Applied Text Analysis for Social Scientists: Texts as Data
Political Science 452, Spring 2011
Wednesday 215-505, the GSL
coursework

Political conflict often occurs in text–either spoken or written. Candidates debate during elections. Representatives write laws. Nations negotiate peace treaties. Clerics issue Fatwas. These examples, and many others, suggest that to understand what politics is about, we need to know what political actors are saying.

This class will provide a set of tools for exploring the content of texts. The focus is applied. Students will learn about tools for analyzing texts quantitatively and intuition for why the tools are useful. Proofs, however, will be avoided.

Students will be evaluated using home work and a final paper. Home work will be assigned weekly and should be completed with a data set of text that students plan to use for their final paper. In the final paper, students should apply techniques learned in this course to analyze texts quantitatively. Homework and the final paper will be given equal weight in grades.

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

Introduction and Acquiring Text Data

- Jackman, Simon. 2006. "Data from the Web Into `R` " *The Political Methodologist*. 14, 2. 11-15

- Berinsky, Adam and Gregory Huber and Gabriel Lenz. 2011. "Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research". Yale University Mimeo.

Representing Texts as Data

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

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

- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification using Machine Learning Techniques" *Proceedings of EMNLP* [coursework]

Dictionary Methods: Inference about Classes with Knowledge About Words

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

- 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 [coursework]

Methods for Identifying Distinctive Words and Phrases: Inference About Words with Knowledge About Classes

- Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". 16(4) [coursework]

- MRS, Section 13.5

- Yano, Tae, Philip Resnik, and Noah Smith. 2010. "Shedding (a Thousand Points of) Light on Biased Language" *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* [coursework]

Vector Spaces, Term Weighting, Distance Measures, and Projection

- MRS 6.2 and 6.3

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

- Spirling, Arthur. 2010. "Bargaining Power in Practice: US Treaty-making with American Indians, 1784-1911". Harvard University Mimeo [coursework]

Unsupervised Learning: Clustering Models

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

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

Unsupervised Learning: Mixed Membership Models

- Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation" Journal of Machine Leanring. [coursework]

- Quinn, Kevin; Burt Monroe, Mike Colaresi, Mike Crespin, and Drago Radev. 2010 "How to Analyze Political Attention with Minimal Assumptions and Costs". AJPS, 54, 1 209-228. [coursework]

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

Supervised Learning: An Introduction

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

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

Supervised Learning: Ensemble Methods

- 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.

Applications/Wildcard