Text as Data: Statistical Text Analysis for the Social Sciences

Brandon Stewart and Clark Bernier *

Class: Tuesdays 2-5pm (McCosh Hall, Room 64)
Precept: Fridays 10am-12pm (Location: Wallace 004).

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

Clark Bernier, Preceptor
cbernier@princeton.edu
Office hours/Precept: Fridays 10a-12p, Wallace 004

1 The Basics

1.1 Overview

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 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 a 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 six week course is organized around three large components of the research process: discovery, measurement, and testing. 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

*The development of this course has been influenced by my frequent collaborators: Justin Grimmer, Molly Roberts and Dustin Tingley. Last Edited: March 21, 2016

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. Testing is the least established area, in which we use text for prediction and causal inference. Text provides an opportunity for granular causal inferences and opens a wide range of questions previously impossible. The pairing of causal inference and text has produced new methodological questions that we highlight and provide initial answers to.

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. Students with a prior course which covers maximum likelihood estimation or Bayesian inference will be most comfortable with the statistical material. Others should consult the instructor. While we will provide statistical details on the model and some training in `R` these will not be required. Gary King's *Unifying Political Methodology* and Larry Wasserman's *All of Statistics-A Concise Course in Statistical Inference*, in addition to Murphy's *Machine Learning* listed below, offer excellent primers on the MLE framework underlying some of these methods.

1.2.1 Processing Text

Unfortunately, the half-semester format of this courses forces us to focus on parts of the toolkit needed to do text analysis at the expense of other parts. Specifically, this course focuses on the how to form good research questions with the available text analysis tools and the optional precept and lab materials give an overview of work-flows for each. These work flows presume, however, starting with a curated and cleaned text dataset. Messing with raw text data can be daunting the first time you do it, and though we won't be able to cover text processing in depth in the course, we encourage you to make use of the additional materials throughout this syllabus. Specifically, *STM* makes use of the quanteda package for text data manipulation which we strongly encourage. Hadley Wickham's Advanced R and Matloff's The Art of R Programming are both *excellent* resources with introductions to working with text.

1.3 Auditing

We are happy to have auditors in the class but we ask that auditors complete a subset of the assignments which are detailed in the Assignments section below, specifically the reading and weekly proposal reviews but not the proposals or end of semester proposal reviews. Auditors are welcome to do proposals as well if they would like feedback on their work but are unable to take the class. Please email Brandon's assistant Kristen Matlofsky (kristent@princeton.edu) to be added to Blackboard.

2 Materials

2.1 Computational Tools

The best way, and often the only way, to learn new statistical procedures is by doing. Precept will cover use of many of these computational tools with programming done in R. These materials are highly recommended but not required.

2.2 Books

The reading will primarily be from relevant articles in the field. Depending on how the semester goes, we may offer chapters from a draft book manuscript by Grimmer, Roberts and Stewart which will be made available on Perusall.

Suggested It is often helpful to see the same material in alternative ways. Thus here are some other texts you might consult.

Natural Language Processing

- Manning, Raghavan, and Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Jurafsky, Daniel and James Martin. 2008. *Speech and Language Processing*. Prentice Hall.

Machine Learning

- Murphy, Kevin, 2012. *Machine Learning: A Probabilistic Perspective*
- Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Hastie, Tibshirani, and Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edition. Springer.

2.3 Articles

Readings will be posted on *Perusall*. Perusall is a new ebook 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.

3 Assignments

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

- Tuesday 5PM: computational lab materials released for the following week.
- Wednesday-Friday: do the reading including collaborative annotations on Perusall

- Friday 10am-12pm: a two hour block for office hours and computational skills with the preceptor.
- Sunday 5pm: proposals due
- Tuesday 9am: proposal reviews due
- Tuesday 2pm-5pm: class (on the topic prepared in the previous week)

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 been assigned. 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 promptly 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. The interdisciplinary nature of these methods means that the articles will be drawn from a variety of different fields and the lack of a single unifying text book treatment means that these articles will often implicitly demand different backgrounds. 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 to the instructor 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 and refine one and receive around 6-7 reply memos total. 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 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 Sunday, 5pm, you need to submit your proposal to Blackboard for peer review. Under Tools:Discussion Board, where is a *Forum* for each week. To submit a proposal for a week, *create a new thread with the title of your submission* and post your proposal as a pdf.

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 selections, 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.

Posting Reviews With 12-15 reviews per week and 40 reviewers, the process of turning around reviews in two days has the potential to get messy quick. Our proposed approach is admittedly cumbersome, but if everyone sticks to it, it will guarantee everyone is getting response to their proposals. This only works if all of use coordinate. We ask you follow the below procedure *carefully* to ensure that reviews are done *timely* and *fairly*:

1. Look through the titles of proposals (*"Threads"*) posted and choose one from those with fewer than two reviews (the number of "Total Posts" with fewer than 3).
2. Open the *Thread* you've chosen, and click *"Reply"*.
3. Add a placeholder message to your reply and click *"Submit"*. This serves as your communication to the rest of the class of which proposal you are electing to review. If you do not do this *before* writing you response, then it's possible that once you've written your response, your chosen proposal will no longer be in need of responses and you'll have to write a second response to another proposal.
4. When you've finished your response, *EDIT* your original placeholder and add the text from your reply.

If every thread already has 3 "Total Posts," then you're welcome to respond to the thread of your choosing. If you post a third response to a thread while there are still threads without two responses, however, you're response won't be counted for credit.

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. In order to give you a bit more room, these proposals will have a 1000 word limit-use your 25% increase wisely! These proposals will be due on Monday May 9 at 5PM.

3.5 Refined Proposal Review

On the Monday May 9 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 until Friday May 13 at 5PM to complete your comments.

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 Lab Structure

4.1 Class

The class is structured such that in a given week we will spend the majority of time covering material on which proposals have already been written and reading has already been done. The allocated class time is 3 hours with a short break in the middle.

In general class time will include: some lecture on particular models or ideas, Q&A on topics of interests to the class, discussion of trouble areas raised on Piazza, discussion of the optional papers and how they applied the described methods and discussion of research proposals submitted by students. Please bring questions or topics of interest to the class as we will have plenty of time for discussion.

For the last part of the class (probably about 40-45 minutes), Brandon will provide a preview of the material for the following week which will help frame and guide the readings.

4.2 Lab

Every week after the first we will provide a handout which sketch a research work flow in `R` for the basic analyses relevant to the week. The lab time which will be 10a-12pm on Fridays will involve the preceptor walking through these materials and helping students with computation. The lab 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. With these handouts will be at least one sample dataset which students can use to gain some intuition for how these methods work in practice.

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 with determining whether or not the methods are right for your research questions.

4.3 Additional Help

4.3.1 Piazza

We encourage you to post questions about the readings to Piazza. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates, the preceptors, and the professor. Unless the question is of a personal nature or completely specific to you, you should not e-mail teaching staff; instead, you should post your questions on Piazza. The course staff will be monitoring the page, but we encourage you to help your classmates as well. The course Piazza site can be accessed through the course BlackBoard page under Tools:Piazza.

4.3.2 Data and Statistical Services

Princeton's DSS Lab is another excellent resource when you hit roadbumps while pushing through megabytes of text. Their walk-in hours this spring are 1-5p Weekdays in Firestone A-12-G.

5 Course Outline

The course takes place over six weeks. We may adjust the schedule due to comprehension, time, and interest. Please note also that readings are subject to change, in particular you should expect that we will add individual articles for discussion in class.

5.1 Core Ideas (March 22)

This first class will cover basic details of the course and motivate the use of text data in the social sciences. We will outline the organizing framework of the class based on the four steps for analyzing text: (1) Identification of Text/Population of Study, (2) Discovery, (3) Measurement, (4) Testing.

Reading

- Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents" *Political Analysis*. 21, 3 267-297.
- Michel et al 2011, "Quantitative analysis of culture using millions of digitized books" *Science*, 331:6014.
- DiMaggio, Paul. "Adapting computational text analysis to social science (and vice versa)." Big Data & Society 2.2 (2015).

Optional Reading

- Schwartz, H. Andrew, et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach." PloS one 8.9 (2013): e73791.
- Lucas C, Nielsen R, Roberts ME, Stewart BM, Storer A, Tingley D. "Computer assisted text analysis for comparative politics." *Political Analysis*. (2015);23(2):254-277
- Monroe, Burt and Phil Schrodt. 2008. "Introduction to the Special Issue: The Statistical Analysis of Political Text". *Political Analysis* 16, 4, 351-355
- Blei, David M. "Build, compute, critique, repeat: Data analysis with latent variable models." *Annual Review of Statistics and Its Application* 1 (2014): 203-232.
- Brendan O'Connor, David Bamman, and Noah A. Smith. "Computational Text Analysis for Social Science: Model Assumptions and Complexity." (2011) *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*
- Romney, D., Stewart, B., & Tingley, D. (2015). Plain Text: Transparency in the Acquisition, Analysis, and Access Stages of the Computer-assisted Analysis of Texts. *Qualitative and Multi-Method Research*, 13(1), 32-37.

5.2 Discovery: Uncovering what we want to study and generating (March 29)

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/embedding methods. Throughout we will emphasize that in discovery we are less concerned with assumptions of model holding and merely generating interesting hypotheses and questions.

Reading

- Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization" *Proceedings of the National Academy of Sciences* 108(7), 2643-2650
- King, Pan and Roberts. 2013 "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*
- Monroe, Colaresi and Quinn (2008) "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict" *Political Analysis* 16: 372-403.

Optional Reading

- Chuang, Jason, Christopher Manning and Jeff Heer. "Without the Clutter of Uninmportant Words: Descriptive Keyphrases for Text Visualization " (2012) *ACM Transactions on Computer-Human Interaction*
- Chang, Boyd-Graber, Gerrish, Wang and Blei (2009) "Reading Tea Leaves: How Humans Interpret Topic Models" *Neural Information Processing Systems*
- Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." *Science* 315.5814 (2007): 972-976.
- Tenenbaum, Joshua, Charles Kemp, Thomas Griffiths and Noah Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction" (2011) *Science*
- Mikolov, Tomas, Ilya Sutskever, Chen, Corrado and Dean "Distributed Representations of Words and Phrases and their Compositionality" (2013) *Advances in Neural Information Processing Systems*
- Hannah, Lauren A. and Hanna M. Wallach "Summarizing Topics: From Word Lists to Phrases" (2014) *NIPS Workshop on Modern Machine Learning and Natural Language Processing*
- King, Gary, Patrick Lam, and Margaret E. Roberts. 2014. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text."

5.3 Measurement: Supervised Methods (April 5)

Shifting into measurement, this class will talk about how to take an 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

- Loughran, Tim and Bill McDonald.2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks" *Journal of Finance* 66, February 35-65

- Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis" *Journal of the American Statistical Association* 108, 755-770 (you may skip Sections 3-4 if you like)

- Grimmer, Justin. 2013. "Evaluating Model Performance in Fictitious Prediction Problems" *Journal of the American Statistical Association.* http://stanford.edu/ jgrimmer/mirc.pdf

- Hopkins, Dan and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science" *American Journal of Political Science*, 54, 1

- Jamal, A., Keohane, R., Romney, D., & Tingley, D. (2015). Anti-Americanism or Anti-Interventionism in Arabic Twitter Discourses. *Perspectives on Politics*, 13(1), 55-73.

Optional Reading

- Soroka, Stuart and Lori Young. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts" *Political Communication* 29: 205-231

- Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." Journal of language and social psychology 29.1 (2010): 24-54.

- Dodds, Peter and Christopher Danforth. 2009. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents". *Journal of Happiness Studies* 11, 4. 441-456

- Mosteller, Frederick and David Wallace. 1963. "Inference in an Authorship Problem" *Journal of the American Statistical Association* 58, 302. 275-309

- Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech". *Journal of Information, Technology, and Politics*. 5(1).

- Stewart, Brandon M. and Yuri M. Zhukov "Use of force and civil-military relations in Russia: an automated content analysis" (2009) *Small Wars & Insurgencies*

- Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. "Diffusion of lexical variation in online social media" (2014) *PLOS-ONE,*

- Dorazio et al. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines *Political Analysis* 22, 2 224- 242

5.4 Measurement: Topic Models (April 12)

Our second week on measurement turns to unsupervised methods. Methodologically we will focus on mixed membership topic models and discuss systematic approaches to testing within this framework using the Structural Topic Model. We will discuss in length the validation and evaluation of unsupervised models.

Reading

- Blei, David. 2012. "Probabilistic Topic Models". *Communications of the ACM*. 55, 4, 77-84

- Roberts, et al "Topic Models for Open-Ended Survey Responses with Application to Experiments" *American Journal of Politcal Science* 2014

- Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". *Political Analysis*, 18(1), 1-35.

- DiMaggio, Paul, Manish Nag, and David Blei. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding." Poetics 41.6 (2013): 570-606.

Optional Reading

- Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation" *Journal of Machine Learning*

- Quinn, Kevin et al. 2010 "How to Analyze Political Attention with Minimal Assumptions and Costs". *American Journal of Political Science*, 54, 1 209-228.

- Wallach, Hanna, David Mimno, and Andrew McCallum. "Rethinking LDA: Why Priors Matter". *Proceedings of the 23rd Annual Conference on Neural Information Processing*

- Chp 5. Wallach, Hanna "Structural Topic Models for Language" `http://people.cs.umass.edu/~wallach/theses/wallach_phd_thesis.pdf`

- Roberts, Margaret, Brandon Stewart, and Edo Airoldi "A Topic Model for Experimentation in the Social Sciences" Forthcoming *Journal of the American Statistical Association* .

- Nelson, Laura K. "Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City" Manuscript

- Young, Daniel Taylor "How Do You Measure a Constitutional Moment? Using Algorithmic Topic Modeling To Evaluate Bruce Ackermans Theory of Constitutional Change" (2013) *Yale Law Journal*

- Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. "Learning to Extract International Relations from Political Context." (2013) *Proceedings of the Association of Computational Linguistics*

- Blaydes, Grimmer and McQueen "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds" (2016) Manuscript.

- Mohr and Bogdanov. 2013. Topic models: What they are and why they matter. *Poetics*.

- Roberts ME, Stewart BM, Tingley D. Navigating the Local Modes of Big Data: The Case of Topic Models. 2016. In: *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press.

5.5 Measurement: Scaling (April 19)

Our final week on measurement will concern scaling from both a supervised and unsupervised perspective. We will discuss different methods as well as the unique challenges of validation in this framework.

Reading

- Lowe, William. (2016) "There's (Basically) Only One Way to Do it" Manuscript.

- Lowe, W. and Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, 21(3):298313.

- Spirling, Arthur (2012). US treaty making with American Indians: Institutional change and relative power, 17841911 *American Journal of Political Science*.

- Nielsen, Richard (2014) "Networks, Careers, and the Jihadi Radicalization of Muslim Clerics" Manuscript pending revision into the book *Deadly Clerics: Blocked Ambition and the Turn to Violent Jihad*

Optional Reading

- Kim, In Song, John Londregan, and Marc Ratkovic. "Voting, Speechmaking, and the Dimensions of Conflict in the US Senate." Annual Meeting of the Midwest Political Science Association. 2014.

- Lowe, Will. 2008. "Understanding Wordscores". *Political Analysis*. 16, 356-371.

- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data". *American Political Science Review*. 97, 2, 311-331

- Jackman, Simon, Joshua Clinton and Doug Rivers. 2004. "The Statistical Analysis of Roll Call Data". *American Political Science Review* 98, 2, 355-370.

- Slapin, Jonathan and Sven-Oliver Prokschk. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts". *American Journal of Political Science*. 52, 3 705-722

- Soroka, Stuart and Lori Young. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts" *Political Communication* 29: 205-231

- Dodds, Peter and Christopher Danforth. 2009. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents". *Journal of Happiness Studies* 11, 4. 441-456

- Beauchamp, Nick. 2012. "Using Text to Scale Legislatures with Uninformative Voting" *Northeastern University Mimeo*

5.6 Testing: Causal Inference and Prediction (April 26)

In our final class, we discuss testing our theories using the framework 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

- Roberts ME, Stewart BM, Nielsen R. (2016) "Matching Methods for High-Dimensional Data with Applications to Text"

- Review: Roberts et al "Structural topic models for open-ended survey responses." *American Journal of Political Science*. 2014;58:1064-1082

- Fong, Christian and Justin Grimmer (2016) "Discovery of Treatments from Text Corpora"

Optional Reading

- Gill, Michael and Andrew Hall. (2016) "How Judicial Identity Changes The Text Of Legal Rulings" Manuscript.

- Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." (2010) *ICWSM*