

# PLSC 498: Text as Data

**Fall 2022**

Tuesday, Thursday 4:35-5:50pm

Advanced Analytics Course  
Penn State University  
Willard Bldg 260

## Instructors

Professor: Kevin Munger

[kmm7999@psu.edu](mailto:kmm7999@psu.edu)

Office hours: Schedule on website (Wednesdays 2-4pm)

Teaching Assistant: Valerie Li

[valerie.li@psu.edu](mailto:valerie.li@psu.edu)

Office hours: Wednesday 1:00pm - 2:00pm

Pond Lab room 224

Current version (subject to change!): September 8, 2022

## Course Overview

The availability of text data has exploded in recent times, and so has the demand for analysis of that data. This course introduces students to the quantitative study of text from a social science perspective, with particular attention paid to political science. This course is applied; we hope to acquire the skills needed to implement some of the advanced techniques developed by others.

We begin by explaining how text can be modeled statistically, and how different texts can be fruitfully compared. We then move to both supervised and unsupervised techniques in some detail, before dealing with some 'special topics' that arise in particular lines of social science research. Ultimately, the goal is to help students conduct their own text as data research projects and this class provides the foundations on which more focused, technical research can be built.

This course is an amalgamation of other Text as Data course I've taken, helped administer or taught. I'm grateful for the commitment to sharing teaching materials in the text as data community, and would particularly like to thank Arthur Spirling, Ken Benoit, Pablo Barberà, Leslie Huang, and Pedro Rodriguez for producing material that I have used in designing this

course. My materials are similarly free to use for anyone interested in teaching courses like this in the future.

## Prerequisites

There are no official requirements for this course, but you will find it difficult without a baseline familiarity with statistics. We will be implementing a number of advanced statistical models in this course, but won't spend too much time on the details of how they're derived.

The coding portion of the course will take place in R, and some knowledge of R is required; you will have to become comfortable with running basic functions throughout the course. In particular, all of the problem sets will need to be completed with R, and the final project will involve analyzing data in R.

If you have zero experience with coding in general, this portion of the course will be **VERY** challenging.

## Course Components and Grading

- **Homeworks:** There will be a series of problem sets throughout the course to ensure that you're keeping up with the instruction and mastering the material. All problem sets will need to be completed in R. Some of these problem sets will entail developing your final projects. (40%)
- **In-Class pRactice:** To give more intermediate coding feedback and ensure that everyone is keeping up to date with the lessons in R, I'll be doing completion checks on some of the in-class R practice. We'll be going over all the answers in class, so this shouldn't be a huge challenge, but just getting everything working on your own machine is a valuable exercise. Each of these that you complete will be worth 10 points towards your average for this portion of the grade. (10%)
- **Final project:** The capstone for this course, the final project will be something we work on all semester. It's important that you have find a question that you're interested in answering and then figuring out how to answer it with text analysis. Further details will be provided during the course, but the purpose of the project is to demonstrate that you have gained an understanding of the types of questions that can be answered using text as data and that you have the skills to provide such an answer. (50%; 10% will be for the 2-page prospectus due November 4, 40% for the final project due December 12.)

## Readings

There is no required textbook for the course; indeed, no appropriate textbook exists. All readings are available either through links on the syllabus or through the course GitHub. Homeworks and in-class assignments can be submitted through Coursera.

## thuRsdays

Each thuRsdays we will be working through code in R. My hope is to have these lessons sync up with the substantive material earlier in the week, but we might end up slower if I feel like we need extra practice with the fundamentals of R.

This will require you to share your screen with me during class. It's essential that I be able to watch you coding and give frequent feedback. I tend to think of coding as partially a *physical* skill, and like a golf coach or piano instructor, I need to watch your form improve in order to create better outcomes.

## Required Syllabi Statements

**ACADEMIC INTEGRITY STATEMENT** Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, the University's Code of Conduct states that all students should act with personal integrity, respect other students' dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts. Academic integrity includes a commitment by all members of the University community not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others.

**DISABILITY ACCOMMODATION STATEMENT** Penn State welcomes students with disabilities into the University's educational programs. Every Penn State campus has an office for students with disabilities. Student Disability Resources (SDR) website provides contact information for every Penn State campus (<http://equity.psu.edu/sdr/disability-coordinator>). For further information, please visit the Student Disability Resources website (<http://equity.psu.edu/sdr/>). In order to receive consideration for reasonable accommodations, you must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: See documentation guidelines at (<http://equity.psu.edu/sdr/guidelines>). If the documentation supports your request for reasonable accommodations, your campus disability services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early as possible. You must follow this process for every semester that you request accommodations.

**COUNSELING AND PSYCHOLOGICAL SERVICES STATEMENT** Many students at Penn State face personal challenges or have psychological needs that may interfere with their academic progress, social development, or emotional wellbeing. The university offers a variety of confidential services to help you through difficult times, including individual and group counseling, crisis intervention, consultations, online chats, and mental health screenings. These ser-

vices are provided by staff who welcome all students and embrace a philosophy respectful of clients' cultural and religious backgrounds, and sensitive to differences in race, ability, gender identity and sexual orientation. Counseling and Psychological Services at University Park (CAPS) (<http://studentaffairs.psu.edu/counseling/>): 814-863-0395 Counseling and Psychological Services at Commonwealth Campuses (<http://senate.psu.edu/faculty/counseling-services-at-commonwealth-campuses/>) Penn State Crisis Line (24 hours/7 days/week): 877-229-6400 Crisis Text Line (24 hours/7 days/week): Text LIONS to 741741

EDUCATIONAL EQUITY/REPORT BIAS STATEMENT Penn State takes great pride to foster a diverse and inclusive environment for students, faculty, and staff. Consistent with University Policy AD29, students who believe they have experienced or observed a hate crime, an act of intolerance, discrimination, or harassment that occurs at Penn State are urged to report these incidents as outlined on the University's Report Bias webpage (<http://equity.psu.edu/reportbias/>)

## Schedule

### Week of August 22: Introductions

On Tuesday, we'll walk through the class and get to know each other, and then Thursday we'll have the first R session.

If anyone has questions about the course (if you're concerned about pre-requisites, for example), we can find some time to talk during my office hours.

### Week of August 29: Representing Text

- Transforming a document into text data
- Feature selection and representation

### Week of September 5: Representing Text 2

- Pre-processing: Stemming and Stopping
- Bag of Words
- Sparseness

### Week of September 12: APSA – First Homework, Catch-Up

### Week of September 19: Getting Data

- Webscraping
- APIs

- Twitter

### **Week of September 26: Descriptive Inference**

- Word distributions: Zipf's law
- Co-occurrence and collocations
- Key words in context
- Similarity measures

### **Week of October 3: Descriptive Inference 2**

- Lexical diversity
- Sophistication/complexity
- Linguistic style and author attribution

### **Week of October 10: Supervised Techniques 1**

- Dictionary approaches
- Sentiment analysis and LIWC
- Event extraction

### **Week of October 17: Supervised Techniques 2**

- Classification of documents
- Evaluation of techniques: precision, recall
- Naive Bayes classification
- Ideological scaling with 'wordscores'

### **Week of October 24: Supervised Techniques 3**

- Basics of machine learning

### **Week of October 31: Supervised Techniques 4**

- k-NN
- Trees and Random Forests

### **Week of November 7: Unsupervised Techniques 1**

- Fundamentals of Unsupervised Learning
- Data reduction
- Clustering

### **Week of November 14: Unsupervised Techniques 2**

- Parametric scaling of political text
- Count models: ‘wordfish’
- Latent Dirichlet Allocation and topic modeling
- Model selection/choosing  $k$

### **Week of November 21: NO CLASS THANKSGIVING**

### **Week of November 28: Personalized/group time to work on final project**

- Final data collection / analysis

### **Week of December 5: Final project presentation and feedback**

- In-class presentations