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!): January 19, 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 April 28th)

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.

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.

Week of January 26: Representing Text

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

Week of February 2: Representing Text 2

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

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

Week of February 16: Descriptive Inference 2

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

Week of February 23: Supervised Techniques 1

- Dictionary approaches
- Sentiment analysis and LIWC
- Event extraction

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 submission and presentation