Syllabus
Text as data
XXXXX XXXXXX

Instructor: Jared Edgerton
Office: 2032 Derby Hall
Email: edgerton.37@osu.edu
Office Hours: TBD
Meeting Place & Time: TBD
Course Web Site: jaredfedgerton.net/classes

Rationale and Scope

This course explores emerging statistical methods for extracting important signals from language stored as text. Topics include text collection and processing; dictionary methods; topic modeling; document clustering; deep learning; and concomitant computational and mathematical challenges.

Politics comprises, in large part, the use of language. Candidates for political office barnstorming for months at a time, making their cases and leveling criticisms of opponents' actions and points of view; legislators draft and debate bills; presidents issue statements; justices hand down legal opinions; information age rabble-rousers peddle conspiracy theories to anyone with an internet connection; and journalists, columnists and bloggers dissect, reframe and disseminate it all to the American public. And those are just a handful of examples from American politics, ignoring the treaties , trade agreements, speeches, and declarations by leaders and media around the world.

For political scholars, the trove of data locked away in transcripts and manuscripts is both a blessing and a curse. The answers to many fascinating questions lie within the written word; yet harnessing the data in ways both efficient and reliable poses considerable methodological challenges. In political science, we have sought for decades to use text as a source of data. Over the past two decades, we have built our capacity to process, quantify, and model text using computational methods. With increased computing power and advanced statistical methodology, scholars have developed new, powerful, and efficient tools to extract patterns from, and test substantive theories using, text as data.

Prerequisites

The course also assumes a working knowledge of that you have a strong understanding of generalized linear models and Bayesian statistics before taking the class.

Evaluation

Your final grade will be based on several problem sets (40%) throughout the semester, a take home test on machine learning (25%), a final paper in which you produce a high quality manuscript (e.g. one that could eventually be published) using the techniques we cover (20%), and the presentation of this paper to the class and a general audience (15%). You should complete the scheduled reading *before the class listed!*.

I subscribe to OSU's grading rubric: A 93-100, A- 90-92.9, B+ 87-89.9, B 83-86.9, B- 80-82.9, C+ 77-79.9, C 73-76.9, C- 70-72.9, D+ 67-69.9, D 60-66.9, E 0-59.

Academic Misconduct

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Students with Disabilities

The University strives to make all learning experiences as accessible as possible. If you anticipate or experience academic barriers based on your disability (including mental health, chronic or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. The Office of Student Life Disability Services is located in 098 Baker Hall, 113 W. 12th Avenue; telephone 614-292-3307, slds@osu.edu; http://slds.osu.edu/.

Course Norms

- Speak up when you have a question.
- Teamwork and collaboration is *highly encouraged* on every aspect of the course. Students may work together on assignments but must write out their own homework and list everyone they worked with. However, you are not allowed to divvy up the problems such that one person does one problem an another the next. You are allowed to collaborate on the final paper if you like (max 2 authors and both get the same grade regardless of real or perceived contributions).
- *All* homework assignments must be written in LaTeX. Assignments not written in LaTeX (or `sweave` if you want to be really fancy) will be returned without a grade.

Texts

There is not current good textbook on text as data so we will have course readings during the semester.

Tentative Schedule

Part 1. Basics of machine learning (prediction, dimension reduction, supervised and unsupervised learning)

- Historical context of speech;
- How to use forecasting to understand ideology;
- Dimension reduction to identify latent clusters.

Week 1 (TBD 8) **Introduction to machine learning** *In this section, we will discuss basic supervised machine learning techniques. Using prediction for inference.*

Week 2 (TBD 15) **Penalized regression** *We will discuss penalized regression and apply the concepts to some political science research.*

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. "Measuring group differences in high-dimensional choices: method and application to congressional speech." Econometrica 87.4 (2019): 1307-1340.

Green, Jon, et al. "Elusive consensus: Polarization in elite communication on the COVID-19 pandemic." Science Advances 6.28 (2020): eabc2717.

Week 3 (TBD 29) **Supervised machine learning** *This lecture introduces SVMs and random forest models. We will spend the class doing applied examples.*

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." Political Analysis 16.4 (2008): 372-403.

Young, Lori, and Stuart Soroka. "Affective news: The automated coding of sentiment in political texts." Political Communication 29.2 (2012): 205-231.

Week 4 (TBD 29) **Unsupervised machine learning** *This lecture introduces neural networks for prediction.*

Week 5 (TBD 5) **Dimension reduction** This lecture will tie in how to data clusters together. This has direct implications for understanding political ideology.

Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

Spirling, Arthur. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." American Journal of Political Science 56.1 (2012): 84-97.

McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

Week 6 (TBD 12) **Clustering** Identifying latent groups.

Du, K-L. "Clustering: A neural network approach." Neural networks 23.1 (2010): 89-107.

Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." Proceedings of the twenty-first international conference on Machine learning. 2004.

Johnson, Stephen C. "Hierarchical clustering schemes." Psychometrika 32.3 (1967): 241-254.

Week 7 (TBD 26) **Midterm take home test**

Part 2. Applications to text

Week 8 (TBD 5) **Topic models** *Identifying topics in a text corpus* Will use an applied example to identify topics in President Obama and Trump's speeches.

Bagozzi, Benjamin E., and Daniel Berliner. "The politics of scrutiny in human rights monitoring: evidence from structural topic models of US State Department human rights reports." Political Science Research and Methods 6.4 (2018): 661-677.

Greene, Derek, Derek O'Callaghan, and Padraig Cunningham. "How many topics? stability analysis for topic models." Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014.

Week 9 (TBD 19) **Structural topic models**

Quinn, Kevin M., et al. "How to analyze political attention with minimal assumptions and costs." American Journal of Political Science 54.1 (2010): 209-228.

Week 10 (TBD 26) **Dimension reduction on text**

Kim, Hyunsoo, Peg Howland, and Haesun Park. "Dimension reduction in text classification with support vector machines." Journal of machine learning research 6.Jan (2005): 37-53.

Howland, Peg, Moongu Jeon, and Haesun Park. "Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition." SIAM Journal on Matrix Analysis and Applications 25.1 (2003): 165-179.

Week 11 (TBD 5) **Neural network and deep learning approaches to text** *This topic is underexplored in political science. We will emerging tools.*

Week 12 (TBD 10) **Scaling language**

Dhillon, Paramveer, Dean P. Foster, and Lyle H. Ungar. "Multi-view learning of word embeddings via cca." Advances in neural information processing systems. 2011.

Week 13 (TBD 2) **Presentation of research papers**

Week 14 (TBD 2) **Presentation of research papers**

Week 15 (TBD 16) **Presentation of research papers**