PLSC 21510/31510: Introduction to Text as Data for Social Science

Spring 2022

Warning!: This syllabus (content, order, etc) may change.

- Time: Wednesdays, 9:30 AM - 12:30 PM
- Location: Pick Hall 506
- Instructor: Rochelle Terman, rterman@uchicago.edu
  - Office hours: Mondays 3:30-5:00pm. Sign up here. All office hours are held on Zoom.
- TA: Maya Van Nuys, mvannuys@uchicago.edu
  - Office hours: Fridays 11:00am-1:00pm. All office hours are held on Zoom.

Course Description

Social scientists increasingly use large quantities of text-based data to address problems in industry and academy. This course provides students with an overview of popular techniques for collecting, processing, and analyzing text data from a social science perspective. We will first learn how to collect text data from a variety of sources, including application programming interfaces (APIs) and web-scraping. The second portion of the class provides an overview of popular methods to analyze text data, including sentiment analysis, topic models, supervised classification, and word embeddings. The course is applied in nature. While many of the techniques we discuss have their origins in computer science or statistics, this is not a CS or statistics course. Ultimately, the goal is to introduce students to modern techniques for computational text analysis and help them apply these methods to their own research.

Prerequisites

At the very least, students should have a first class in statistics and/or inference under their belt before taking this course. Basic knowledge of probability, densities, distributions, statistical tests, hypothesis testing, the linear model, generalized linear models, and maximum likelihood is assumed.

We will also assume basic knowledge of the `R` programming language. Students with no prior experience with `R` will struggle with the assignments. Students must have basic computer skills, must be familiar with their computer's file system, and must be comfortable entering commands in a command line environment.

Please check with the instructor if you unclear as to whether you are qualified for this course.

Assessment

Your final grade will be based on:

1. **Assignments (50%)**:

   The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for applied work. The assignments will be frequent, but each of them should be fairly short.

   You are encouraged to work in groups, but the work you turn in must be your own. It is not acceptable to submit homework as a group or to turn in copies of the same code or output. While you are encouraged to use the internet to help you debug, do not copy and paste large chunks of code that you do not understand. Remember, the only way you actually learn how to write code is by writing code!

   Portions of the homework in `R` should be completed using R Markdown, a markup language for producing well-formatted documents with embedded R code and outputs. To submit your homework, knit the R Markdown file to PDF and then submit the PDF file through Canvas (unless otherwise noted).

2. **Class Participation (25%)**

   The class participation portion of the grade can be satisfied in one or more of the following ways:

   - Attending the lectures.
   - Asking and answering questions in class.
   - Attending office hours.
   - Contributing to class discussion on the Piazza site.
   - Collaborating with the computing community by attending a workshop or meetup, submitting a pull request to a GitHub repository (including the class repository), answering a question on StackExchange, or other involvement in the social computing/digital humanities community.

3. **Final Project (25%)**

   Students have two options for class projects:

   1. **Data project**: Use the tools we learned in class on your own data of interest. Collect and/or clean the data, perform some analysis, and visualize the results. Post your reproducible code on GitHub.
   2. **Method project**: Create a tutorial on a method or tool we did not cover in class. Ideas include: tidytext, random forests, bagging/boosting, part-of-speech tagging, named entity recognition, etc. Post it on GitHub.

   Students are required to write a short proposal by **May 4** (no more than 2 paragraphs) in order to get approval/feedback from the instructors.

   Project materials (i.e., a GitHub repository) will be due by end of day on **TBD**. We will specify submission details in class.

   On **TBD**, we will have a **lightning talk session** where students can present their projects in a maximum 5-minute talk.

Activities and Materials

Class Format

Classes will follow a "workshop" style, combining lecture, demonstration, and coding exercises. We envision the class to be as interactive/hands-on as possible, with students programming every session. I anticipate spending 1/2 the class lecturing and 1/2 practicing with code challenges. Bring your laptops.

It is important that students **complete the requisite reading** before class.

Course Notes and Code

All classroom materials will be available on GitHub, including slides, code demonstrations, sample data, etc.

Download the materials on your computer by running the following code in RStudio. Note that for this to work, you will need to have tidyverse installed.

```
# Install tidyverse if you have not already done so.
# install.packages("tidyverse")

library("usethis")
use_course("https://github.com/rochelleterman/TAD-S22/archive/main.zip")
```

Canvas and Piazza

We will use Canvas for turning in assignments and distributing readings.

We will use Piazza for communication (announcements and questions). You should ask questions about class materials and assignments through Piazza so that everyone can benefit from the discussion. We encourage you to respond to each other's questions as well. Questions of a personal nature can be emailed to us directly.

Find our Piazza class signup link at: http://piazza.com/uchicago/spring2022/plsc2151031510

Tech Requirements and Software

See the Installation Page page for detailed information on the software we will be using. Please download and install the required software before the first class.

If you have difficulties installing, please post a question on Piazza with details on what you are trying to install, what actions you took, any error messages, etc.

Policies

Late Policy and Incompletes

All deadlines are strict. Late assignments will be dropped a full letter grade for each 24 hours past the deadline. Exceptions will be made for students with a documented emergency or illness. I will only consider granting incompletes to students under extreme personal/family duress.

Academic Integrity

All work in this course is governed by the University of Chicago's standards for academic integrity. Students are responsible for familiarizing themselves with, and following, university policies regarding proper student conduct. Being found guilty of academic dishonesty is a serious offense and may result in a failing grade for the assignment in question, and possibly for the entire course.

Professionalism

This class is committed to creating an inclusive environment in which everyone can participate regardless of background. Everyone is expected to follow basic norms of professional intellectual exchange. Please be respectful in all your communications, including class discussions. Misconduct based on race, gender, religion or sexual orientation are not acceptable.

Accessibility

Students who have Letters of Accommodation in this class are encouraged to contact me as early in the semester as possible to ensure that such accommodations are implemented in a timely fashion. For those without Letters of Accommodation, assistance is available to eligible students through Student Disability Services. Please contact Student Disability Services at 773-702-6000 or disabilities@uchicago.edu for more information. All discussions will remain confidential. If you have a particular concern surrounding inclusiveness or accessibility, please see me as soon as possible so that we can make proper arrangements.

Communications

If you have a question that will require more than one minute to answer, I would much rather answer that question on Piazza so that all students can benefit from the exchange. Also be advised that I do not check my email on the weekends.

If you have a preferred name or pronoun you think I should know about, please let me know.

Acknowledgments

Portions of this course adapt materials from the following organizations and individuals. Thank you!

- Pete Cuppernull
- Justin Grimmer
- Laura Nelson
- Allen Riddell
- Arthur Spirling
- Hadley Wickham

Class Schedule

Week 1: Intro and String Data

- About this class
- About text as data
- String data and regex

Readings

- Grimmer, Justin, and Brandon M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21, no. 3 (July 1, 2013): 267–97.

Week 2: Collecting Texts

- Web APIs
- Webscraping

Readings

- Densmore, James. "Ethics in Web Scraping." Toward Data Science, July 23, 2019. https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01.

Week 3: Representing Texts

- Vector-space model of texts
- Preprocessing recipes
- Bag of words assumption and alternatives
- Sparseness
- Non-English/multiple languages

Readings

- Lucas, Christopher,Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. "Computer-Assisted Text Analysis for Comparative Politics." Political Analysis 23, no. 2 (2015): 254–77.
- 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, no. 2 (2018): 168–89.

Week 4: Describing Texts

- Word distributions and Zipf's law
- Co-occurances, collocations, keywords in context
- Lexical diversity, complexity, readability

Readings

- Ban, Pamela, Alexander Fouirnaies, Andrew B. Hall, and James M. Snyder. "How Newspapers Reveal Political Power." Political Science Research and Methods 7, no. 4 (2019): 661–78.
- Hengel, Erin, and University College London. "Publishing While Female." Working Paper.
- Spirling, Arthur. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." The Journal of Politics 78, no. 1 (2016): 120–36.

Week 5: Supervised Methods 1

- Supervised vs. unsupervised methods
- Dictionary methods
- Sentiment analysis
- Distinctive / discriminating words
- Scaling using 'wordscores'

Readings

- Mosteller, Frederick, and David L. Wallace. "Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers." Journal of the American Statistical Association 58, no. 302 (1963): 275–309.
- 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.
- Dodds, Peter Sheridan, and Christopher M. Danforth. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." Journal of Happiness Studies 11, no. 4 (2010): 441–56.

Week 6: Supervised Methods 2

- Intro to classification using machine learning
- Readme
- Popular classifiers (Naive bayes, Support vector machines, etc).
- Evaluation (precision, recall)

Readings

- Hopkins, Daniel, and Gary King. "A Method of Automated Nonparametric Content Analysis for Social Science." American Journal of Political Science 54, no. 1 (2010): 229–47.
- King, Gary, Jennifer Pan, and Margaret E. Roberts. "How Censorship in China Allows Government Criticism but Silences Collective Expression." American Political Science Review 107, no. 02 (2013): 326–43.
- Katagiri, Azusa, and Eric Min. "The Credibility of Public and Private Signals: A Document-Based Approach." American Political Science Review 113, no. 1 (2019): 156–72.

Week 7: Unsupervised Methods 1

- Distance and similarity
- Principal components and multidimensional scaling
- Latent semantic analysis
- Scaling using 'wordfish'
- Fully automated clustering (kmeans and hierarchical)

Readings

- Nelson, Laura K. "Computational Grounded Theory: A Methodological Framework." Sociological Methods & Research 49, no. 1 (2020): 3–42.

Week 8: Unsupervised Methods 2

- LDA and topic models
- Model selection and evaluation

Readings:

- Blei, David M. "Probabilistic Topic Models." Communications of the ACM 55, no. 4 (2012): 77–84.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science 58, no. 4 (2014): 1064–82.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. "How to Analyze Political Attention with Minimal Assumptions and Costs." American Journal of Political Science 54, no. 1 (2010): 209–28.
- Terman, Rochelle. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of U.S. News Coverage." International Studies Quarterly 61, no. 3 (2017): 489–502.

Week 9: Embeddings

- Intro to neural nets
- Word2Vec
- Doc2Vec
- The frontier

Readings

- Kozlowski, Austin C., Matt Taddy, and James A. Evans. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." American Sociological Review 84, no. 5 (2019): 905–49.
- Nelson, Laura K. "Leveraging the Alignment between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century US South." Poetics 88 (2021): 101539.