Dirichlet Allocation (LDA) model.
Tues	2 Dec	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

Session 0: Introduction

This topic will introduce the goals and logistics of the course, provide an overview of the topics to be covered, and the software to be used. Since text analysis courses typically students diverse in their prior experience, applied fields, programming expertise, and statistical knowledge, this session allows us to get a feel for the class and what to expect. (It also helps me pitch the level of the remaining sessions.)

Required Reading:

Vignette and instructions at http://github.com/kbenoit/quanteda
Grimmer and Stewart (2013)

Session 1: Quantitative text analysis overview and fundamentals

This session will cover fundamentals, including the continuum from traditional (non-computer assisted) content analysis to fully automated 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 discuss issues including 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, and stop-words.

Required Reading:

Krippendorff (2013, Ch. 1–2, 5, 7)
Grimmer and Stewart (2013)
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)

Exercise:

Working with Texts in quanteda

Session 2: 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. We will also discuss weighting strategies for features, such as *tf-idf*.

Required Reading:

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

Recommended Reading:

DuBay (2004)

Exercise

Selecting, weighting, and summarizing texts and their features.

Session 3: 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)

Exercise

Comparing texts and their features.

Session 4: 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)

Exercise

Applying, modifying, and creating dictionaries for the analysis of political texts.

Session 5: Document classifiers and supervised scaling models.

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. The topic also introduces validation and reporting methods for classifiers and discusses where these methods are applicable. Building on the Naive Bayes classifier, we introduce the "Wordscores" method of Laver, Benoit and Garry (2003) for scaling latent traits, and show the link between classification and scaling.

Required Reading:

Manning, Raghavan and Schütze (2008, Ch. 13)
Evans et al. (2007)
Laver, Benoit and Garry (2003)
Benoit and Nulty (2013.)

Recommended Reading:

Statsoft, "Naive Bayes Classifier Introductory Overview," http://www.statsoft.com/textbook/naive-bayes-classifier/.
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)
Martin and Vanberg (2007)
Benoit and Laver (2008)
Lowe (2008)

Exercise:

Classifying legal documents and legislative speeches.

Session 6: Unsupervised 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. We also cover non-parametric methods such as correspondence analysis and discuss the similarity to parametric (Poisson-scaling) models.

Required Reading:

Slapin and Proksch (2008)
Lowe and Benoit (2013)

Recommended Reading:

Clinton, Jackman and Rivers (2004)

Exercise:

Using "Wordfish" and correspondence analysis to scale documents.

Session 7: Clustering methods and topic models

Topic extraction clustering for textual data, including nonparametric models based on principal components methods, and the parametric Latent Dirichlet Alloca- tion (LDA) model.

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)

Exercise:

Using LDA to estimate document topics in political party programmes.

Session 8: Working with Social Media 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)

Exercise

Using Twitter to analyze sentiment in political blogs.

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.

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.

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.