Text as Data
Political Science 452, Fall 2014
Tuesday, Thursday 9:30am-11:00am, GSL

Instructor: Justin Grimmer, Political Science Department

Office: Encina Hall West, Room 414

Contact: jgrimmer@stanford.edu, 617-710-6803. Gchat; justin.grimmer@gmail.com

Office Hours: My door is almost always open during normal business hours. If you absolutely need to see me at a particular time, please schedule an appointment

Programming TA: Frances Zlotnick, Political Science Department

Office/Programming Section: Encina Hall West, Room 417

Contact: Zlotnick@stanford.edu

Office Hours: 230-430 pm and by appointment

Language is the medium for politics and political conflict. Candidates debate during elections. Representatives write laws. Nations negotiate peace treaties. Clerics issue Fatwas. Citizens express their opinions about politics on social media sites. These examples, and many others, suggest that to understand what politics is about, we need to know what political actors are saying and writing.

This course introduces techniques to collect, analyze, and utilize large collections of text for social science inferences. The ultimate goal of the course is to introduce students to modern quantitative text analysis techniques and provide the skills necessary to apply the methods in their own research. In achieving this ultimate goal, students will also learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems. They will also have the opportunity to develop their programming abilities and develop an original research project or to participate in an ongoing research project.

Prerequisites

At a minimum, students should have completed coursework on univariate inference and linear regression. The ideal student will have also taken a course on model based inference. The course will develop student's programming skills. Prior experience with `R`, `Python`, or a related language is strongly recommended. If you have any questions about whether you're ready for the course, please speak with me.

Evaluation

Students will be evaluated across three areas.

Homework Students will be asked to complete a weekly homework assignment. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for their work. Even if students are auditing the course, I hope they'll attempt the homework assignments.Portions of the homework completed in R should be submitted using R markdown, a markup language for producing well-formated HTML documents with embedded R code and outputs. R markdown requires installation of the knitr package. We recommend using Rstudio, an IDE for R, which is set up well for the creation of R markdown documents. Python assignments, including code and output, should be submitted in LaTeX or similar document preparation language. We recommend downloading and installing the Enthought python distribution, which includes many of the most commonly used packages and the Canopy python IDE.

More about `RStudio` can be found here:
http://www.rstudio.com/

And `R Markdown` can be found here:
http://rmarkdown.rstudio.com/

Final Project Students will have the opportunity to complete a final project. One option for the final project is to complete an original research project and (in the best case scenario) the project will contribute to completing their dissertation, field paper, or ongoing research. This is not always possible, particularly for students who are just beginning their research program. A second option is that students can participate in an ongoing research project that has some large text analytic component. I have several examples of such projects that cover a wide range of substantive interests and from a diverse set of faculty. On the first day of class I'll introduce some of the projects and if students are interested, they should stop by my office to discuss the project further.

Political science is an increasingly collaborative discipline. So, students will be allowed (and encouraged) to complete the final project as a two-person team.

Students who are taking the 3-unit version of the course are not required to complete the final project. That said, I'm happy to advise a student on a project if they want to begin work on a paper that might later be a publication, field paper, or component of dissertation.

Students will present their final project during a class wide poster session on the final class meeting, where all faculty and graduate students will be invited to attend. Poster sessions provide the opportunity to receive a lot of feedback from many people and (I think) are the best way to present research to receive actual feedback. After the poster session students will submit a paper. Specifics about the paper will be discussed in class.

Participation Students are expected to attend each class and to ask questions regularly. To encourage questions and discussion we will use Piazza. You should enroll in the course

(Poli Sci 452) at the following link:
piazza.com/stanford/fall2014/polsci452

Books

There are no required books for the class. But there are many books on Text Analysis and Machine Learning you may find useful.

Natural Language Processing

- Manning, Raghavan, and Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
  Available at `http://nlp.stanford.edu/IR-book/information-retrieval-book.html` (hereafter MRS)
- Jurafsky, Daniel and James Martin. 2008. *Speech and Language Processing*. Prentice Hall.

Machine Learning

- 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.
- McLachlan and Peel. 2000 *Finite Mixture Models* Wiley.
- McLachlan and Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd Edition Wiley.

Computer Languages

- Lutz, Mark. 2010. *Programming Python*. 4th Edition O'Reilly [python on cover]
- Lutz, Mark. 2009. *Learning Python*. 4th Edition O'Reilley [mouse on cover]

Class Outline

9/23: Text as Data: Characterizing the Haystack

- 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.

- Monroe, Burt and Phil Schrodt. 2008. "Introduction to the Special Issue: The Statistical Analysis of Political Text". *Political Analysis* 16, 4, 351-355

9/25: A Statistics, Computing Refresher

- `http://thomasleeper.com/Rcourse/Intro2R/Intro2R.pdf`

9/30: Acquiring and Manipulating Text Data

- Berinsky, Adam and Gregory Huber and Gabriel Lenz. 2011. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk". *Political Analysis* 20, 3. 351-368

- Porter, MF. 2001. "Snowball: A Language for Stemming Algorithms" `http://snowball.tartarus.org/texts/introduction.html`

- `http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html`

10/2: Dictionary Methods: Measuring Weighted Word Usage

- 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

- 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

10/7: Methods for Finding Discriminating Words and Applications

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

- 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)

- Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis" *Journal of the American Statistical Association* 108, 755-770

- MRS, Section 13.5

10/9: The Vector Space Model and the Geometry of Text

- Hand out, refresher on linear algebra [course work]

10/14: Principal Components, Multi-dimensional Scaling, and Texts

- 14.8. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]

- Spirling, Arthur. US Treaty-Making with American Indians: Institutional Change and Relative Power 1784-1911 *American Journal of Political Science* 56, 1, 84-97.

10/16: Counts, Proportions, and Distributions (Getting to Know the Dirichlet Distribution and Other Distributions on the Simplex)

- Chapter 2 Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning* (Sections 2.1, 2.2 especially) [coursework]

- Katz, Jonathan and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data" *American Political Science Review* 93, 1, 15-32.

10/21: Clustering Methods 1: Fully Automated Clustering Models

- 14.3. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]

- Chp 9. Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning* (Sections 2.1, 2.2 especially) [coursework]

10/23: Clustering Methods 2: Interpretation and Computer Assisted Clustering

- Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization" *Proceedings of the National Academy of Sciences* 108(7), 2643-2650

10/28: Topic Models 1: Vanilla LDA

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

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

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

10/30: Topic Models 2: Structural Topic Models

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

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

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

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

- Roberts, Margaret, Brandon Stewart, and Edo Airoldi "Structural Topic Models" *Harvard University Mimeo* .

11/4: Guest Lecture: Hanna Wallach (UMASS-Amherst CS, Microsoft Research). "The Case for Hierarchical Bayesian Modeling in Text

11/6: Supervised Methods 1: Classifying Documents, Training Coders and LASSO

- Policy Agendas Codebook `http://www.policyagendas.org/page/topic-codebook`

- Grimmer, Justin, Gary King and Chiara Superti "Patterns of Partisan Tauting in the US Senate" *Stanford University Mimeo*

- 3.4. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [coursework]

11/11: Supervised Methods 2: Naive Bayes, Support Vector Machines, and Read Me

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

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

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

11/13: Model Fit, Complexity, and Cross Validation

- Chp 7. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer.

11/18: Ensembles of Classifiers

- Hillard, Dustin, Stephen Purpura and John Wilkerson. 2007. "Computer Assisted Classification for Mixed Methods Social Science Research". *Journal of Information, Technology, and Politics*.

- 7.10. Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning* Springer. [Coursework]

- van der Laan, Mark and Eric Polley and Alan Hubbard. 2007. "Super Learner" *Statistical Applications in Genetics and Molecular Biology* 6, 1.

- Chapter 3. Grimmer, Justin, Sean Westwood, and Solomon Messing. 2014. "The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability" *Princeton University Press*

11/20: Using Text to Measure Ideology: Word Scores

- 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

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

12/2: Using Text to Measure Ideology: Item Response Theory

- 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

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

12/4: Poster Session