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-formatted 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
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

- Jurafsky, Daniel and James Martin. 2008. Speech and Language Processing. Prentice Hall.

Machine Learning


Computer Languages


Class Outline

9/23: Text as Data: Characterizing the Haystack


9/30: Acquiring and Manipulating Text Data

10/2: Dictionary Methods: Measuring Weighted Word Usage
- 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

10/7: Methods for Finding Discriminating Words and Applications
- 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]
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]


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


10/28: Topic Models 1: Vanilla LDA


10/30: Topic Models 2: Structural Topic Models


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


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

11/18: Ensembles of Classifiers


12/2: Using Text to Measure Ideology: Item Response Theory
- Beauchamp, Nick. 2012. “Using Text to Scale Legislatures with Uninformative Voting” Northeastern University Mimeo

12/4: Poster Session