Text Analysis for Applied Social Science

Molly Roberts

USCD Political Science

May 7, 2015

1 Overview

Statistical analysis of text data has become increasingly common in the social sciences (Grimmer and Stewart, 2013). Applications can be found in political science, economics, sociology, and psychology, for example. In this week long workshop we introduce scholars to the necessary tools for doing text analysis in a rigorous, replicable, way. We cover both pragmatic aspects but also cover the statistical details of workhorse text analysis models.

2 Day 1: Introduction to Text Analysis

This introduction will introduce an overview of text analysis as a methodology. It will begin to introduce text and the basics of text processing necessary to use these tools. This unit will cover:

- Text analysis: an introduction (Monroe and Schrodt, 2008)
- Overview of approaches text analysis methods (Grimmer and Stewart, 2013)
- Collection of text-scraping software (Jackman, 2006; Pilgrim, 2000) (the latter at `http://www.diveintopython.net/`)
- Introduction to basic stemming and lemmatization `http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html`

Assigned Reading

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* p. mps028.

Jackman, Simon. 2006. "Data from the Web into R." *The Political Methodologist* 14(2):11–15.

Monroe, Burt L and Philip A Schrodt. 2008. "Introduction to the special issue: The statistical analysis of political text." *Political Analysis* 16(4):351–355.

Pilgrim, Mark. 2000. "Dive Into Python.".

3 Day 2: Word Counts and Basic Text Manipulations

This unit will also discuss using word counts for text data. It will introduce software to count words and software to identify discriminating words. This unit will cover:

- Yoshikoder: software for word counts `http://www.yoshikoder.org/courses/apsa2006/apsa-yk.pdf`
- Word counts in action (Young and Soroka, 2012)
- The pitfalls of word counts (Loughran and McDonald, 2011)
- Multinomial Inverse Regression (Taddy, 2013)

Assigned Reading

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.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503):755–770.

Young, Lori and Stuart Soroka. 2012. "Affective news: The automated coding of sentiment in political texts." *Political Communication* 29(2):205–231.

4 Day 3: Supervised Text Methods

This unit will focus on supervised methods for text analysis. Supervised methods leverage some form of human training or guidance which is then used directly in the analysis of textual data. We will cover the statistical foundations of the models and describe their use. This unit will cover:

- ReadMe (Hopkins and King, 2010)
- Classifying political parties from speech (Yu, Kaufmann and Diermeier, 2008)
- Classifiers and ensembles (Hillard, Purpura and Wilkerson, 2008)
- RTextTools (Jurka et al., 2011)

Assigned Reading

Hillard, Dustin, Stephen Purpura and John Wilkerson. 2008. "Computer-assisted topic classification for mixed-methods social science research." *Journal of Information Technology & Politics* 4(4):31–46.

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

Jurka, Timothy P, Loren Collingwood, Amber E Boydstun, Emiliano Grossman and Wouter van Atteveldt. 2011. “Rtexttools: A supervised learning package for text classification.”.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. “Classifying party affiliation from political speech.” *Journal of Information Technology & Politics* 5(1):33–48.

5 Day 4: Unsupervised Text Methods

This unit will focus on unsupervised methods for text analysis. Unsupervised methods leverage use statistical tools to discover common patterns in textual data, which then require human interpretation and validation. We will cover the statistical foundations of the models and describe their use. This unit will cover:

- Introduction to inference for latent variable models (Bishop et al., 2006, Chapter 1)
- Latent Dirichelet Allocation (Blei, Ng and Jordan, 2003)
- Structural Topic Models (Roberts et al., 2014, 2013)
- Clustering (Grimmer and King, 2011)

Assigned Reading

Bishop, Christopher M et al. 2006. *Pattern recognition and machine learning*. Vol. 1 springer New York.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. “Latent dirichlet allocation.” *the Journal of machine Learning research* 3:993–1022.

Grimmer, Justin and Gary King. 2011. “General purpose computer-assisted clustering and conceptualization.” *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Roberts, Margaret, Brandon Stewart, Dustin Tingley and Edoardo Airoldi. 2013. “The structural topic model and applied social science.” *Neural Information Processing Society (peer reviewed conference paper)* .

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58(4):1064–1082.

6 Day 5: New Applications

This final day will cover applications of the text analysis methods described above to interesting social science questions.

- Reverse engineering censorship in China (King, Pan and Roberts, 2013, 2014)

- Text analysis for comparative politics (Lucas et al., 2015)
- Measuring political communication (Grimmer, 2010)
- Measuring anti-Americanism (Jamal et al., 2014)

Assigned Reading

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1–35.

Jamal, Amaney, Robert Keohane, David Romney and Dustin Tingley. 2014. "Anti-Americanism or Anti-Interventionism? Evidence from the Arabic Twitter Universe." *Perspectives on Politics* .

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(02):326–343.

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer and Dustin Tingley. 2015. "Computer assisted text analysis for comparative politics." *Political Analysis* 23:254–277.

7 Statistical Packages

Throughout the course we will leverage several statistical packages that we or others have contributed to the open source community. These packages will be helpful for students wishing to complete optional workshops. These include:

- `Python` for web-sraping
- `Python` package BeautifulSoup
- `Yoshikoder` for word counts `http://www.yoshikoder.org/`
- `R` package `textir` for multinomial inverse regression
- `R` package `ReadMe`
- `R` package `RTextTools` for classifiers
- `R` package implements the Structural Topic Model (Roberts, Stewart and Tingley, Submitted) (available at www.structuraltopicmodel.com)

Assigned Reading

Roberts, Margaret, Brandon Stewart and Dustin Tingley. Submitted. "stm: R package for Structural Topic Models." *Working paper* .