Quantitative Text Analysis

Kenneth Benoit
Department of Political Science
Trinity College Dublin
kbenoit@tcd.ie

Version: February 4, 2016

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 covers 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 must have completed at least one prior course in quantitative methods.
Students in this course will strongly benefit from prior experience with the R statistical package. All methods will be implemented in R, using primarily the (instructor's) R package `quanteda` available from `http://github.com/kbenoit/quanteda` and from CRAN.

Grading

Formative coursework

Exercises from the computer classes will be submitted for marking. All weeks will be marked, although the first two weeks will be basically marked 100% for participation. The marks from these problem sets will form 60% of the course grade.

Project

(TO BE CONFIRMED)
A final project of 3,000 words will be due at the end of ST, and form 40% of the course grade. This will be an original analysis of texts using some of the methods covered in class, and may focus on replicating or extending a published work. Additional guidelines will be issued in week 2.

Detailed Outline

Teaching

Lectures will meet for four sessions, consisting of about 2.5 hours of lectures, a break, and then some supervised "lab" sessions. Lab sessions will consist of supervised problem sets, with some questions to be completed outside of class. These will involve computer exercises applied to texts supplied by the instructor. These will be marked to provide 60% of the course grade.

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. 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 (slowly) writing one, entitled (*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 and book excerpts, which I will make available on Moodle as pdf files.

Detailed Course Schedule

5 Feb: 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:

- Vignette and instructions at http://github.com/kbenoit/quanteda
- Grimmer and Stewart (2013)
- Manning, Raghavan and Schütze (2008, 117–120)
- Krippendorff (2013, Chs. 9–10)
- Dunning (1993)

Recommended Reading:

- Krippendorff (2013, Ch. 1–2, 5, 7)
- 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)
- Däubler et al. (2012)

Exercise: Getting started with `quanteda`

Working with Texts in `quanteda`.

12 Feb: Quantitative methods for comparing texts

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*. The emphasis will be on comparing texts, through concordances and keyword identification, dissimilarity measures, association models, and vector-space models.

Required Reading:

- Krippendorff (2013, Ch. 10)
- Lowe et al. (2011)
- Manning, Raghavan and Schütze (2008, Section 6.3)

Recommended Reading:

- Seale, Ziebland and Charteris-Black (2006)

Exercise

Comparing texts and their features.

11 Mar: 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)
- Tausczik and Pennebaker (2010)
- Loughran and McDonald (2011)

Exercise

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

8 Feb: Machine Learning for Texts.

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. We also cover applications of penalized regression to score and scale texts.

Required Reading:

- Manning, Raghavan and Schütze (2008, Ch. 13)
- Lantz (2013, Ch. 3–4)
- Evans et al. (2007)
- Laver, Benoit and Garry (2003)

Recommended Reading:

- Lantz (2013, Ch. 10)
- 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)
- Zumel and Mount (2014, Ch. 5–6)
- Benoit and Nulty (2013.)
- Martin and Vanberg (2007)
- Benoit and Laver (2008)
- Lowe (2008)

Exercise:

Classifying legal documents and legislative speeches.

References

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.

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.

Dunning, Ted. 1993. "Accurate methods for the statistics of surprise and coincidence." *Computational linguistics* 19:61–74.

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):1007–1039.

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.

Lantz, Brett. 2013. *Machine Learning with R*. Packt Publishing Ltd.

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):35–65.

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

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

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):1272–1283.

Seale, Clive, Sue Ziebland and Jonathan Charteris-Black. 2006. "Gender, cancer experience and internet use: A comparative keyword analysis of interviews and online cancer support groups." *Social Science & Medicine* 62(10):2577–2590.

Tausczik, Y R and James W Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1):24–54.

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

Zumel, Nina and John Mount. 2014. *Practical Data Science with R*. Manning Publications.