ICPSR Session 2: Data Science and Text Analysis

KELSEY SHOUB AND MATTHEW DENNY THURSDAY 29TH JULY, 2021

Contact and Office Hours

- **Instructor for weeks 1 and 3:** Matthew Denny, email: matthewjdenny@gmail.com (pronouns: he/him). He goes by Matt, or Dr. Denny if you prefer.
- **Instructor for weeks 2 and 4:** Kelsey Shoub, email: kelsey.shoub@gmail.com (pronouns: she/her). She goes by Kelsey, or Dr. Shoub, if you prefer.
- **Course Teaching Assistants:** Bethany Leap and Jared Edgerton
- **Office Hours:** TBD, please check course webpage.
- **Instructor Websites:** `https://www.kelseyshoub.com/` and `http://www.mjdenny.com`.

Course Overview

The social sciences have experienced an explosion of interest in data science, with a special emphasis on the quantitative analysis of textual data, over the past decade. This interest has not been misplaced, as data science techniques and text-as-data studies have revealed important social phenomena across a wide range of disciplines, and at a previously inconceivable scale. Yet, as with any emerging fields, there are many poorly documented pitfalls, as well as significant technical and theoretical challenges. This course seeks to arm its participants with the theoretical background, practical experience, and technical capacity to pursue cutting edge social science research using text data. This course is designed to cover the key technical aspects of conducting research with text data with a special emphasis on techniques drawn from the emerging field of data science: from data collection and preprocessing, through to description and inferential analysis. Students are expected to have some experience with R programming, and some background in statistical analysis for the social sciences.

Goals

By the end of the course, participants should possess the basic skills necessary to perform most data collection and management tasks they are likely to encounter over the course

of a research project intended for publication as a scholarly journal article. They should also have the skills and experience to collect and preprocess digital text data, and apply a number of standard text analysis techniques to their data.

Prerequisites

Some background in statistical inference, regression analysis, etc. with some basic R programming experience a plus.

- This course is intended for graduate students in the social sciences with some background in statistical analysis for the social sciences. We expect that you will have some experience with hypothesis testing, regression analysis and some basic background in statistics, as we do not intend to cover this material during the course. It is important that you have this background so that you will be able to follow along with the statistical models for text we cover in the later part of the course.
- We do not require any programing experience, although having some experience with R will certainly make the early part of the course less challenging. The first week of the course will be focussed on getting everyone up to speed on R programming, including those with no background, so if you do not have experience using R or with programming, this course is still for you. For those with some experience using R, you may find the first week to week and a half to be more review, before we dive into text analysis.

Homework

Students will have a weekly homework assignment (for a total of four assignments), typically assigned on Wednesday and due on Monday. A writeup with instructions for each assignment will be posted on the course website. Each student will submit their assignment on Monday and we will leave time for discussion of the assignments. More details will be provided in the assignment, but the writeup will typically be 2-5 pages, single spaced (including figures). Below is a brief overview of the four assignments for the class:

- **Programming Assignment**: Your first assignment will involve several coding exercises. You will submit an R script with your solutions, and comment in that R script as to how you came up with the solution.
- **Data Collection and Description**: Your second assignment will involve collecting a corpus of documents that you wil analyze over the rest of the course. We will go over a variety of techniques for collecting these data in class. You will be expected to provide a writeup on the dataset you selected (including some basic descriptive statistics), and to provide a copy of your dataset in a format we will describe in the homework assignment. You may collect any corpus you like, and we will provide examples for those who do not have one in mind. Your corpus should satisfy the following criteria:

  - Your dataset should be 100+ documents.
  - Your dataset should be 100+ pages (30,000+ words) in total. This is important because some of the algorithms we use in this class require a reasonably large amount of text to produce actually interpretable results.
  - Your dataset should have one categorical variable and one continuous variable per document for at least 100 documents. You can hand-code these for a subset of documents if you are working with a larger dataset and there is not an easy way to gather these covariates automatically/programatically.

- **Text Preprocessing and Exploration**: In this assignment, you will apply some of the preprocessing and description techniques we go over in the class to the corpus you collected in your previous homework assignment.
- **Text Analysis**: In the final assignment, you will apply some of the inferential methods we discuss in the final week of the class to your corpus.

Grades

Grades will be assigned based on performance on the homework assignments. Each assignment will be graded (0 - no submission, 1 - does not demonstrate comprehension of the methodology, 2 - adequate comprehension demonstrated, 3 - excellent comprehension demonstrated). Grade Scale (based on sum of assignment grades):

- A+ (12)
- A (10-11)
- A- (8-9)
- B+ (6-7)
- B- (<6)

Schedule

More details, along with links to assigned readings and class lab materials will be provided on the course Canvas page. Each class will typically start with a lecture component, followed by a lab. We will post the R scripts for each lab on the course Canvas page so that participants may follow along and adapt the provided code to help with their assignments.

Week 1: Programming and Data Management

This week will be taught by Dr. Denny.

- **Monday**: No class.

- **Tuesday**: Course Introduction, Setting up R, and R Basics.
  - Additional Resources: What we cover in the first class is also written up as part of this basic R tutorial: `http://mjdenny.com/R_Tutorial.html`
- **Wednesday**: Looping and Conditionals, Data I/O and R Packages
  - R Packages: `rio`
- **Thursday**: Data Management
  - R Packages: `statnet`
- **Friday**: Data Management Part 2
  - R Packages: `stringr, tidyr`

Week 2: Web Scraping, String Manipulation and Corpus Creation

This week will be taught by Dr. Shoub.

- **Monday**: Manipulating Strings
  - Required Readings: (1) The introduction for the regular expressions tutorial on Regular-Expressions. (`https://www.regular-expressions.info/tutorial.html`)
  - Recommended Readings: (1) Nield. 2017. "An Introduction to Regular Expressions." O'Reilly. (`https://www.oreilly.com/content/an-introduction-to-regular-expressions/`); (2) Work through the lessons on RegexOne (`https://regexone.com/lesson/introduction_abcs`) to develop a better sense of the logic; (3) For the Tidyverse code and explanation, see Chapter 14 from R for Data Science (`https://r4ds.had.co.nz/strings.html`)
  - R Packages: `stringr`
- **Tuesday**: Web Scraping
  - Required Readings: Wilkerson, John, and Andreu Casas. "Large-scale computerized text analysis in political science: Opportunities and challenges." Annual Review of Political Science 20 (2017): 529-544.
  - R Packages: `httr, stingr`
- **Wednesday**: Web Scraping (Part 2)
  - Required Readings: (1) Barberá and Steinert-Threlkdeld. 2020. "How to Use Social Media Data for Political Science Research." in *The Sage Handbook of Research Methods in Political Science and International Relations* eds Curini and Franzese.; (2) Greene et al. 2020. "Elusive Consensus: Polarization in Elite Communication on the COVID-19 Pandemic." Science Advances.; (3) Twitter's Terms of Service
  - Recommended Readings: (1) Zachary C. Steinert-Threlkeld. 2018. Twitter as Data. Cambridge University Press.; (2) The "Congress Soars to New Heights on Social Media" report by Pew Research Center

  - R Packages: `ROAuth, devtools, ggplot2, maps, streamR, rtweet, readr`

- **Thursday**: Loading, Representing and Describing Text Data
  - Recommended Readings: Quanteda Reference Materials (`https://quanteda.io/articles/pkgdown/quickstart.html`)
  - R Packages: `quanteda, ggplot2, quanteda.textplots, quanteda.textstats`

- **Friday**: Dictionaries
  - Required Readings: (1) Muddiman, Ashley, Shannon C. McGregor, and Natalie Jomini Stroud. "(Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries." Political Communication 36.2 (2019): 214-226.; (2) Young, Lori, and Stuart Soroka. "Affective news: The automated coding of sentiment in political texts." Political Communication 29.2 (2012): 205-231.
  - Recommended Readings: (1) Guo et al. 2016. "Big Social Data Analysis in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling." Journalism and Mass Communication Quarterly.; (2) Dun, Lindsay, Stuart Soroka, and Christopher Wlezien. "Dictionaries, Supervised Learning, and Media Coverage of Public Policy." Political Communication (2020): 1-19.; (3) Quanteda Reference Materials
  - R Packages: `quanteda, quanteda.textstats, ggplot2`

Week 3: Text Preprocessing and Exploration

This week will be taught by Dr. Denny.

- **Monday**: Text Preprocessing
  - Required Readings: (1) Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." Political Analysis 26.2 (2018): 168-189.; (2) Chapter 4 from Manning and Schutze's *Foundations of Statistical Natural Language Process*
  - R Packages: `quanteda, stringr, stopwords, devtools` and `matthewjdenny/preText` from GitHub

- **Tuesday**: Basic NLP: Parts of Speech and Named Entity Recognition
  - Required Readings: (1) Chapter 3 from Manning and Schutze's *Foundations of Statistical Natural Language Process*; (2) Handler, Abram, Matthew Denny, Hanna Wallach, and Brendan O'Connor. "Bag of what? simple noun phrase extraction for text analysis." In Proceedings of the First Workshop on NLP and Computational Social Science, pp. 114-124. 2016.; (3) Justeson, John S., and Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text." Natural language engineering 1, no. 1 (1995): 9-27.

  - For additional help on setting the Java home: (1) `https://www.r-statistics.com/2012/08/how-to-load-the-rjava-package-after-the-error-java_home-cannot-be-determined-from-the-registry/`; (2) `https://stackoverflow.com/questions/9120270/how-can-i-install-rjava-for-use-with-64bit-r-on-a-64-bit-windows-computer`; (3) `https://conjugateprior.org/2014/12/getting-r-and-java-18-to-work-together-on-osx/`
  - R Packages: `slanglab/phrasemachine/R/phrasemachine` from GitHub

- **Wednesday**: Term Category Associations
  - Required Readings: (1) O'Connor, Brendan. "MiTextExplorer: Linked brushing and mutual information for exploratory text data analysis." In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 1-13. 2014.; (2) 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, no. 4 (2008): 372-403.; (3) Denny "Revisiting Fightin' Words"
  - R Packages: `slam, ggplot2` and `matthewjdenny/SpeedReader` grom GitHub

- **Thursday**: Text Reuse
  - Required Readings: (1) Wilkerson, John, David Smith, and Nicholas Stramp. "Tracing the flow of policy ideas in legislatures: A text reuse approach." American Journal of Political Science 59, no. 4 (2015): 943-956.; (2) Casas, Andreu, Matthew J. Denny, and John Wilkerson. "More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process." American Journal of Political Science 64, no. 1 (2020): 5-18.; (3) Denny "Assessing Editing Patterns Across Document Versions"

- **Friday**: Word Embeddings
  - Required Readings: (1) Rodriguez, Pedro, and Arthur Spirling. "Word Embeddings: What works, what doesn't, and how to tell the difference for applied research." (2021).; (2) Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pp. 3111-3119. 2013.; (3) Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences.
  - Recommended Readings: Rodriguez, Spirling, and Stewart. *Working Paper.* "Embedding Regression: Models for Context-Specific Description and Inference"
  - R Packages: `"text2vec","Rtsne","ggrepel"`

Week 4: Text Analysis

This week will be taught by Dr. Shoub.

- **Monday**: Topic Models
  - Required Readings: (1) Blei, D. (2012). Probabilistic Topic Models. Communications of the ACM, 55(4), 7784.; (2) Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A. (2011). Optimizing semantic coherence in topic models. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (2), 262272
  - Recommended Readings: Fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008." American Sociological Review 82.5 (2017): 879-909
  - R Packages: `quanteda, SpeedReader, ggplot2, MCMCpack, stm, quanteda.corpora, cowplot, igraph`
- **Tuesday**: Topic Models (Part 2)
  - Required Readings: (1) Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209228.; (2) Roberts, M. E. et al. (2014). Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4), 10641082.; (3) Roberts et al. (2014) stm: R Package for Structural Topic Models
  - Recommended Readings: Wallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D. (2009). Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09, (4), 18.
  - R Packages: `quanteda, stm, rtweet, ggplot2`
- **Wednesday**: Supervised Learning
  - Required Reading: (1) Jukra et al. RTextTools: A Supervised Learning Package for Text Classification `https://research.vu.nl/ws/portalfiles/portal/828993/310585.pdf`; (2) Chapter 4 from Jacob Eisenstein. (2018). Natural Language Processing. `https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes-june-1.pdf`
  - Recommended Readings: (1) Chapter 4 through to the end of part 4.3 from: James et al. (2017) An Introduction to Statistical Learning. `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.`; (2) An extensive vignette for glmnet: `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`
  - R Packages: `quanteda, rtweet, ggplot2, glmnet, caret, ROCR, pROC`

- **Thursday**: Supervised Learning (Part 2)
  - Supplemental Resources: An overview of SVMs and regression trees: `CURINI_FRANZESE_V2_Chp56_final.pdf`; (2) More information on the caret package (`https://topepo.github.io/caret/index.html`); (3) An easy walk through of how to use caret (`https://towardsdatascience.com/create-predictive-models-in-r-with-caret-12baf9941236`)
  - R Packages: `xgboost`, `quanteda`, `rtweet`, `ggplot2`, `glmnet`, `caret`, `ROCR`, `pROC`