Intro to Text Analysis

Yongjun Zhang, Ph.D.

Department of Sociology and IACS

Stony Brook University SUNY

Today's Agenda

- The Moment of Text as Data
- A Brief Discussion of Your Textual Data
- Lab Training on Collecting and Processing Textual Data

The Moment of Text as Data

Data Sources for Textual Data?

- Newspaper Data (e.g.,Nexis Uni or ProQuest)
- Government Records (e.g.,congressional hearings, speeches, gov websites)
- Corporate Statements (e.g., statement, press release)
- Social Media (e.g., Twitter, Facebook)
- Other Online Platforms (e.g., Reddit, Github,PrePrints, WebofScience)
- Archive Data and Existing Corpora (e.g.,historical documents, undigitized texts)
- More (*Give me some examples*). . . . . .

But Big Data Is Not About Data!

"The revolution is not about the data. It's about the analytics that we can come up with and that we now have to be able to understand what these data say."

Gary King,Harvard University

Then The Big Question Is. . .

What kind of toolkit do we have for text Analysis as social scientists?

- Give me some examples of methods you have heard of. . .

Advance Access publication January 22, 2013

Political Analysis (2013) 21:267–297
doi:10.1093/pan/mps028

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Justin Grimmer
Department of Political Science, Stanford University, Encina Hall West 616 Serra Street, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu (corresponding author)

Brandon M. Stewart
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: bstewart@fas.harvard.edu

Edited by R. Michael Alvarez

Advance Access publication February 16, 2009

Political Analysis (2008) 16:372–403
doi:10.1093/pan/mpn018

Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict

Burt L. Monroe
Department of Political Science, Quantitative Social Science Initiative, The Pennsylvania State University, e-mail: burtmonroe@psu.edu (corresponding author)

Michael P. Colaresi
Department of Political Science, Michigan State University, e-mail: colaresi@msu.edu

Kevin M. Quinn
Department of Government and Institute for Quantitative Social Science, Harvard University, e-mail: kevin_quinn@harvard.edu

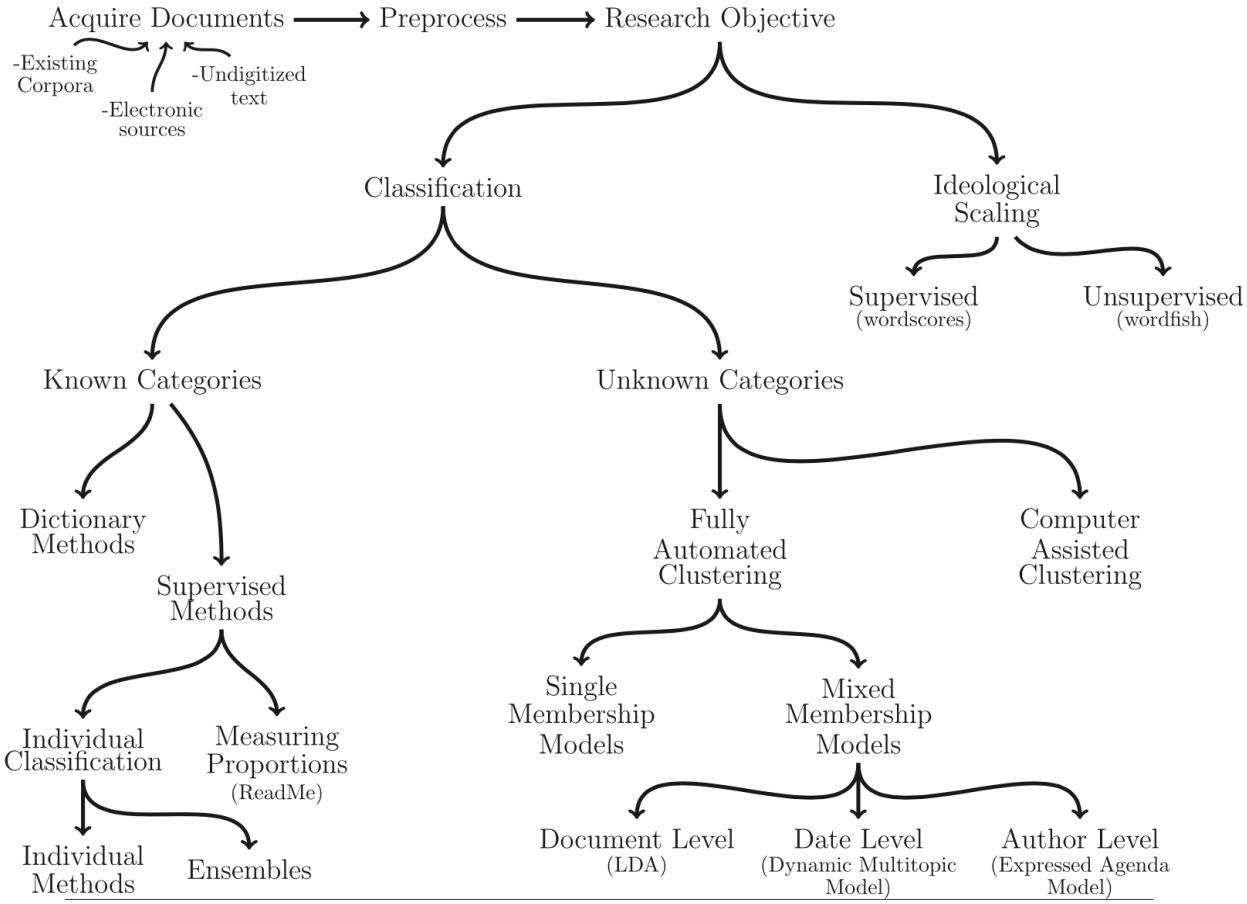


Fig. 1 An overview of text as data methods.

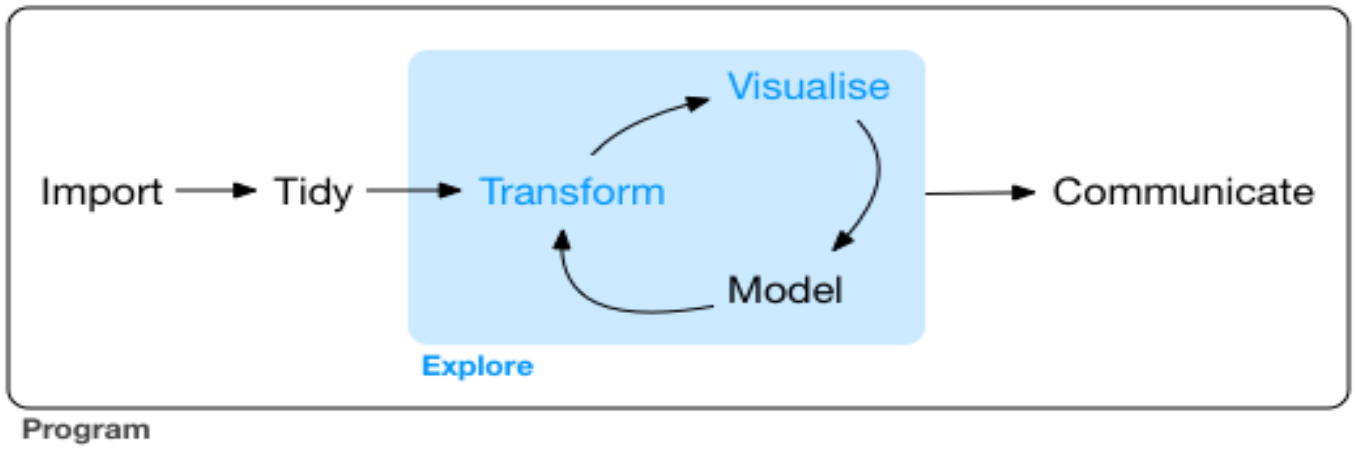


Figure 1: Hadley Wickham and Garrett Grolemund 2016

How to Represent Text in NLP

bag-of-words

The intuition of the classifier is shown in Fig. 4.1. We represent a text document as if it were a **bag-of-words**, that is, an unordered set of words with their position ignored, keeping only their frequency in the document. In the example in the figure, instead of representing the word order in all the phrases like "I love this movie" and "I would recommend it", we simply note that the word *I* occurred 5 times in the entire excerpt, the word *it* 6 times, the words *love*, *recommend*, and *movie* once, and so on.

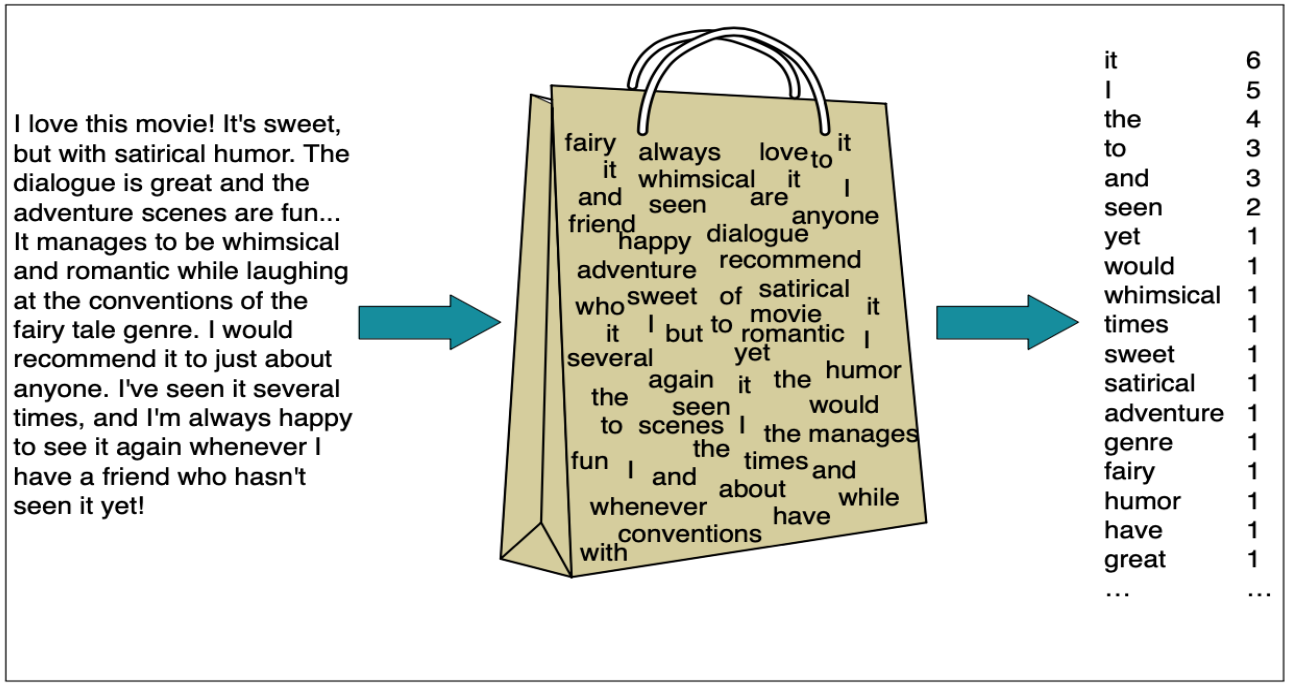


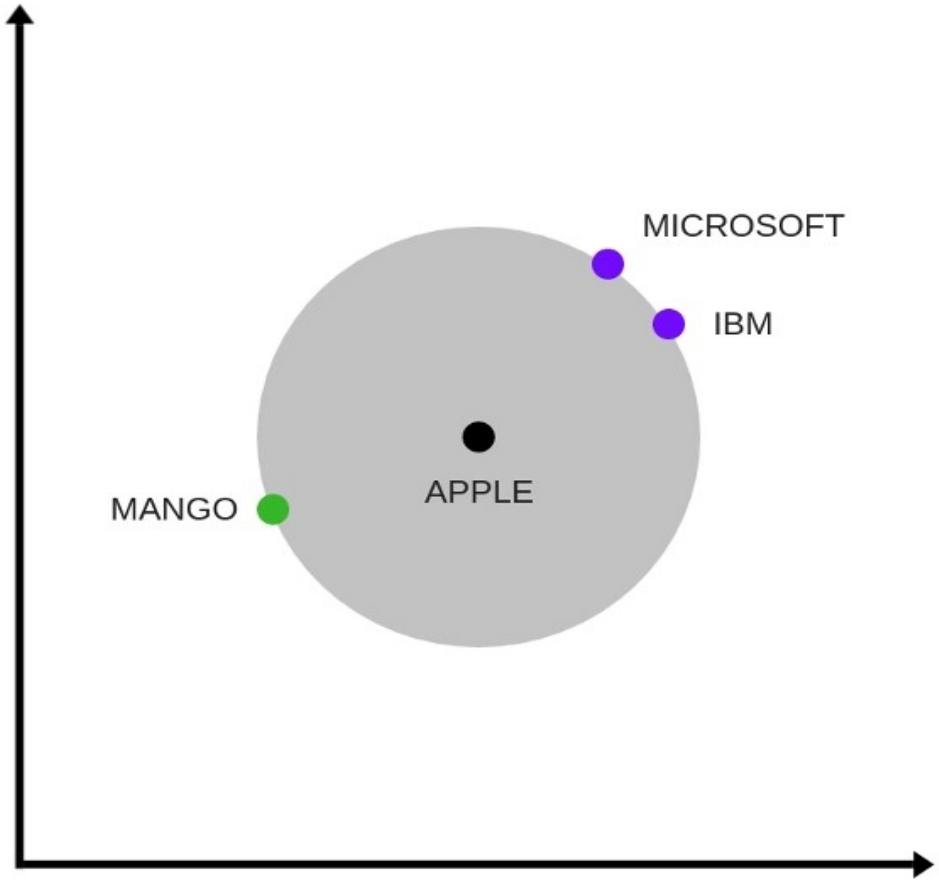
Figure 4.1 Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.

Figure 2: Dan Jurafsky and James Martin 2019

All Models Are Wrong, But Some Are Useful.

But How to Represent Meaning of Words

- Word Embedding (e.g.,Word2vec, GloVe)



How to Preprocess Textual Data (We will cover this in Lab)

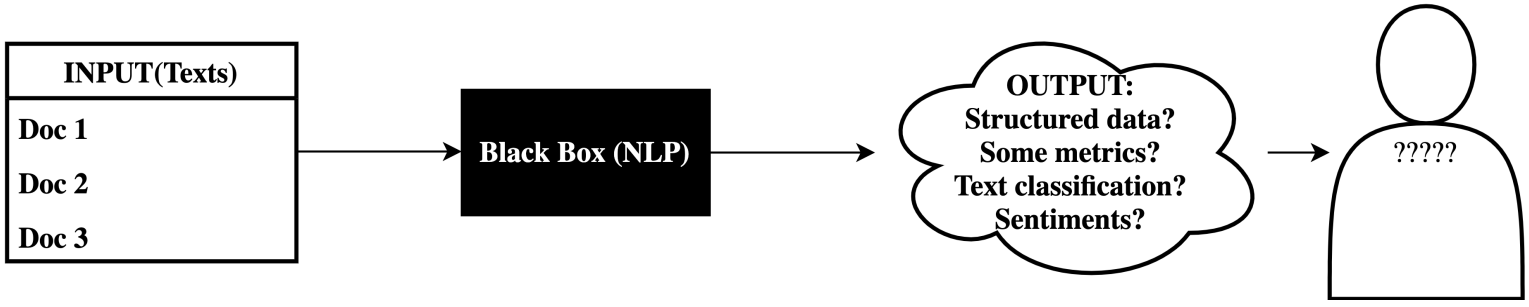
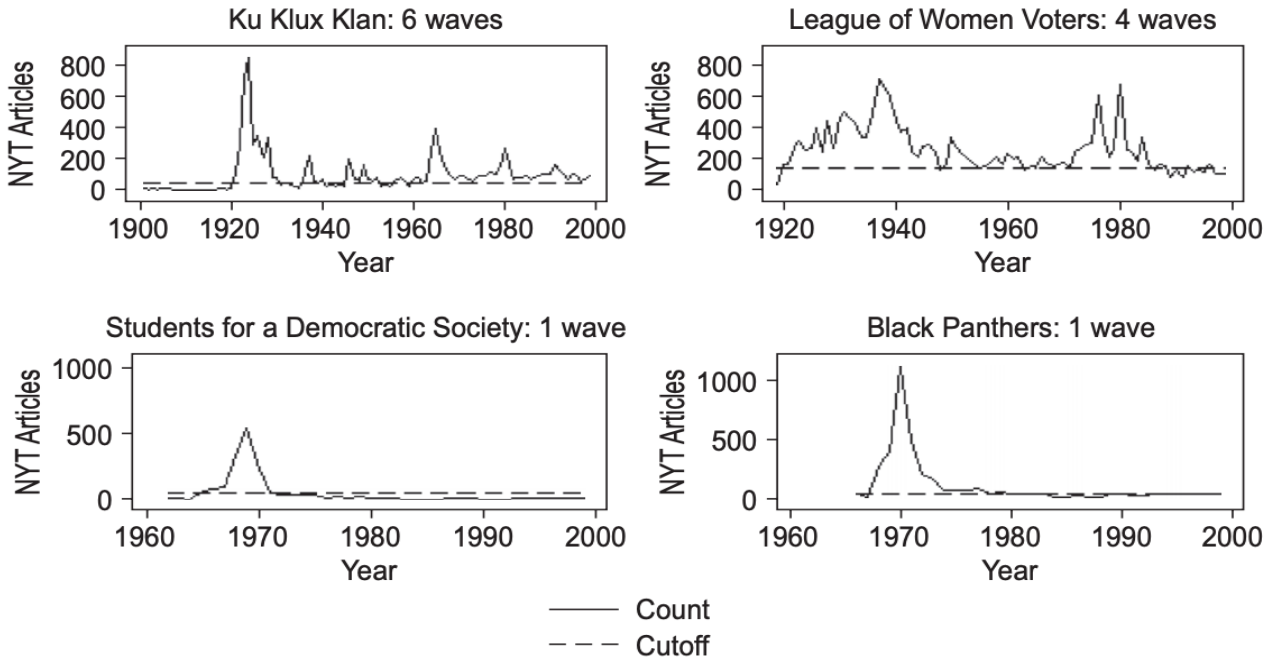


Figure 3: A Model of Thinking Texts in Social Science

- Examples: Lexicon Methods, Document-Term Matrix,etc.

Some Examples (Newspapers: NYT)

Figure 2. Cascades measurement strategy



Note: The figure illustrates the measurement of news cascades. Cascades occur in years above the article count cutoff. The size of cascades is the total number of articles in all consecutive years above the cutoff. There are 411 cascades identified in the *New York Times* data. These cases were chosen to show how the cascades measurement works in a variety of conditions, as well as to provide context for the Black Panthers' media cascade with the media attention of some well-known SMOs.

http://www.charlieseguin.com/uploads/4/1/2/7/41271621/social_forces-2015-seguin-sf-sov085.pdf

Some Examples (Quarterly Earnings Call Transcripts)

Firm-Level Political Risk: Measurement and Effects

Tarek A Hassan, Stephan Hollander, Laurence van Lent, Ahmed Tahoun

The Quarterly Journal of Economics, Volume 134, Issue 4, November 2019, Pages 2135–2202, https://doi.org/10.1093/qje/qjz021
Published: 26 August 2019

Abstract

We adapt simple tools from computational linguistics to construct a new measure of political risk faced by individual U.S. firms: the share of their quarterly earnings conference calls that they devote to political risks. We validate our measure by showing that it correctly identifies calls containing extensive conversations on risks that are political in nature, that it varies intuitively over time and across sectors, and that it correlates with the firm's actions and stock market volatility in a manner that is highly indicative of political risk.

https://sites.google.com/view/firmrisk/home?authuser=0

Some Examples (Congressional Speeches)

Econometrica, Vol. 87, No. 4 (July, 2019), 1307–1340

MEASURING GROUP DIFFERENCES IN HIGH-DIMENSIONAL CHOICES: METHOD AND APPLICATION TO CONGRESSIONAL SPEECH

MATTHEW GENTZKOW
Department of Economics, Stanford University and NBER

JESSE M. SHAPIRO
Department of Economics, Brown University and NBER

MATT TADDY
Amazon

We study the problem of measuring group differences in choices when the dimensionality of the choice set is large. We show that standard approaches suffer from a severe finite-sample bias, and we propose an estimator that applies recent advances in machine learning to address this bias. We apply this method to measure trends in the partisanship of congressional speech from 1873 to 2016, defining partisanship to be the ease with which an observer could infer a congressperson's party from a single utterance. Our estimates imply that partisanship is far greater in recent years than in the past, and that it increased sharply in the early 1990s after remaining low and relatively constant over the preceding century.

KEYWORDS: Partisanship, polarization, machine learning, text analysis.

https://www.brown.edu/Research/Shapiro/pdfs/politext.pdf

Some Examples (Survey Responses)

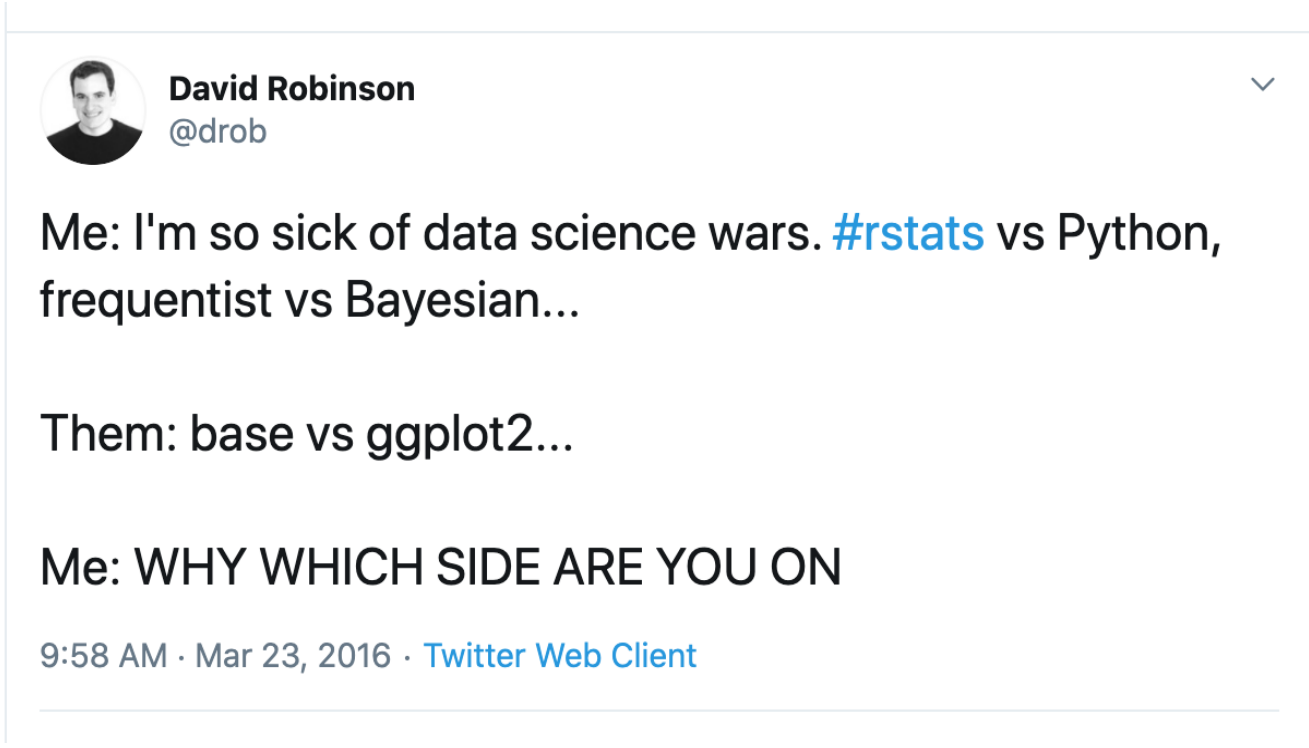
Structural Topic Models for Open-Ended Survey Responses

Margaret E. Roberts University of California, San Diego
Brandon M. Stewart Harvard University
Dustin Tingley Harvard University
Christopher Lucas Harvard University
Jetson Leder-Luis California Institute of Technology
Shana Kushner Gadarian Syracuse University
Bethany Albertson University of Texas at Austin
David G. Rand Yale University

Collection and especially analysis of open-ended survey responses are relatively rare in the discipline and when conducted are almost exclusively done through human coding. We present an alternative, semiautomated approach, the structural topic model (STM) (Roberts, Stewart, and Airoldi 2013; Roberts et al. 2013), that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This article focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with analysis of text from surveys and experiments.

https://scholar.harvard.edu/dtingley/files/topicmodelsopenendedexperiments.pdf

Today's Lab



https://twitter.com/drob/status/712639593703542785

Thank You

Yongjun Zhang, Ph.D.

Yongjun.Zhang@stonybrook.edu

https://yongjunzhang.com