POLS 8500: Text As Data

Maymester 2022

Professor:	Joe Ornstein	**Time:**	MTWThF 11:00am – 1:45pm
Email:	jornstein@uga.edu	**Place:**	101D Baldwin Hall
Website:	maymester-text-as-data-2022		

So much information about our political world is stored not in tidy datasets, but in *texts*. Party manifestos, social media posts, city council minutes, email correspondence, treaties, ordinances, judicial decisions, presidential debates, Congressional speeches....the practice of politics in its myriad forms ultimately comes to be recorded in words on a page. How do we take all this unstructured text, far too extensive for any one researcher to read, and convert it into useful information for scientific inquiry?

In this class we will explore the cutting edge of "text-as-data", with a focus on how to retrieve, represent, and analyze political texts as part of a research project. We will aim to cover both the conceptual and the practical, spending class time writing code to replicate and extend significant results from the past decade of political science research.

Course Prerequisites and Objectives

I will assume that before taking this course you have taken the POLS 7012 and 7014 or their equivalents (introduction to statistics and linear regression) and that you are familiar with the basics of the `R` programming language. By the end of this course, you will be able to:

- Retrieve and clean text data from a variety of sources
- Thoughtfully measure and quantify concepts of interest to your research using text data
- Fit machine learning models for clustering, sentiment analysis, topic classification, and prediction

Assignments & Grading

There will be daily coding activities in class designed to apply the concepts we learn to practical research problems. Your grade will be based on the successful completion of these assignments.

Office Hours

Since we're meeting for 3 hours every day, I will not hold formal office hours, but I will generally be available before and after class to discuss questions and will respond to emails from 9am to 4pm.

Textbooks

For our tour through the world of text as data, we will rely on two textbooks. The first is a a broad conceptual overview of the field of text as data, and the second is a practical introduction to the `R` code you can use to work with text (and is available online for free at the link below).

- Grimmer, Justin, Brandon M. Stewart, and Margaret E. Roberts. *Text As Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press, 2022.
- Silge, Julia, and David Robinson. *Text Mining with R: A Tidy Approach.* First edition. O'Reilly, 2017.

Tentative Course Outline and Readings

Please read and markup these readings before class each day.

Day 1: Getting Started

What Are We Doing Here? A review of R and RStudio.

- GSR Chapters 1-2

Day 2: Creating a Corpus

Where and how do we get text data? Playtime with Twitter.

- GSR Chapters 3-4
- SR Chapter 1: Tidy Text Data

Day 3: The Bag of Words

What if we ignored everything we know about language and just counted the words? Would that get us anywhere?

- GSR Chapter 5
- SR Chapter 5: Document-Term Matrices

Day 4: Modeling The Bag of Words

Probabilistic vs. Algorithmic Models, The Federalist Papers

- GSR Chapters 6-7

Day 5: Text Reuse

Let's take a model designed to detect plagiarism and use it to see how ideas spread in political texts over time.

- GSR Chapter 9.1
- Wilkerson, John, David Smith, and Nicholas Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." *American Journal of Political Science* 59(4): 943–56.
- Jansa, Joshua M., Eric R. Hansen, and Virginia H. Gray. 2019. "Copy and Paste Lawmaking: Legislative Professionalism and Policy Reinvention in the States." *American Politics Research* 47(4): 739–67.

Day 6: Sentiment Analysis

Dictionaries and Supervised Learning, Supervised Learning

- SR Chapter 2

Day 7: Word Embeddings

What if words were actually just a bunch of numbers?

- GSR Chapter 8
- Rodriguez, Pedro L., and Arthur Spirling. 2021. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research" *The Journal of Politics.*

Day XXX: Clustering

Day XXX: Topic Models

Latent Dirichlet Allocation (LDA): Not Quite As Frightening As That Name Makes It Seem

- GSR Chapter 13
- SR Chapter 6

Day 13: Crowd-Coding

What if we let other people do the work? MTurk, SentimentIt

- GSR Chapter 18
- Carlson, David, and Jacob M. Montgomery (2017). "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts." *American Political Science Review* 111, no. 4: 835–43.

Day 14: Foundation Models

Here we bring it all together. Take a massive semi-supervised neural network, train it on the entire Internet, and see what happens.

- GSR Section 8.5
- Ornstein, Blasingame, and Truscott (2022). "How To Train Your Stochastic Parrot." *Working Paper.*
- Alexander, Scott. 2019. "GPT-2 As Step Toward General Intelligence." *Slate Star Codex.*
- Alexander, Scott. 2020. "The Obligatory GPT-3 Post." *Slate Star Codex.*

Day 15: Bonus Day

Catch-up and Topics By Popular Demand

Mental Health and Wellness Resources

- If you or someone you know needs assistance, you are encouraged to contact Student Care and Outreach in the Division of Student Affairs at 706-542-7774 or visit https://sco.uga.edu. They will help you navigate any difficult circumstances you may be facing by connecting you with the appropriate resources or services.
- UGA has several resources for a student seeking mental health services or crisis support.
- If you need help managing stress anxiety, relationships, etc., please visit BeWellUGA for a list of FREE workshops, classes, mentoring, and health coaching led by licensed clinicians and health educators in the University Health Center.
- Additional resources can be accessed through the UGA App.