PLSC 497: Text as Data

Spring 2021

Tuesday, Thursday 1:35-2:50pm

Advanced Analytics Course
Penn State University
Zoom: `https://psu.zoom.us/j/91042700760`

Instructors

Professor: Kevin Munger
`kmm7999@psu.edu`
Office hours: Schedule by email (usually Tuesdays)

Teaching Assistant: Angel Villegas Cruz
`amv7518@psu.edu`
Office hours: Friday 11:00 - 12:00
Zoom: https://psu.zoom.us/j/98203528709

Current version (subject to change!): February 25, 2021

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 while no 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 may prove to be the most challenging portion of the course, but if you're willing to put in the work, you'll be able to succeed.

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 March 16, 40% for the final project due May 5)

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. I will not be using Canvas.

Using Github to manage code is an essential part of the toolkit for putting these skills to use, and while it is not intuitive at first, it is necessary for this course.

thuRsdays

Each thuRsday 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. I won't ever require you to have your laptop camera on (although it is encouraged — it makes my job easier and more friendly!), but 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 coun-

seling, crisis intervention, consultations, online chats, and mental health screenings. These services 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 January 19: 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 on Tuesday.

Zoom Recordings (email if you don't have password):

```
https://psu.zoom.us/rec/share/4QoRx_pEhvkwEZH7i4izda1yFjH5agcNpipd612nyWxI7La6AoVw3FhK3fH_
GWrC.hu5zutBDxnaDKefO
```

```
https://psu.zoom.us/rec/share/7Gvk275kwQ_vGAF04bCvdILMUOwkh494joXdYFKz2-xu7kp2gyKmkjK5NNphqU
Ueb7t02U3qw-CE32
```

Week of January 26: Representing Text

- Transforming a document into text data

- Feature selection and representation

```
https://psu.zoom.us/rec/share/o0rO7u3BGFUIL0bXKYOhqYoE5PkO3C956meguxlabz16vusMMKzccvpr4_
QSrn1d.SaNY3JYNKLz-1zrv
```

```
https://psu.zoom.us/rec/share/vq_DfSXT2XDuyu0SDFZTODhICWikJtGXZwVbNXppz01sRNBqk7qKAqX7jjJJNI
WD6KYnDyQC5mFArm
```

Week of February 2: Representing Text 2

- Pre-processing: Stemming and Stopping

- Bag of Words

- Sparseness

    https://psu.zoom.us/rec/share/aQZs29zTOwLMcluwlfehWrAAhc-iQmj5HQCNE2S2DW7Pfsd_VHqF9WRwYrAyQ(
gnJ74RqClOk8ni_v

    https://psu.zoom.us/rec/share/JQi5GfgMZnpDO9KviPhK8e5aQBgyggTdaYNag5kV6OhnQL8Oswfiwe4REeFfzS
Db3nZDN9nWUhQG3Y

Week of February 9: Descriptive Inference

No class on Tuesday for Wellness day

- Word distributions: Zipf's law

- Co-occurance and collocations

- Key words in context

- Similarity measures

    https://psu.zoom.us/rec/play/CncDCP5D8aRb64sNzWDltY2duTkmvjcgzNkZ7hg68GIFkXrArtGXVMEnfOHn404
VKmaUlwiE7XxfbDX

Week of February 16: Descriptive Inference 2

- Lexical diversity

- Sophistication/complexity

- Linguistic style and author attribution

    https://psu.zoom.us/rec/share/O6uZGQJwh778sBPSRBsOL7y82JkQqvfCMO5f3gMshMuplr9FaTmlbJUPzDyNsz
yI8UyBAvna7SU6Pq

    https://psu.zoom.us/rec/share/X1uzd9S1jEuFZNVerSChhm05VdY1ghfgJlzmDSx_z7li1XAkBs76P1Y3dF8Qyt
QhQi4oyRH8J3amOh

Week of February 23: Supervised Techniques 1

- Dictionary approaches

- Sentiment analysis and LIWC

- Event extraction

```
https://psu.zoom.us/rec/share/CyRQHGP-PRPoT-cXVTE1OVeeOObd1l3jDlQdhxmIxH3156BsTJ5h9zTrD_
E808HR.LCUsZL6x7p7m3nB2
```

Lecture part 1: `https://psu.zoom.us/rec/share/GjHjgLEZ-PC-ob6oruTEEP2-s5oVZOUZCcwZTZLY-`
`yjmm-wI8wzEHY48UI6iaDNK.sXdvhfHm2MUb4VA-`

Lecture part 2: `https://psu.zoom.us/rec/share/bfXbikjCwOwObkyPOlSyJNfBGj9mpcZp3ZtN6u-`
`V_3cIAQz_WtjeZsUhUnl35ytb.X3Q7N6SlaArR79vb`

Week of March 2: Supervised Techniques 2

- Classification of documents

- Evaluation of techniques: precision, recall

- Naive Bayes classification

- Ideological scaling with 'wordscores'

Week of March 9: Supervised Techniques 3

No class on Thursday for Wellness day

- Basics of machine learning

Week of March 16: Supervised Techniques 4

- k-NN

- Trees and Random Forests

Week of March 23: Unsupervised Techniques 1

- Fundamentals of Unsupervised Learning

- Data reduction

Week of March 30: Unsupervised Techniques 2

- Clustering

- Parametric scaling of political text

- Count models: 'wordfish'

Week of April 6: Unsupervised Techniques 3

- Plate notation

- Latent Dirichlet Allocation and topic modeling

- Model selection/choosing k

Week of April 13: Catch up and final paper discussion

- Revisit any topic that needs more attention

- Present final project prospectus and get feedback

Week of April 20: Advanced topics

- Structural Topic Models

- Word Embeddings: word2vec

- Video as Data

Week of April 27: Final project presentation and feedback