Text as Data

Alexandra Siegel
alexandra.siegel@colorado.edu

Spring 2021

Meeting Time: Fridays 9:10-11:40

Office Hours: Thursdays 2:00-4:00

Course Overview

This course is an introduction to the quantitative analysis of text as data, a rapidly growing field within the social sciences. The availability of textual data—from archival documents and politician's speeches to social media posts and online news sources—has grown massively in recent years. In this course, students will learn how to quantitatively analyze text from a social-science perspective. Students will learn different methods to acquire text, transform it to data, and analyze it to shed light on important research questions. Each week we will cover different methods, including dictionary construction and application, sentiment analysis, scaling and topic models, and machine learning classification of text. Lectures will be accompanied by hands-on exercises that will give students practical experience working with real-world texts. The goal of the course is to provide students with the skills to use a broad range of text as data methods in their own research.

Prerequisites

Students should have taken, at minimum, an introductory class in statistics before taking this course. Basic knowledge of probability, distributions, hypothesis testing, and linear models will be essential for understanding the concepts discussed in class. In addition, students should have experience working with the R programming language.

Course Logistics

The course will meet synchronously every Friday on Zoom from 9:10-11:40am MST. Class time will be a mixture of lecture, live coding exercises, and updates on student projects. Readings and assignments for each week will be posted on Canvas. Please complete each week's assigned reading before class.

Assignments and Grading

Evaluations will revolve around two components:

(a) **Reading, Attendance and Participation (10%)**: Students should come to class each week having read the assigned required readings. Students should be prepared to contribute to in-class discussion.

(b) **Problem Sets (40%)**: There will be three problem sets, all of which will involve working with various forms of text data. Students will select a dataset of their choosing to apply skills we have learned in class to complete each problem set. Ideally these exercises will contribute toward students' final projects. Students can work together or individually, depending on their preferences. Students will be assigned to peer review problem sets to provide an initial round of feedback. Students auditing the course are welcome to complete homework assignments as long as they participate in the peer review process.

(c) **Final Project (50%)**: The class is designed to build up to a final research project. Throughout the semester, students will develop a final project in which they answer a research question with textual data. Students are free to choose the topic, and are encouraged to acquire their own data for the project. To make this project as useful as possible toward ultimately producing publishable research, final projects can take several forms: 1) "Everything but the framing"; 2) A research proposal + proof of concept analysis; 3) A conference poster; 4) A complete paper; 5) Some other pre-approved format. In the last two weeks of the semester, students will present their final projects to the class in a conference-style presentation. A 1 page summary of the proposed project will be due midway through the semester on **March 12, 2021**. The final project will be due on **April 30, 2021**. Final projects must be accompanied by a replication file. Students are welcome to collaborate on final projects or combine the final project for PSCI-7108 with a class paper for a substantive course.

Course Materials

Many of the readings for this course will come from a brand new (unpublished) textbook, *Text as Data* by Justin Grimmer, Molly Roberts, and Justin Stewart. A PDF draft of the textbook is posted on Canvas. All other course readings will be provided on Canvas. I may update the readings throughout the semester depending on students' interests.

Course Outline

Week 1: Course Overview (Jan. 22)

- Make sure to have R and R Studio installed and working
- For a review of R basics, see: https://datacarpentry.org/r-socialsci/

Week 2: Introduction to Text as Data (Jan 29)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 1-2.
- Henry E. Brady. 2019. "The Challenge of Big Data and Data Science." Annual Review of Political Science.
- Grimmer, Justin 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.
- Michel et al. 2011, "Quantitative analysis of culture using millions of digitized books" Science, 331:6014.
- DiMaggio, Paul. "Adapting computational text analysis to social science (and vice versa)." Big Data & Society 2.2 (2015).

Week 3: Selecting, Acquiring, Representing Texts (Feb 5)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 3-5.
- Denny, Matthew and Arthur Spirling, 2017. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It."
- Gill, Michael and Arthur Spirling "Estimating the Severity of the WikiLeaks U.S. Diplomatic Cables Disclosure", 2015. Political Analysis 23(2), 299-305.
- Barbera, Pablo. and Rivero, Gonzalo., 2015. "Understanding the Political Representativeness of Twitter Users." Social Science Computer Review, 33(6), pp.712-729.

Week 4: Dictionary Methods and Measurement (Feb 12)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 15-16
- Loughran, Tim and Bill McDonald. 2011. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks" Journal of Finance 66, February 35-65
- Tausczik, Y. R. and Pennebaker, Jamie. W. (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 29(1), 24(54).
- 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

Week 5: Similarity & Complexity Measures (Feb 19)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapter 7
- Spirling, Arthur. 2016. "Democratization and Linguistic Complexity", Journal of Politics.
- Benoit, K., Munger, K. and Spirling, A. 2017. Measuring and Explaining Political Sophistication Through Textual Complexity
- R Peng and N Hengartner. 2002. "Quantitative Analysis of Literary Styles", The American Statistician, Volume 56.

Week 6: "Out of the Box" Sentiment Analysis (Feb 26)

- Lori Young and Stuart Soroka 2012 "Affective News: The Automated Coding of Sentiment in Political Texts." Political Communication, 29:2, 205-231.
- Naldi, M. 2019. "A review of sentiment computation methods with R packages."
- Miazga, Justyna, and Tomasz Hachaj. 2019. "Evaluation of Most Popular Sentiment Lexicons Coverage on Various Datasets." In Proceedings of the 2019 2nd International Conference on Sensors, Signal and Image Processing, pp. 86-90.
- Sentiment analysis with tidy data https://www.tidytextmining.com/sentiment.html

Week 7: Supervised Machine Learning (March 5)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 17-20
- Siegel, Alexandra, et al. "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath." (Forthcoming, in The Quarterly Journal of Political Science).
- Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." Journal of Information, Technology, and Politics. 5(1).
- 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.

Week 8: Learning Clustering and Topic Models (March 12)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 12-14.
- Roberts, Margaret E., et al. "Topic models for open ended survey responses with applications to experiments." American Journal of Political Science 58 (2014): 1064-82.
- Catalinac, Amy. "From pork to policy: The rise of programmatic campaigning in Japanese elections." The Journal of Politics 78.1 (2016): 1-18.
- Roberts, M.E., Stewart, B.M. and Tingley, D., 2019. STM: An R package for structural topic models. Journal of Statistical Software, 91(2).

Week 9: Meetings to Discuss Final Projects (March 19)

Week 10: Word Embeddings (March 26)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapter 8.
- Rodriguez, Pedro L., and Arthur Spirling. 2020. "Word Embeddings."
- Meyer, David. How Exactly Does Word2Vec Work?
- Rodman, E., 2020. A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Analysis, 28(1), pp.87-111.

Week 11: Using Text to Measure Ideology (April 2)

- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data". American Political Science Review. 97, 2, 311-331
- Lowe, Will. 2008. "Understanding Wordscores". Political Analysis. 16, 356-371.
- Barberá, Pablo. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." Political Analysis 23.1 (2015): 76-91.
- Jackman, Simon, Joshua Clinton and Doug Rivers. 2004. "The Statistical Analysis of Roll Call Data." American Political Science Review. 98, 2, 355-370.
- Slapin, Jonathan and Sven-Oliver Prokschk. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." American Journal of Political Science. 52, 3 705-722

Week 12: Causal Inference with Text (April 9)

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapters 24-27.

Week 13: Images and Video as Data (April 16)

- Won, D., Steinert-Threlkeld, Z.C. and Joo, J., 2017. "Protest activity detection and perceived violence estimation from social media images." In Proceedings of the 25th ACM international conference on Multimedia (pp. 786-794).
- Torres, Michelle., 2018. "Give me the full picture: Using computer vision to understand visual frames and political communication."
- Knox, Dean. and Lucas, Christopher, 2019. "A dynamic model of speech for the social sciences."

Week 14: Student Presentations (April 23)

Week 15: Student Presentations (April 30)

University Polices

Classroom Behavior

Both students and faculty are responsible for maintaining an appropriate learning environment in all instructional settings, whether in person, remote or online. Those who fail to adhere to such behavioral standards may be subject to discipline. Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with race, color, national origin, sex, pregnancy, age, disability, creed, religion, sexual orientation, gender identity, gender expression, veteran status, political affiliation or political philosophy. For more information, see the policies on classroom behavior and the Student Code of Conduct.

Requirements for COVID-19

As a matter of public health and safety due to the pandemic, all members of the CU Boulder community and all visitors to campus must follow university, department and building requirements, and public health orders in place to reduce the risk of spreading infectious disease. Students who fail to adhere to these requirements will be asked to leave class, and students who do not leave class when asked or who refuse to comply with these requirements will be referred to Student Conduct and Conflict Resolution. For more information, see the policies on COVID-19 Health and Safety and classroom behavior and the Student Code of Conduct. If you require accommodation because a disability prevents you from fulfilling these safety measures, please see the "Accommodation for Disabilities" statement on this syllabus. All students who are new to campus must complete the COVID-19 Student Health and Expectations Course. Before coming to campus each day, all students are required to complete the Buff Pass. Faculty, add if applicable: In this class, you may be reminded of the responsibility to complete the Buff Pass and given time during class to complete it. Students who have tested positive for COVID-19, have symptoms of COVID-19, or have had close contact with someone who has tested positive for or had symptoms of COVID-19 must stay home. In this class, if you are sick or quarantined, Faculty: insert your procedure here for students to alert you about absence due to illness or quarantine. Because of FERPA student privacy laws, do not require students to state the nature of their illness when alerting you. Do not require "doctor's notes" for classes missed due to illness; campus health services no longer provide "doctor's notes" or appointment verifications.

Accommodation for Disabilities

If you qualify for accommodations because of a disability, please submit your accommodation letter from Disability Services to your faculty member in a timely manner so that your needs can be addressed. Disability Services determines accommodations based on documented disabilities in the academic environment. Information on requesting accommodations is located on the Disability Services website. Contact Disability Services at 303-492-8671 or dsinfo@colorado.edu for further assistance. If you have a temporary medical condition, see Temporary Medical Conditions on the Disability Services website.

Preferred Student Names and Pronouns

CU Boulder recognizes that students' legal information doesn't always align with how they identify. Students may update their preferred names and pronouns via the student portal; those preferred names

and pronouns are listed on instructors' class rosters. In the absence of such updates, the name that appears on the class roster is the student's legal name.

Honor Code

All students enrolled in a University of Colorado Boulder course are responsible for knowing and adhering to the Honor Code. Violations of the policy may include: plagiarism, cheating, fabrication, lying, bribery, threat, unauthorized access to academic materials, clicker fraud, submitting the same or similar work in more than one course without permission from all course instructors involved, and aiding academic dishonesty. All incidents of academic misconduct will be reported to the Honor Code (honor@colorado.edu); 303-492-5550). Students found responsible for violating the academic integrity policy will be subject to nonacademic sanctions from the Honor Code as well as academic sanctions from the faculty member. Additional information regarding the Honor Code academic integrity policy can be found at the Honor Code Office website.

Sexual Misconduct, Discrimination, Harassment and/or Related Retaliation

The University of Colorado Boulder (CU Boulder) is committed to fostering an inclusive and welcoming learning, working, and living environment. CU Boulder will not tolerate acts of sexual misconduct (harassment, exploitation, and assault), intimate partner violence (dating or domestic violence), stalking, or protected-class discrimination or harassment by members of our community. Individuals who believe they have been subject to misconduct or retaliatory actions for reporting a concern should contact the Office of Institutional Equity and Compliance (OIEC) at 303-492-2127 or cureport@colorado.edu. Information about the OIEC, university policies, anonymous reporting, and the campus resources can be found on the OIEC website. Please know that faculty and graduate instructors have a responsibility to inform OIEC when made aware of incidents of sexual misconduct, dating and domestic violence, stalking, discrimination, harassment and/or related retaliation, to ensure that individuals impacted receive information about options for reporting and support resources.

Religious Holidays

Campus policy regarding religious observances requires that faculty make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required attendance.