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 meth-
ods, 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 ma-
trix; 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 an-
notation 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.


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

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Topic(s)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tues</td>
<td>16 Sept</td>
<td>Course overview and introduction to software</td>
<td>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.</td>
</tr>
<tr>
<td>Tues</td>
<td>23 Sept</td>
<td>Quantitative text analysis overview and fundamentals, defining documents and textual features</td>
<td>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.</td>
</tr>
<tr>
<td>Tues</td>
<td>30 Sept</td>
<td>Descriptive statistical methods for textual analysis</td>
<td>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.</td>
</tr>
<tr>
<td>Tues</td>
<td>14 Oct</td>
<td>Quantitative methods for comparing texts</td>
<td>Quantitative methods for comparing texts, such as keyword identification, dissimilarity measures, association models, vector space models; “keyness” association with labels or classes.</td>
</tr>
<tr>
<td>Tues</td>
<td>21 Oct</td>
<td>Automated dictionary methods</td>
<td>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.</td>
</tr>
<tr>
<td>Tues</td>
<td>28 Oct</td>
<td>Document classifiers and supervised scaling models</td>
<td>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.</td>
</tr>
<tr>
<td>Tues</td>
<td>4 Nov</td>
<td>Unsupervised models for scaling texts</td>
<td>Correspondence analysis; Poisson scaling models (aka “wordfish”) of latent word and document traits, and their applications.</td>
</tr>
<tr>
<td>Tues</td>
<td>18 Nov</td>
<td>Clustering methods and topic models</td>
<td>Topic extraction clustering for textual data, including nonparametric models based on principal components methods, and the parametric Latent Dirichlet Allocation (LDA) model.</td>
</tr>
<tr>
<td>Tues</td>
<td>2 Dec</td>
<td>Mining Social Media: An application to textual analysis of Twitter data.</td>
<td>Methods for extracting text and meta-data from Twitter feeds and applying sentiment analysis to these feeds.</td>
</tr>
</tbody>
</table>
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:

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, dis-
similarity 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 dic-
tionaries. 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:
An online article by Paul Graham on classifying spam e-mail. http://www.paulgraham.com/spam.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 Allocation (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


