DACSS 758
Text-as-Data
Fall 2023

Instructor: Dr. Rosemary Pang
Email: mrpang@umass.edu
Office Location: Bartlett 263 or Zoom

Course Time and Location:

Monday & Wednesday 4:00 - 5:15 pm	Multimodal	Machmer W-13 or Zoom

The first session starts on September 6, Wednesday, at 4:00 PM

Some suggested practices for Zoom:

- Please mute your window unless you have a question to avoid noise interference
- Please use a headset with microphone if available
- Please raise your hand to be recognized (available in the participant window)
- Please indicate you understand something by vocal response or thumbs up/down (in the participant window), particularly if video is not enabled
- Feel free to ask and answer questions via the chat window

Office Hours:

Regular office hours will be held 1:00 PM - 3:00 PM Monday & Wednesday in Bartlett 263 or on Zoom. Please book in advance through Calendly. Make sure you specify the course and summarize the question you have. If this time does not work, please send me an email for appointment.

Attendance Policy:

All class sessions will be recorded and be made available to all students. Students who are taking this course synchronously (both in-person or online) are required to attend lectures and participate in discussions. Students who are taking this course asynchronously are required to watch lecture recordings and posted materials, and also participate in discussions on Piazza.

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.

Learning 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.

Textbook:

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.

Feedback and Questions:

Students use Google Form to provide feedback and ask questions about course material every week. The instructor will address these questions in the following week.

Course Structure and Grading:
Final grades will be based on:

- **Attendance and Participation (10%):**
  As a group, synchronous students will meet for lecture every Monday and Wednesday. 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. In class, students are expected to participate regularly, and participation should reflect careful consideration of the readings and topic. Asynchronous students are required to watch lecture recordings and participate in discussions on Piazza.
- **Modules & Quizzes (20%):**
  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 that 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.

- **Final Project Check-in Assignments (50%):**
  Students are required to complete five final project check-in assignments during the course of the semester. These assignments 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 check-in should detail your general interests and the research question(s) you plan to explore. After that, assignments should simply reflect the material for that week. Therefore, you might detail in one check-in your experience with getting the corpus for your final project into R and formatted to your liking. The check-in assignments 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 assignments should be render into PDF files using Quarto. Authoring your documents in qmd is recommended. For information on Quarto visit https://quarto.org/.

- **Final Project (20%):**
  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.

- **Grade Scale:**
  A: 94-100; A-: 90-93; B+: 86-89; B: 81-85; B-: 77-80; C+: 74-76; C: 70-73; FAIL: 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.

University Policies:

- Academic Honesty Statement:
  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. 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 (http://www.umass.edu/dean_students/codeofconduct/acadhonesty/).

- Accommodation Statement:
  The University of Massachusetts Amherst is committed to providing an equal educational opportunity for all students. If you have a documented physical, psychological, or learning disability on file with Disability Services (DS), you may be eligible for reasonable academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, please notify me within the first two weeks of the semester so that we may make appropriate arrangements. For further information, please visit Disability Services (https://www.umass.edu/disability/).

- Title IX Statement:
  In accordance with Title IX of the Education Amendments of 1972 that prohibits gender-based discrimination in educational settings that receive federal funds, the University of Massachusetts Amherst is committed to providing a safe learning environment for all students, free from all forms of discrimination, including sexual assault, sexual harassment, domestic violence, dating violence, stalking, and retaliation. This includes interactions in person or online through digital platforms and social media. Title IX also protects against discrimination on the basis of pregnancy, childbirth, false pregnancy, miscarriage, abortion, or related conditions, including recovery. There are resources here on campus to support you. A summary of the available Title IX resources (confidential and non-confidential) can be found at the following link: https://www.umass.edu/titleix/resources. You do not need to make a formal report to access them. If you need immediate support, you are not alone. Free and confidential support is available 24 hours a day / 7 days a week / 365 days a year at the SASA Hotline 413-545-0800.

Class Schedule and Readings

The schedule is tentative and subject to change. We may adjust the schedule due to time or interest.

Sept 6 Introduction

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

Sept 11 & 13 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*

Sept 18 & 20 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.

September 24th: Check-in 1 due

Sept 25 & 27 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].

Oct 2 & 4 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.

October 8th: Check-in 2 due

Oct 10 & 11 Representing Texts: Bag(s) of Words & Word Embeddings

No class on Oct 9. Oct 10 Follows Monday schedule

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].

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]

Oct 16 & 18 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.

October 22nd: Check-in 3 due

Oct 23 & 25 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*

Oct 30 & Nov 1 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.

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.*

November 5th: Check-in 4 due

Nov 6 & 8 BERT, eLMO, & Transformers

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."

Nov 13 & 15 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.

November 19th: Check-in 5 due

Nov 20 Review

No class on Nov 22: Thanks Giving Holiday

Nov 27 & 29 Wrapping Up for Final Project

Dec 4 & 6 Poster Presentation

December 10th: Final Project due