POSC 207: Quantitative Text Analysis for the Social Sciences

Loren Collingwood

Fall, 2020

E-mail: loren.collingwood@ucr.edu
Web: **https://github.com/lorenc5/POSC-207**
Office Hours: T 2-5pm
Class Hours: T 9-11:50am
Office: NA
Class Room: https://ucr.zoom.us/j/92778212788?pwd=bFdjUjBZVWd1OXNIakpXdE9JR3FLUT09
Meeting ID: 927 7821 2788
Passcode: 87863

Course Description

This course investigates how to use digitized texts – news articles, speeches, laws, press releases, party manifestos/platforms, transcripts, open-ended surveys, Tweets, etc. – as sources of data for social science research.

We begin with overviews of the "text as data" field in political science – which is heavily influenced by computer science branch of natural language processing (NLP). The idea is to get you data as a poor graduate student for free that you can then use in your own research to answer questions of theoretical interest.

We then discuss theory/mechanics of converting text into data. This will include topics like preprocessing text and related NLP tasks (e.g., stemming, tokenizing) and representing text as data (e.g., bag-of-words, measures of association), etc. Text data is often "messy" so handling that will be a large part of this course (e.g., web scraping, file encodings, file formats, extracting only relevant text from strings, etc.).

We'll then turn to the major approaches to measuring social science concepts with textual data, including rule-based methods, supervised learning from human-coded or known examples, and un-supervised methods. As we go, we will discuss particular measurement objectives like classification, scaling, topic modeling, and analysis of sentiment and stance, as well as ways of validating our models.

Depending on time, student interest and capacity, we may learn about the neural network / deep learning approach that has come to dominate NLP in recent years.

The course will assume students have some graduate level work in statistical inference, quantitative social science methodology, or machine learning, and at least know what R is but ideally

some experience with R.

Required Materials

- All readings are posted to course website

Course Objectives

Successful students will learn how to:

1. Improve R programming capacity
2. Webscrape and organize their own textual data
3. Select appropriate text method to analyze data
4. Use data to answer/assess theoretical questions of interest
5. Be confident moving forward they can use whatever text-based methods they choose

Course Structure

Class Structure

Lecture (1/3), Discussion (1/3), Code (1/3)

Assessments

1. Weekly assignments
2. Research Design and Analysis
3. Final Presentation

Research Design/Analysis and Final Presentation

All students will turn in a research design/analysis of a project of their choice. If ongoing project, text as data methods must be incorporated in addition to existing design.

Grading Policy

- **50%** of your grade will be determined by weekly assignments
- **25%** of your grade will be determined by research design and analysis
- **25%** of your grade will be determined by final presentation

Course Policies

During Class

Log in on time. The first 5-10 minutes we might do some sort of "ice-breaker" just to get it rolling in the zoom land. Ask questions whenever you have them and I will try to answer as best I can. I will often have code and data available in advance. So go to the course website and download anything in advance of each class.

Attendance Policy

Attendance is expected in all classes. Valid excuses for absence will be accepted before class, please just email me in advance.

Policies on Incomplete Grades and Late Assignments

If throughout the quarter you suspect you will not be able to complete the course or hand in all the material, please discuss the possibility of an Incomplete with me. Turn all assignments in on time before the class starts. Extensions may be granted in extenuating circumstances.

Academic Integrity and Honesty

Students are required to comply with the university policy on academic integrity. In effect, in this class, for all papers and research designs you generate, while your work can be based on others', the content must be solely yours.

Accommodations for Disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with the Disability Services Office and let me know in advance.

Discrimination based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation will not be tolerated. Harassment of any person (either in the form of quid pro quo or creation of a hostile environment) based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation also is not be tolerated. Retaliation against any person who complains about discrimination is also prohibited.

Schedule and weekly learning goals

The schedule is tentative and subject to change. The learning goals below should be viewed as the key concepts and code you should grasp after each week, and also as a study guide before each exam, and at the end of the semester. Each exam will test on the material that was taught up until 1 week prior to the exam (i.e. vorticity will not be tested until exam 2). The applications in the second half of the semester tend to build on the concepts in the first half of the semester though, so it is still important to at least review those concepts throughout the semester.

- **Week 01, 10/05 - 10/09** Topic 1 - Intro, Text Harvesting, and Cleaning
  - Use package Rvest, understand selector gadget
  - Scraping Twitter (currently problem with api)
  - Handling media data from Nexus or Proquest
  - Hitting NYT API
  - Possible Special Guest: Sean Long
  - Readings: Wilkerson and Casas (2017); Grimmer and Stewart (2013)
  - Assignment: Develop a text corpus you will use for rest of course.
- **Week 02, 10/12 - 10/16** Topic 2 – Preprocessing and Feature Representation
  - Use quanteda package, possibly preText package
  - Tokenizing, bag of words
  - Document Term Matrix
  - Readings: Denny and Spirling (2017); Diermeier et al. 2011; Jivani 2011
  - Assignment: Convert your corpus into document term matrix, and clean that matrix
- **Week 03, 10/19 - 10/23** Topic 3 – Comparing Texts, Scaling
  - Use textreuse package, insights into RCopyFind package; quanteda package
  - Word co-occurence; Cosine Similarity
  - Text reuse
  - Special Guest: Stephanie DeMora
  - Readings: DeMora, Collingwood, and Ninci 2019; Wilkerson, Smith, and Stramp 2015; Laver, Benoit, and Garry (2003); Lowe (2008); Slapin and Proksch (2008)
  - Assignment: Conduct some type of text comparison/scaling analysis depending on your data.
- **Week 04, 10/26 - 10/30** Topic 4 – Sentiment and Dictionary Methods
  - quanteda package
  - Collingwood example code from CGU
  - Special Guest: Dr. Kassra Oskooii
  - Readings: Young and Soroka (2013); Oskooii, Lajevardi, and Collingwood (2019)
  - Assignment: Develop a dictionary or find a dictionary that makes sense to what you are doing.
- **Week 05, 11/2 - 11/06** Topic 5 – Supervised Machine Learning
  - Use RTextTools package
  - Readings: Collingwood and Wilkerson 2012; Jurka et al. 2013
  - Suggested Readings: Krippendorff, 2004: *Content Analysis: An introduction to its methodology*
  - Assignment: Hand-code a portion of your data (if not already coded), then conduct supervised learning on it and predict onto virgin text. NOTE: you will likely not have enough hand-coded data to do this well but the exercise is useful.
- **Week 06, 11/09 - 11/13** Topic 6 – Unsupervised Machine Learning

  - LDA Model
  - STM Model
  - Readings: Roberts, Stewart, Tingley, 2013; Roberts et al. 2014; Blei 2012; Blei, Ng, and Jordan 2003;
  - Suggested Readings: Bagozzi and Berliner 2018; Berliner, Bagozzi, and Rubin
  - Assignment: Conduct some type of topic model on your data.

- **Week 07, 11/16 - 11/20** Topic 7 – Neural Networks and Deep Learning
  - Special Guest: Dr. Sarah Dreier
  - Readings: Dreier et al. (Working Paper); Willems: https://www.datacamp.com/community/tutorials/keras-r-deep-learning

- **Week 08, 11/23 - 11/27** Topic 8 – Text analysis and causal inference
  - Use packages stm and textmatching (still somewhat new)
  - Readings: Egami et al., 2018; Roberts et al. 2020

- **Week 09, 11/30 - 12/04** Topic 9 – Word Embeddings (Word 2 Vec) and contextualized word embeddings
  - Use packages text2vec; word2vec
  - Readings: Smith 2019; Mikolov et al 2013

- **Week 10, 12/07 - 12/11** Topic 10 – Final Presentations; LC Special: Bayesian Improved Surname Geocoding (BISG)
  - Student presentations, 15-20 minutes each
  - Research Design/Analysis due by end of finals week
  - use package eiCompare
  - Readings: Elliot et al (2008); Imai and Khanna (2016)