Introduction to Text Analysis

① Motivation

② Conceptual Overview

③ Extended Example

④ Additional Resources
Text is important for politics and political science

- Most of politics passes through language at some point
- When we have language, text is often a useful abstraction
- Most political data sets are derived from text
- Most qualitative work is text-based
Motivation

Quantitative text analysis methods can assist both quantitative and qualitative modes of inquiry

• Systematizes rote tasks that a human can do
  ▶ Equivalent to thousands of cheap RAs with moderate, known error rates (force multiplier)
  ▶ My interest in text analysis stems from experience as a human RA (with moderate, unknown error rates)

• Makes helpful proposals for further reading, coding, etc

• Anyone who says it is magic is selling you something
  ▶ Not a magic method for generating insights
  ▶ Not a substitute for reading or deep knowledge
Motivation

I’ve used it on almost every project, even “non-text” projects

- Book
  - Measured Jihadism using cleric fatwas
  - Explored topics in a Jihadi corpus
  - Characterized the population of Muslim clerics on Wikipedia
  - Improved speed of qualitative coding of CV data

- Classified NYT articles as being about human rights abuses or not (ISQ 2013)

- Searched for all instances of praise for ratification in US and EU press releases (ISQ 2015)
② Conceptual Overview

No math.
Conceptual Overview

Four Principles of Quantitative Text Analysis (Grimmer and Stewart, 2013)

(1) All quantitative models of language are wrong—but some are useful.
(2) Quantitative methods for text amplify resources and augment humans.
(3) There is no globally best method for automated text analysis.
(4) Validate, Validate, Validate.

My additions:

(5) All text analysis is dimension reduction.
(6) There is no free lunch.
   (supervised methods require front-end work, unsupervised require back-end work)
2 Conceptual Overview

Analyzing massive collections of text has been essentially impossible for all but the most well-funded projects. We show how automated content methods can make possible the previously impossible in political science: the systematic analysis of large-scale text collections without massive funding support. Across all subfields of political science, scholars have developed or imported methods that facilitate substantively important inferences about politics from large text collections. We provide a guide to this exciting area of research, identify common misconceptions and errors, and offer guidelines on how to use text methods for social scientific research.

We emphasize that the complexity of language implies that automated content analysis methods will never replace careful and close reading of texts. Rather, the methods that we profile here are best thought of as amplifying and augmenting careful reading and thoughtful analysis. Further, automated content methods are incorrect models of language. This means that the performance of any one method on a new data set cannot be guaranteed, and therefore validation is essential when applying automated content methods. We describe best practice validations across diverse research objectives and models.

Before proceeding we provide a road map for our tour. Figure 1 provides a visual overview of automated content analysis methods and outlines the process of moving from collecting texts to applying statistical methods. This process begins at the top left of Fig. 1, where the texts are initially collected. The burst of interest in automated content methods is partly due to the proliferation of easy-to-obtain electronic texts. In Section 3, we describe document collections which political scientists have successfully used for automated content analysis and identify methods for efficiently collecting new texts.

With these texts, we overview methods that accomplish two broad tasks: classification and scaling. Classification organizes texts into a set of categories. Sometimes researchers know the categories beforehand. In this case, automated methods can minimize the amount of labor needed to classify documents. Dictionary methods, for example, use the frequency of key words to determine a document’s class (Section 5.1). But applying dictionaries outside the domain for which they were developed can lead to serious errors. One way to improve upon dictionaries are...
Extended Example
4. Additional Resources

Textbooks:
- Tibshirani, Hastie, and Friedman. *ESL* (link)

Papers:
- Text as Data (link)
- Computer Assisted Text Analysis for Comparative Politics (link)
- Work by Kevin Quinn, Arthur Spirling, Molly Roberts, Justin Grimmer, Gary King, Ben Lauderdale, Hanna Wallach, others.