Text-as-Data

POLSCI 797TA
University of Massachusetts Amherst
Spring 2021

Contact Information

Dr. Douglas Rice
Department of Political Science
Thompson Hall 218
University of Massachusetts Amherst
Amherst, MA 01003
E-mail: drrice@legal.umass.edu

Course Time: Tuesdays & Thursdays, 10:00 – 11:15 a.m. ET
Course Location: https://umass-amherst.zoom.us/j/99642249981
Office Hours: Tuesday 2:30 - 4:00 p.m. ET, or by appointment

Course Description

With the recent explosion in availability of digitized text, social scientists increasingly are turning to computational tools for the analysis of text as data. In this three credit course, students will first learn how to convert text to formats suitable for analysis. From there, the course will introduce and proceed through tutorials on a variety of natural language processing approaches to the treatment of text-as-data. This will include relatively simple dictionary approaches for measurement, supervised learning approaches for document classification, vector representations, contextualized embeddings, and more.

Course Objectives

- Equip students to become knowledgeable consumers of text data research, capable of critically analyzing research that employs text data and text-as-data techniques.
- Provide students with the tools to design and complete basic and advanced text-as-data research, from converting text to formats appropriate for analysis to estimating text-as-data models.
- Develop students ability to work individually and collaboratively on subjects with important real-world relevance.
- Enable students to communicate — both orally and in written format — clearly and appropriately the results or shortcomings of text-as-data research.

Text

There is no required text for this course. All required readings are posted to the course website. This syllabus outlines general areas of study throughout the semester, as well as listing specific reading assignments on a daily basis. It is vital that you keep current with the readings, as they will provide the basis for in-class lectures and discussions.

Grading

Grades are calculated as follows:

Modules & Quizzes (30%)

Blog Posts (30%)

Research Project (30%)

Participation (10%)

Modules & Quizzes Students are required to complete a series of 10 short interactive modules online through Google Colab. Each will familiarize students with a different text-as-data approach, and how to implement said approach in `R`. After completing the module, students will complete a short quiz. Quizzes are primarily graded on completeness rather than having correct responses to each question; therefore, it is imperative that you complete each quiz.

Blog Posts Students are required to complete blog posts every other week during the course of the semester. The blog posts should detail your progress working with text-as-data with the corpus that you are using for your research project. As such, they should serve as ongoing documentation of (a) your growing expertise with text-as-data, and (b) your growing progress on the final project. Your first blog post should detail your general interests and the research question(s) you plan to explore. After that, blog posts should simply reflect the material for that week. Therefore, you might detail in one blog post your experience with getting the corpus for your final project into R and formatted to your liking. The blog posts could include plots and screen caps of successful work, but might also include details on unforeseen challenges, terrifying error messages, or just what you think is an unbelievably ugly plot you mistakenly created. The goal is to document and reflect on your progress.

The blog posts will be written and submitted to the professor as an R markdown (.Rmd) document. *Plots should be submitted separately as PDF files.* Each blog post is due to be submitted on Sundays by 11:59 p.m.

Research Project The class is tailored around aiding you in producing a research paper appropriate for submission and presentation at an academic conference. Students are expected to complete an original research project that features both a strong theoretical grounding, research design, and original analysis that features a text-as-data component.

Participation As a group, we will meet for lecture every Tuesday. During this time, we will cover the central concepts for that week and discuss examples of research engaging in that type of text-as-data analysis. Students must contact the professor *before* the class session they will miss in order to ensure the absence is excused. In class, students are expected to participate regularly, and participation should reflect careful consideration of the readings and topic.

On Thursdays, we will **not** meet as a group. Instead, this time should be considered lab times during which — if you have not already — you complete the tutorials and quizzes and work on completing the the work for that week with your respective text corpora and associated blog posts. Students interested in meeting may sign up for meetings using the link on the course website; during these meetings, we can discuss any challenges, concerns, questions, or just general thoughts that are arising as you are carrying out your analyses. **Students are required to schedule at least three meetings during this time over the course of the semester.**

Final letter grades are assigned using the University's Plus-Minus Grading Scale according to following rubric:

A (94-100%)

A- (90-93%)

B+ (86-89%)

B (81-85%)

B- (77-80%)

C+ (74-76%)

C (70-73%)

F (Below 70%)

Software

Students in this class will use `R` and `RStudio`. The software is free and available online; the course website includes a guide for installing both on your machine. The course assumes no familiarity with the `R` programming language, though that is helpful. We will also discuss and utilize Python but at a much smaller scale.

Academic Honesty

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst.

Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. The procedures outlined below are intended to provide an efficient and orderly process by which action may be taken if it appears that academic dishonesty has occurred and by which students may appeal such actions.

Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent.

For more information about what constitutes academic dishonesty, please see the Dean of Students' website:
http://umass.edu/dean_students/codeofconduct/acadhonesty/

Statement on Disabilities

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus.

If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website at http://www.umass.edu/disability/.

Course Schedule

Daily reading assignments are listed in parentheses. Note that reading assignments are listed according to the day on which the subject matter will be discussed; they should therefore be read prior to that date.

February 2nd & 4th: Introduction

Discuss syllabus, background material, and quantitative versus qualitative reading of texts. Introduction to `R` and initial set-up.

February 9th & 11th: Using Text as Data

Michel et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*.

Kenneth Benoit. 2019. "Text as Data: An Overview" *Sage Handbook of Research Methods in Political Science & International Relations.*

Margaret Roberts. 2016. "Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science." *Political Analysis*

February 14th: Blog Post 1 due

February 16th & 18th: Acquiring Texts: Scraping & APIs

John Wilkerson and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529-544.

Alex Bradley and Richard James. 2019. "Web Scraping Using R" *Advances in Methods and Practices in Psychological Science*

Deen Freelon. 2018. "Computational Research in the Post-API Age." *Political Communication* pp. 665-668.

February 23rd & 25th: Natural Language Processing

Jacob Eisenstein. 2021. Chapter 1, *Natural Language Processing*, available [here].

Daniel Jurafsky and James Martin. 2020. Chapter 2, *Speech and Natural Language Processing*, available [here].

February 28th: Blog Post 2 due

March 2nd & 4th: Preprocessing

Matthew Denny and Arthur Spirling. 2018. "Text Processing for Unsupervised Learning: Why It Matters, Why It Misleads, and What to Do About It."

Alexandra Schofield and David Mimno. "Comparing Apples to Apple: The Effects of Stemmers on Topic Models" Transactions of the Association for Computational Linguistics 4 (2016): 287-300.

March 9th & 11th: Representing Texts & Bag(s) of Words

Brendan O'Connor, David Bamman, and Noah A. Smith (2011) "Computational Text Analysis for Social Science: Model Assumptions and Complexity." *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*

Justin Grimmer 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.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. 2009. Chapter 6, *Introduction to Information Retrieval*, available [here].

March 14th: Blog Post 3 due

March 16th & 18th: Dictionary

Peter Dodds and Christopher Danforth. 2009. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and President." *Journal of Happiness Studies* 11(4):441-456.

Tim Loughran and Bill McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66(1): 35-65.

Leah Windsor, Nia Dowell, Alistair Windsor, and John Kaltner. 2018. "Leader Language and Political Survival Strategies." *International Interactions* 44(2): 321-336.

March 23rd & 25th: Supervised Learning

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment classification using machine learning techniques." *Proceedings of EMNLP*

D'Orazio et al. (2013) "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines" *Political Analysis* 22(2):224-242.

Theocharis, Y., Barber, P., Fazekas, Z., Popa, S. A. and Parnet, O. 2016. "A Bad Workman Blames His Tweets: The Consequences of Citizens Uncivil Twitter Use When Interacting With Party Candidates." *Journal of Communication*

March 28th: Blog Post 4 due

March 30th & April 1st: Scaling

Will Lowe and Ken Benoit (2013) "Validating estimates of latent traits from textual data using human judgment as a benchmark." *Political Analysis* 21(3): 298-313.

Burt Monroe, Michael Colaresi and Kevin Quinn (2008) "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict" *Political Analysis* 16:372-403.

Benjamin Lauderdale and Alex Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis*.

April 6th & 8th: Topic Models

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

Roberts et al. (2014) "Structural topic models for open-ended survey responses." *American Journal of Political Science*. 58:1064-1082.

April 11th: Blog Post 5 due

April 13th & 15th: Word Embeddings

Mikolov et al. (2013) "Distributed Representations of Words and Phrases and their Compositionality" *Advances in Neural Information Processing Systems*

Rice, Rhodes, and Nteta (2019) "Racial Bias in Legal Language" *Research & Politics*

Sebastian Ruder. 2018. "NLP's ImageNet moment has arrived." available [here]

April 20th & 22nd: BERT, eLMO, & Transformers

NOTE: There will be no class on April 20th.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. "A Primer in BERTology: What we know about how BERT works." *TACL*, available [here].

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations."

April 25th: Blog Post 6 due

April 27th & 29th: Causal Inference

Katherine Keith, David Jensen, and Brendan O'Connor (2020) "Text and Causal Inference: A

Review of Using Text to Remove Confounding from Causal Estimates." *Transactions of the Association for Computational Linguistics.*

Margaret Roberts, Brandon Stewart, and Richard Nielsen. 2020. "Adjusting for Confounding with Text Matching." *American Journal of Political Science*

Reagan Mozer, Luke Miratrix, A. Kaufman, and Jason Anastasopolous. 2020. "Matching with text data: An experimental evaluation of methods for matching documents and measuring match quality." *Political Analysis* 28(4):445-468.

May 4: Poster Presentations

No reading assignment. During this class period, students will present their research posters.