PLSC 497: Text as Data

Fall 2019

Monday, Wednesday, Friday 10:10-11am

Advanced Analytics Course
Penn State University

Instructor

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

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 offical 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. (40%)
- **Class Presentation:** Once a semester, each student will present a summary of the assigned paper for that week. (Depending on enrollment, we might split some papers between two students.) This will give you exposure to cutting edge research and an opportunity to practice presenting research to an expert audience. (10%)
- **Final project:** The capstone for this course, the final project can be completed in teams of up to two people. 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 22, 40% for the final project due December 19)

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.

Paper Presentations

Each Wednesday, we'll spend the last 20 minutes of class hearing one (or possibly two) students present that week's assigned papers; plan for 15 minutes of presentation and 5 minutes of Q&A. There is relatively little reading for this course, but I want us to take it seriously. We'll see a prototype of the kind of presentation I'm looking for in the first week of class. Note that some (most) of these papers may be too technical for you to understand fully. That's perfectly fine. Our goal is to become comfortable encountering difficult material and learning as much as possible.

fRidays

Each Friday 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. Please be sure to bring your laptops to class on Fridays.

Schedule

Week of August 26: Introductions

The timing of the beginning of the semester is unfortunate. The largest political science conference of the year is this week, so I'll be gone Wednesday-Friday.

On Monday, we'll walk through the class and get to know each other, and I'll run a diagnostic to evaluate everyone's proficiency in R. This will allow me to provide materials at the appropriate level for you to work through on Wednesday and Friday.

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 September 2: Representing Text

No class Monday for Labor Day Holiday.

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

Reading: Grimmer and Stewart 2013 —- introduction to using text as data

Week of September 9: Representing Text 2

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

Reading: Denny and Spirling 2018 — overview of impacts of pre-processing

Week of September 16: Descriptive Inference

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

Week of September 23: Descriptive Inference 2

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

Reading: Benoit, Munger and Spirling 2019 — New method for analyzing textual complexity

Week of September 30: Supervised Techniques 1

- Dictionary approaches
- Sentiment analysis and LIWC
- Event extraction

Reading: Tausczik and Pennebaker 2010 — Understanding psychological meaning of words

Week of October 7: Supervised Techniques 2

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

Reading: Hopkins and King 2010 — Adjusting classification methods for social science applications

Week of October 14: Supervised Techniques 3

- Basics of machine learning

Reading: Domingos 2012 — Introduction to machine learning for classification

Week of October 21: Supervised Techniques 4

- k-NN
- Trees and Random Forests

Reading: Jones and Linder ND — Using random forests for exploratory data analysis

Week of October 28: Unsupervised Techniques 1

- Fundamentals of Unsupervised Learning
- Data reduction

Reading: Jones and Linder ND — ???

Week of November 4: Unsupervised Techniques 2

- Clustering
- Parametric scaling of political text
- Count models: ‘wordfish’

Reading: Grimmer and King 2011 — Comparing clustering approaches

Week of November 11: Unsupervised Techniques 3

- Plate notation
- Latent Dirichlet Allocation and topic modeling
- Model selection/choosing k

Reading: Change et al 2009 — How humans evaluate topic models

Week of November 18: Catch up and final paper discussion

- Revisit any topic that needs more attention
- Present final project prospectus and get feedback

Week of November 25: No Class, Thanksgiving break

Week of December 2: Unsupervised Techniques 4

- Structural Topic Models
- Word Embeddings: word2vec

Reading: Roberts et al 2014 — Applying STM to survey responses

Week of December 9: Special Topics

- Bursts and memes
- Plagiarism detection and text reuse
- Video as data