Spring 2019, Soc592:
Text as Data: Statistical Text Analysis for the Social Sciences[*]

Class: Thursday 1:30-4:30pm (Wallace 165)
Precept: Monday 4:30pm-6:30pm (tentative time, location TBD)

Brandon Stewart
bms4@princeton.edu, scholar.princeton.edu/bstewart
Phone: 609-258-5094
Office Hours: When the door is open (Wallace Hall 149)

Cambria Naslund, Preceptor
cnaslund@princeton.edu

This is a half-semester graduate-level course on statistical text analysis. The course meets once a week in a three hour block for six weeks total. The goal of the class is to learn about the use of large quantities of text in research design. While the main course does not cover details of practical implementation, we offer a two hour optional precept each week that will provide hands-on experience with the methods covered during the week's class in the programming language `R`.

The major assignments involve reading, writing short research proposals and responding to proposals of your peers. Students of all departments and backgrounds are welcome.

The last day of class is March 14 and the final assignment is due April 1st.

As per the norm in Sociology mini-courses, there is assigned reading to be completed before the first class. See below.

1 Course Overview

1.1 The Idea of the Course

Never before in human history has so much information been so easy to access. The promise of this wealth of information is immense, but because of its pure volume it is difficult to summarize and interpret. However, a burgeoning array of algorithms and statistical methods are beginning to make analysis of this information possible. These new forms of data and new statistical techniques

[*] Last Edited: January 31, 2019. The development of this course has been influenced by my frequent collaborators: Justin Grimmer, Molly Roberts and Dustin Tingley.

provide opportunities to observe behavior that was previously unobservable, to measure quantities of interest that were previously unmeasurable, and to test hypotheses that were previously impossible to test.

In this course we will introduce the *social science* logic for how text can be included in every stage of the research process. Our goal is to describe the prevalence of a social behavior or phenomenon and make inferences about its origins. We explain how the abundance of text and new statistical methods facilitate these inferences. The goal of inference in social science research is qualitatively different than the goals that have been often used to evaluate text analytic methods, which often focus on performing a specific task. The focus on inference will push us to reconsider when and how some methods are useful, suggest new ways to evaluate methods, and will present new open questions in the use of text as data.

This course is organized around around the tasks in the research process: discovery, measurement and inference. Discovery is the process of hypothesis generation and often where scholars begin a research project. We will discuss methods that suggest ways of organizing texts that are specific to the task of discovery and help the researcher through the process of understanding the contours of the data. Measurement is the process of capturing the degree or extent of some behavior. We will introduce methods specifically focused on measurement, but also explain how we modify methods used for discovery to methods that are more specific to the goal of measurement. Ultimately we measure properties in text because we want to draw inferences about the world. We cover predictive and causal inference using text.

The goal of the course is to provide students with an overview of the literature while developing an understanding of what is possible. While the time scale does not permit a deep mathematical understanding of every approach, students will learn about tools for analyzing texts quantitatively and intuition for why the tools are useful.

1.2 Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. The lecture and readings will use mathematical notation and cover statistical material. To fully follow all this material would require some coursework or other study of Bayesian inference (at the level of Gelman et al. *Bayesian Data Analysis* or Murphy *Machine Learning: A Probabilistic Perspective*). We don't expect everyone to have this background coming in and we understand that everyone will likely feel lost at some point during the class. If you are interested in using text analysis to do research, this is still the class for you as long as you are willing to do your best to follow along.

The precepts will be taught using the `R` programming language and so if you want to attend and fully experience these optional sessions, it is probably necessary to have some experience with `R` coming in.

1.3 Auditing

If you are a student, we would *strongly* prefer that you actually enroll in the class. If you aren't a student (staff member, post-doc, etc.) we are happy to have you as an auditor but we ask that you complete a subset of the work which is detailed in the Assignments section below. We ask that you do the reading, participate in class and provide comments on the proposals of your classmates during the semester. You don't need to write proposals of your own though or participate in the end of semester proposal reviews. Please email Brandon's assistant Kristen Cuzzo (kcuzzo@princeton.edu) to be added to Blackboard.

2 Materials

2.1 GRS Manuscript

Throughout the semester I will provide working drafts of chapters from a book that I am completing with Justin Grimmer and Molly Roberts which I will denote GRS. This is a work of progress with rewrites likely being done shortly before it is posted, so it will need your help! Please feel free to pass along any suggestions, comments, pointed criticisms etc. This is a particularly good time to use Perusall to annotate specific lines you don't understand. While I'm happy to take typo types of corrections, I'm not trying to crowd-source a copy-edit so I wouldn't recommend burning time on that.

2.2 Books

Some people really like to have physical books for background reading. If that describes you, here are some books for your reading enjoyment.

Suggested Background Reading

Content Analysis

- Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology.* 3rd ed. Sage.
- Neuendorf, Kimberly. 2017. *The Content Analysis Guidebook.* 2nd ed. Sage.
- Boyd-Graber, Jordan, Yuening Hu, and David Mimno. 2017. "Applications of Topic Models." *Foundations and Trends in Information Retrieval* 11(2-3):143-296.

Natural Language Processing

- Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3nd ed. Prentice Hall. [Dan Jurafsky has drafts of the updated 3rd Edition posted on his website `https://web.stanford.edu/~jurafsky/slp3/` which is definitely worth a look.]
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press. [online edition available at `https://nlp.stanford.edu/IR-book/`]
- Silge, Julia, and David Robinson. 2017. *Text Mining with R: A Tidy Approach.* O'Reilly Media. [available at `https://www.tidytextmining.com/`]

Machine Learning

- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning.* Springer.

- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. [available at `https://www.deeplearningbook.org/`]
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer. [available at`https://web.stanford.edu/~hastie/ElemStatLearn/`]

Computational Social Science

- Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

2.3 Articles

Readings will be posted on *Perusall.* Perusall is a platform with collaborative annotation that allows you to post and answer questions directly in the text itself. This gives us the opportunity to answer questions outside of class in the text itself. So asking good questions not only helps you, it helps your classmates. If you know the answer to a question that another student posted, please make a contribution to the class and try to answer it!

We will send an email to the class with the code for joining and providing basic instructions.

2.4 Slack

Slack is a collaboration tool which we will be using for class as a way to post reviews, ask questions and encourage collaboration and discussion across the class. The class slack is `textasdataprinceton`. We will provide information for signing up on the first day of class. For those who haven't used it before, Slack is a little like a series of chatrooms. It has a desktop client as well as an app for your phone. See `slack.com`.

3 Assignments

We want this class to help you get your research done. Thus, the major assignment for this class is to write a few short research proposals and give feedback on proposals for others.

Each week will follow a similar schedule with a slight amendment for the first week. A typical week (starting from class) will involve the following tasks which are given in more detail below:

- Thursday 1:30-4:30pm: class
- Friday 5:00pm: deadline for signing up to do a proposal for the coming week, precept materials posted for the following week.
- Monday 4:30-6:30pm (tentative): precept covering practical implementation.
- Tuesday 5:00pm: proposals due
- Thursday 9:00am: proposal reviews due, proposal authors read all reviews in preparation for class.

There are five types of assignments in the course. They are listed below with how many times they will be done:

1. *Collaborative Annotation and Reading:* (6×) This will be done every week in the Perusall online system.

2. *Research Proposal:* (2×) For two of the weeks you will submit a short (max 800 words) proposal of how the methodology described that week could be applied to a research topic of interest to you.

3. *Proposal Reviews:* (5×) Each week except the first you will respond/review a research proposal of one of your colleagues which you have selected. We will talk in class about how to write effective reviews.

4. *Refined Proposal:* (1×) At the end of the semester you will submit a refined/updated version of one of your research proposals for comments and review from your class mates and teaching staff. You can write a new proposal if you wish- whatever would be most helpful to your research.

5. *Refined Proposal Review:* (2×) At the end of the semester you will be assigned two refined proposals to review and comment on.

We describe each of these assignments in more detail below.

For the first week you will only need to do the reading. For the subsequent five weeks you will a) do the reading, b) do either a proposal (twice, three others are off weeks), and c) do a review of a classmate's proposal. Due to the nature of the timing completing assignments on time is *extremely important* and so we ask your diligence in meeting deadlines. The course should afford you opportunities to get an enormous amount of feedback on your work.

3.1 Collaborative Annotation and Reading

The majority of the readings will be from articles in the field and the excerpts from my in-progress book manuscript. We will use the Perusall system to help each other out and work together on these readings through collaborative annotations. This will not only allow classmates to help each other understand the material, it will also highlight for us what material would best be covered in class time.

3.2 Research Proposal

The research proposal is the major assignment for the class. In at most 800 words, your goal is to lay out how you could apply the class of methods discussed in the current week's readings to a research topic of interest to you. This is a great opportunity to get feedback on potential research projects from both the teaching staff and your fellow students. During the course you will write two proposals, refine one, and receive around 6-7 reply memos. The higher the quality of your proposal, the more helpful your feedback will be.

Guidance on the Proposal There aren't many hard and fast rules for the research proposals other than the sharp 800 word cutoff. However, we expect that strong research proposals will include: a research question or area of interest, a plausible corpus of documents to analyze, a proposal for a method, and a clear statement of how the method will help you learn about the question of interest. It is this last step- connecting the method to the substantive topic of interest that can be very challenging. We want you to gain an intuition for what these methods do well and thinking about them in the context of your own research is a big part of that. If you have preliminary analyses on a set of data you have already collected you are more than welcome to include them but we do not expect that most people will have had the opportunity to do so. You may also pose questions to your reviewers on areas of the project where you need some help.

On the Length Cutoff and Deadlines We are asking your classmates to provide a thoughtful review of your proposal within a fairly short time-frame. Thus it is extremely important that you complete your proposal by the deadline and within the word limit. We explicitly use a word count rather than a page limit so that you can include figures if this would be helpful. Please do not mistake the 800 word limit as an indication that this is a simple assignment. Writing clearly and concisely is challenging and it may take you several drafts to articulate your idea well. There is less reading than a typical mini so we'd like you to reinvest that energy in these proposals.

Choosing a Proposal Topic You are only required to write two proposals. Try to choose topics which fit well with your research interests. Note that choosing a less popular week will allow you to get more feedback as the total number of reviews is constant and the number of proposals is variable. Finally, the proposals are intended to be applications of existing methods to new social science questions. However, you are also welcome to propose the development of new methods within the area discussed that week if you are so inclined.

Posting your Proposals By Tuesday at 5PM, you need to submit your proposal to Blackboard for peer review. We will provide a handout detailing the procedure for submitting a proposal.

3.3 Proposal Reviews

Each week you will review a proposal written by one of your classmates. The goal of the review is to provide constructive feedback on the proposed research project. The comments could address core areas such as corpus selection, the applicability of the methods or the theoretical relevance. Here again we have very few hard and fast rules because we want you to give the best comments that you can give based on your unique skills and background. These reviews will be assessed on how helpful they are to the author.

We do ask that you start each review with a short summary of what you think the author is proposing to do and why the author believes it is important. This establishes a common baseline so that the author knows how their work was read and what you are actually responding to. You would be amazed how often this part of the review is one of the most useful pieces!

Posting Reviews With 40% of the class on average posting a proposal each week and everyone writing a review, the process of turning around reviews in two days has the potential to get messy quick. Please carefully follow the instructions we provide on how to post reviews and do so on time.

3.4 Refined Proposal

At the end of the course you will submit a final research proposal. This is intended to be a refined/updated version of one of your two previously submitted proposals but can also be a new piece of work. If a new piece of work, feel free to weave together methods from across the different weeks if you like.

We have three requirements for the proposal:

1. Take one concrete step
   While the refined proposal is still a proposal, we ask that you take at least *one concrete step* towards completing the project. This might be collecting a sample of the data, testing out the method on some related but easier to access data, verifying that the data you want actually exists, or writing out some of the math for a new method you are proposing. The more forward progress you can make, the better.

2. Respond to the reviews
   Respond, at least implicitly, to the reviews you received in earlier rounds. You don't have to do what they were asking, but justify why you aren't.

3. Keep it to 1000 words or less
   This is 25% more words than the original proposal. Use it wisely!

These proposals will be due on Monday March 25 at 5PM.

3.5 Refined Proposal Review

On the Monday March 25 deadline for the refined proposals you will receive the refined proposal for two other classmates. You will submit written comments on the work as previously done for the regular proposals. You will have one week, until Monday April 1st at 5pm to complete this work.

3.6 Grading

Final grades will be assessed based on the assignments described above according to the following breakdown:

1. Class Participation, Collaborative Annotation and Reading: 25%

2. Research Proposals: 25%

3. Final Research Proposal: 25%

4. All Proposal Reviews: 25%

By default the class is graded on a PDF basis. If you need a letter grade there is a form which you can get from your graduate program coordinator. Bring that form to Brandon and he will happily sign it.

4 Class and Precept Structure

4.1 Class

Class meets once a week for three hours and will include a mix of lecture and group discussion. Each class will cover the current day's topic and a preview of the coming week's topic. Thus when you go to do the reading you will have already seen a short introductory lecture, and after you have done the reading you will have more lecture and discussion to help fill in the details.

Internally most classes will use the following approximate structure

- Lecture (1 hour)
  a wrap-up lecture on the current week's topic which builds from the readings you did before class.
- Discussion (25 minutes)
  a small group discussion on the current week's topic including material raised in lectures and readings. There will be time for groups to report back to the larger class.
- Proposal Discussion (15 minutes)
  a small group discussion where proposal authors will discuss with their reviewers. This will begin with the author providing a short summary of the issues raised by the reviewers and then a discussion between the authors and reviewers about the proposal.
- Break (10 minutes)
  three hours is a long time without a break.
- Q&A (10 minutes)
  a question and answer session wrapping up any loose ends on the day's topic.
- Preview Lecture (1 hour)
  a lecture providing a preview and introduction to the material that you will be reading about and writing proposals about over the next week.

4.2 Precept

Every week after the first we will provide an optional precept where the preceptor will cover some of the practicalities of how you implement the week's material in `R`. We will try to provide the materials the Friday before so that you can have a look and decide if you want to join. The precept will presume that you have done the reading for the following week; the emphasis is on how the technique looks in practice, rather than an introduction to it.

While these sessions are not required, we highly recommend that you attend them in order to gain intuition about the approaches. We expect that seeing how the methods work on actual data will help understanding how the methods can be applied to your research question.

4.3 Additional Help

We ask that you post questions for the instructional staff or your classmates in the appropriate channel in Slack. We also highly encourage you to create your own channels in the Slack and build a community together.

5 Course Outline

The course takes place over six weeks. We may adjust the schedule and/or readings throughout the semester. The last time I taught this course, we added articles at the end of the semester that hadn't been written when the semester began!

The reading listed for a given date is the reading you should do *before* coming to that day's class.

(1) Text as Data in Social Science (February 7)

This first class will cover basic details of the course and motivate the use of text data in the social sciences. During the time that would usually be devoted to proposal discussions here, we will cover some class logistics.

Reading

- GRS Chapter 2
- Grimmer, Justin, and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents." *Political Analysis* 21(3): 267-97
- Evans, James, and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1): 21-50.
- Grimmer, Justin. 2015 "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48(1): 80-83.
- DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* 2(2): 1-5.
- Wallach, Hanna. 2018. "Computational Social Science $\neq$ Computer Science + Social Data." *Communications of the ACM* 61(3): 42-44.

Optional Reading

- Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology.* 3rd ed. Sage.
- Lazer, David, and Jason Radford. 2017. "Data Ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43(1): 19-39.
- Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." NBER Working Paper #23276.
- Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20(1): 529-44.
- Schwartz, H. Andrew et al. 2013. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PLoS ONE* 8(9): 1-16.
- Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2): 254-77.
- Michel, Jean-Baptiste et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331(6014): 176-82.
- Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3): 465-82.
- Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press.

(2) Representation: how we turn text into numbers (February 14)

This class follows several distinct threads including how we select a corpus, extract text, and convert that text into a numerical representation.

Reading

- GRS Chapter 3
- Denny, Matthew J., and Arthur Spirling. 2018. “Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to do About It.” *Political Analysis* 26(2): 168-89.
- Hirschberg, Julia, and Christopher D. Manning. 2015. “Advances in Natural Language Processing.” *Science* 349(6245): 261-66.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” In *Advances in Neural Information Processing Systems 26*: 3111-19.
- Blog Posts from “Off the Convex Path”
  - “Semantic Word Embeddings”
    `http://www.offconvex.org/2015/12/12/word-embeddings-1/`
  - “Word Embeddings: Explaining their properties”
    `http://www.offconvex.org/2016/02/14/word-embeddings-2/`
  - “Simple and efficient semantic embeddings for rare words, n-grams, and language features”
    `http://www.offconvex.org/2018/09/18/alacarte/`
- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Science* 356(6334): 183-86.
- King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” *American Journal of Political Science* 61(4): 971-88.

Optional Reading

- Turney, Peter D., and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37(1), 141-88.
- Schofield, Alexandra, and David Mimno. 2016. “Comparing Apples to Apple: The Effects of Stemmers on Topic Models.” *Transactions of the Association for Computational Linguistics* 4: 287-300.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 1532-43.
- Wilkerson, John, David Smith, and Nicholas Stramp. 2015. “Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach.” *American Journal of Political Science* 59(4): 943-56.
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. (2018). “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes.” *Proceedings of the National Academy of Sciences* 115(16): E3635-44.
- Benoit, Kenneth et al. 2018. “quanteda: An R Package for the Quantitative Analysis of Textual Data.” *Journal of Open Source Software* 3(30): 774.

(3) Discovery: uncovering what we want to study (February 21)

This class will discuss methods of discovery: or ‘How do we organize our texts and generate hypotheses for our work?’ We will discuss methods for identifying discriminating words and unsupervised clustering. Throughout we will emphasize that in discovery we are less concerned with assumptions of model holding and merely generating interesting hypotheses and questions.

Reading

- GRS Chapter 4.
- Grimmer, Justin, and Gary King. 2011. “General Purpose Computer-Assisted Clustering and Conceptualization.” *Proceedings of the National Academy of Sciences* 108(7): 2643-50.
- Taddy, Matt. 2013. “Multinomial Inverse Regression for Text Analysis.” *Journal of the American Statistical Association* 108(503): 755-70. (you may skip Sections 3-4 if you like)
- Grimmer, Justin. 2013. “Comment: Evaluating Model Performance in Fictitious Prediction Problems.” *Journal of the American Statistical Association* 108(503): 770-71.
- Blei, David M. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55(4): 77-84.
- Nelson, Laura K. 2017. “Computational Grounded Theory: A Methodological Framework.” *Sociological Methods & Research.*
- Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. “Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?” *Journal of the Association for Information Science and Technology* 68(6): 1397-1410.

Optional Reading

- Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16(4): 372-403.
- Chuang, Jason, Christopher D. Manning, and Jeff Heer. 2012. “Without the Clutter of Unimportant Words: Descriptive Keyphrases for Text Visualization.” *ACM Transactions on Computer-Human Interaction* 19(3): 19:1-19:29.
- Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. “Reading Tea Leaves: How Humans Interpret Topic Models.” In *Advances in Neural Information Processing Systems 22.*
- 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(2): 326-43.

(4) Measurement with Unsupervised Methods: learning measurement categories (February 28)

Shifting into measurement, the next two classes will discuss how to use text as data methods to measure some property of interest. The ultimate purpose may be description, causal inference or prediction. We start by discussing the implications of re-appropriating and extending some of the methods we covered in discovery for measurement purposes.

Reading

- GRS Chapter 5 part 1
- Quinn, Kevin M. et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1): 209-28.
- Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1): 1-35.
- Roberts, Margaret E. et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064-82.
- Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515): 988-1003.
- Spirling, Arthur. 2012. "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784-1911." *American Journal of Political Science* 56(1): 84-97.

Optional Reading

- O'Connor, Brendan, Brandon M. Stewart, and Noah A. Smith. 2013. "Learning to Extract International Relations from Political Context." (2013) In *Proceedings of the Association of Computational Linguistics.*
- Blaydes, Lisa, Justin Grimmer, and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *The Journal of Politics* 80(4): 1150-67.
- Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Computational Social Science: Discovery and Prediction.* Cambridge University Press.
- Young, Daniel Taylor. 2013. "How Do You Measure a Constitutional Moment? Using Algorithmic Topic Modeling To Evaluate Bruce Ackerman's Theory of Constitutional Change." *Yale Law Journal* 122(7): 1990-2054.
- Catalinac, Amy. 2016. *Electoral Reform and National Security in Japan: From Pork to Foreign Policy.* Cambridge University Press.
- DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6): 570-606.
- Bail, Christopher A. 2016. "Combining Natural Language Processing and Network Analysis to Examine How Advocacy Organizations Stimulate Conversation on Social Media." *Proceedings of the National Academy of Sciences* 113(42): 11823-28.

(5) Measurement with Supervised Methods: using known categories (March 7)

This class will talk about how to take a known organizational structure for our data and measure quantities of interest such as individual document classifications or proportions over a corpus. We will cover dictionary methods (and their limits), hand coding procedures, and methods to learn classifiers from hand coding. We will also discuss the importance and difficulties of validation.

Reading

- GRS Chapter 5 part 2
- 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.
- Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229-47.
- Hengel, Erin. 2018. "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review."
- Nielsen, Richard A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad.* Cambridge University Press. (excerpts TBD)
- Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2012. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity* 2012(2): 1-81.
- Spirling, Arthur. Comment on: "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity* 2012(2): 70-79.
- Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." NBER Working Paper #22423.

Optional Reading

- Jamal, Amaney A., Robert O. Keohane, David Romney, and Dustin Tingley. 2015. "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses." *Perspectives on Politics* 13(1), 55-73.
- Voigt, Rob et al. 2017. "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect." *Proceedings of the National Academy of Sciences* 114(25): 6521-26.
- Tausczik, Yla 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.
- Dodds, Peter, and Christopher M. Danforth. 2009. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." *Journal of Happiness Studies* 11(4): 441-56.
- Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem" *Journal of the American Statistical Association* 58(302): 275-309.
- Benoit, Kenneth, Kevin Munger, and Arthur Spirling. 2018. "Measuring and Explaining Political Sophistication through Textual Complexity."
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. "Diffusion of Lexical Change in Social Media." *PLoS ONE* 9(11): 1-13.

(6) Causal Inference and Prediction: understanding the world (March 14)

In our final class, we discuss testing our theories using the frameworks of causal inference and prediction. We will discuss how text work fits into the Rubin Causal Model and how to think about text as response, treatment and confounder. We will also discuss text as a predictor of non-text events.

Reading

- GRS Chapter 6-7.
- Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. 2019. "Adjusting for Confounding with Text Matching."
- Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2019. "How to Make Causal Inferences Using Texts."
- Fong, Christian J., and Justin Grimmer. 2019. "Causal Inference with Latent Treatments."

Optional Reading

- O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).*
- Eichstaedt, Johannes C. et al. 2015. "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality." *Psychological Science* 26(2): 159-69.
- Mueller, Hannes, and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review* 112(2): 358-75.

In the period of class that would typically preview the next week, we will talk about the connections between "text as data" methods and approaches to networks, images, audio and video.

Readings for Networks, Images, Audio

- Torres, Michelle. "Understanding Visual Messages: Visual Framing and the Bag of Visual Words."
- Joo, Jungseock, and Zachary C. Steinert-Threlkeld. "Image as Data: Automated Visual Content Analysis for Political Science."
- Zhang, Han, and Jennifer Pan. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology.*
- Knox, Dean, and Christopher Lucas. "A Dynamic Model of Speech for the Social Sciences."