Term Information

Effective Term	Autumn 2017

General Information

Course Bulletin Listing/Subject Area	Political Science
Fiscal Unit/Academic Org	Political Science - D0755
College/Academic Group	Arts and Sciences
Level/Career	Graduate
Course Number/Catalog	7781
Course Title	Text as Data
Transcript Abbreviation	TEXT AS DATA
Course Description	Social interaction and conflict comprises the artful (and sometimes painfully in-artful) use of language. This course explores emerging statistical methods for extracting important signals from language stored as text. Topics include text collection and processing; dictionary methods; topic modeling; document clustering; deep learning; and concomitant computational and mathematical challenges.
Semester Credit Hours/Units	Fixed: 3

Offering Information

Length Of Course	14 Week, 12 Week, 8 Week, 7 Week, 6 Week, 4 Week
Flexibly Scheduled Course	Never
Does any section of this course have a distance education component?	No
Grading Basis	Letter Grade
Repeatable	No
Course Components	Seminar
Grade Roster Component	Seminar
Credit Available by Exam	No
Admission Condition Course	No
Off Campus	Never
Campus of Offering	Columbus

Prerequisites and Exclusions

Prerequisites/Corequisites

Exclusions

Cross-Listings

Cross-Listings

Subject/CIP Code

Subject/CIP Code	45.1001
Subsidy Level	Doctoral Course
Intended Rank	Doctoral

Requirement/Elective Designation

The course is an elective (for this or other units) or is a service course for other units

Course Details

Course goals or learning objectives/outcomes

- Students develop an understanding of methodologies used to extract and model processes using text
- Students learn to gather, process, clean and use textual data
- Students develop capacity to employ methods to produce novel scientific insights

Content Topic List

- Finding text, file formats, scraping the web
- Representing text quantitatively
- Supervised Discrimination Methods
- Multidimensional scaling of texts
- Hand-labeled documents
- Supplemental information on Bayesian inference
- Clustering texts
- Topic models
- Neural network and deep learning approaches to text
- Beyond atomic representations

Attachments

- POLITSC_7781_syllabus.pdf: syllabus

(Syllabus. Owner: Smith,Charles William)

Comments

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Smith,Charles William	02/15/2017 12:48 PM	Submitted for Approval
Approved	Herrmann,Richard Karl	02/15/2017 01:04 PM	Unit Approval
Approved	Haddad,Deborah Moore	02/15/2017 03:41 PM	College Approval
Pending Approval	Nolen,Dawn Vankeerbergen,Bernadette Chantal Hanlin,Deborah Kay Jenkins,Mary Ellen Bigler	02/15/2017 03:41 PM	ASCCAO Approval

Political Science 7781—Fall 2017

Analysis of Text as Data

Professor Brice D. L. Acree

Information

- Meeting: M/W 12:45-14:05, Derby 125
- Office: Derby Hall 2126
- Office hours: 11:00-12:00 M/W, 14:30-15:30 W, or by appointment
- Email: acree.11@osu.edu
- Phone: 859.221.1782

Text as a source of data

Politics comprises, in large part, the artful (and sometimes painfully inartful) use of language. Candidates for political office barnstorm for months at a time, making their cases and leveling criticisms of opponents' actions and points of view; legislators draft and debate bills; presidents issue statements; justices hand down legal opinions; information age rabble-rousers peddle conspiracy theories to anyone with an internet connection; and journalists, columnists and bloggers dissect, reframe and disseminate it all to the American public. And those are just a handful of examples from American politics, ignoring the treaties , trade agreements, speeches, and declarations by leaders and media around the world.

For political scholars, the trove of data locked away in transcripts and manuscripts is both a blessing and a curse. The answers to many fascinating questions lie within the written word; yet harnessing the data in ways both efficient and reliable poses considerable methodological challenges.

In political science, we have sought for decades to use text as a source of data. Over the past two decades, we have built our capacity to process, quantify, and model text using computational methods. With increased computing power and advanced statistical methodology, scholars have developed new, powerful, and efficient tools to extract patterns from, and test substantive theories using, text as data.

Objectives

This course has three broad goals:

1. To build student understanding of methodologies used to extract and model processes using text
2. To train students to gather, process, clean, and use textual data in `R`
3. To build student capacity to *employ* methods to produce novel scientific insights.

Prerequisites

This course is intended for advanced graduate students. You should be comfortable with:

- Basic linear algebra, differentiation, integration
- Generalized linear models

- Probability theory
- The R statistical language

Students in political science should have taken the methods sequence in the department. If you have concerns, please contact me.

Support

I am available during my office hours and by appointment. If you have questions, e-mail them to me. If I can answer quickly, I will resolve the issue over e-mail. If the problem is more involved, I may ask you to come to my office.

Please note: if you are e-mailing me regarding programming in R, you will need to include a reproducible code that throws the error. Otherwise it will be difficult for me to help.

Materials

Lecture slides, R scripts, and other sundry instructional materials will be posted to the Carmen site.

Software

You will need to install R and RStudio on your personal computer. Both are free and open source.

- To install R, visit http://www.r-project.org. For instructions on how to install R, see http://socserv.mcmaster.ca/jfox/Courses/R/ICPSR/R-install-instructions.html.
- To install RStudio, visit https://www.rstudio.com/products/rstudio/download/.
- You also need to install the `quanteda` package from Kenneth Benoit (with Kohei Watanabe, Paul Nulty, Adam Obeng, Haiyan Wang, Ben Lauderdale, and Will Lowe). You can install it from `CRAN` or from `Github`. https://github.com/kbenoit/quanteda

Why R? We will be using the R programming language. There are many statistical programs (e.g., Stata, SPSS, SAS), programming environment-languages (e.g., R, Matlab, Julia) and languages (e.g., C, C++, Fortran). R is quite powerful, flexible and popular.

For some advanced applications, you may need to use languages like Python, which is admittedly much better than R for text. (If that is the case, schedule a time to meet with me.) For many applications, however, R will work—and has the advantage of already being taught in our methods sequence.

Text

There is **no required textbook for the course**. Indeed, there is currently no great textbook for analyzing text as data. (There are some in process, but none ready for us.)

Instead, I will post materials for reading to Carmen. Materials in a folder for the current week need to be read before the first class of that week.

An optional book, to which we will refer often, is Manning, Raghavan, and Schutze's 2008 text, ***Introduction to Information Retrieval***. You can find the book for free at http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Expectations

In general, I expect you to put considerable effort into this course. We will cover many topics and, at times, will move quickly through material. If you put your full effort into this course, you will make it through in good shape. By full effort, I mean:

- Ask questions! Clarify points of confusion whenever they arrive.
- Come to office hours! Don't be afraid to seek help with material.
- Be professional! Arrive on time, pay attention in class, be aware of professional etiquette, submit polished work. If you would not show an e-mail/homework/project to your advisor or to a potential employer, don't submit it to me.

Grading

You will be evaluated on three basic criteria:

- Homework: 35 percent
  - 4-6 assignments
- Reactions: 20 percent
  - Two reaction papers
- Project: 35 percent
  - [05 percent] Proposal presentation
  - [20 percent] Paper
  - [10 percent] Presentation
- Participation: 10 percent

Homework

Homework assignments will cover the use of methods taught in class. As such, these assignments will be primarily *applied* rather than theoretical or proof-based. You may still be responsible, however, for some mathematical explanation.

Homeworks will need to be submitted using `RMarkdown`. This allows for seamless integration of (a) easy markup; (b) LaTeXwhen needed; (c) `R` code and output. We will cover `RMarkdown` in the first week of class.

Reactions

Twice during the semester, you will read an assigned article for class, write a short (two single-spaced pages at maximum) reaction paper about the article, and be responsible for discussion in-class of the methods described in the papers.

Project

Your project should address a question related to your own research. I have to baseline criteria for your project: (1) It should address a novel question; and (2) it should be as useful to your career as possible. You should use this project to build on your prior work; or as part of your dissertation; or try to develop it into a publishable paper. Anything less is, in too many ways, a waste of your time.

This comes with a corollary: you should work dilligently on this paper. Leaving your project to the last week will result in a substandard product, a wasted effort on your part, and least importantly, a poor grade.

Obtaining, processing, and modeling text data is *messy*. It does not lend itself to last-minute efforts. Start early, work hard.

The project proposal will be 1-2 pages explaining the question you seek to address; what type of data you have or plan to collect; how you will collect the data; and your strategy for analysis. The paper should be written as a traditional article, complete with at least preliminary analysis and results. The presentation will be conference-style, with 10 minutes for you and 5 minutes for another student serving as discussant.

Participation

I do not take attendance, but I expect you to attend most class sessions. I also expect basic professional decorum—being on time, participating in discussion, et cetera.

Outline

Week 1: 22 - 25 August

Introduction and syllabus.

Lecture: The basics: what do we mean by analyzing text?

Lab: Introduction to `R`, `RMarkdown`

Week 2: 28 August - 01 September

Lecture: Finding text; file formats; scraping the web; text and theory

Lab: Using `R` to scrape text

Week 3: 04 - 08 September

Lecture: Representing text quantitatively; processing text, and the vagaries of forking paths

Lab: Using `quanteda` to process text

No class 04 September: Merged class/lab on 06 September

Week 4: 11 - 15 September

Lecture: Supersupervised methods (i.e., dictionaries)

Lab: Dictionaries in `R`

Week 5: 18 - 22 September

Lecture: Supervised discrimination methods, including penalized regression, inverse regression

Lab: Discovering disriminating words/phrases in `R`

Week 6: 25 - 29 September

Lecture: Presentation of project proposals

Lecture: Words in multiple dimensions: texts and linear algebra

Week 7: 02 - 06 October

Lecture: Multidimensional scaling of texts

Lab: Principal components, factor analyis in R

Week 8: 09 - 13 October

Lecture: Supervised learning: hand-labeled documents, obtaining gold-standard labels

Lecture: Supervised learning: building and evaluating several classifiers

Week 9: 16 - 20 October

Lab: Project workshop

Lab: Document classification (and misclassification) in R

Week 10: 23 - 27 October

Lecture: Supervised learning: Naïve Bayes; supplemental information on Bayesian inference, Bayes' rule, and the probability simplex

Lab: Programming a naïve Bayes routine

Week 11: 30 October - 03 November

Lecture: Unsupervised methods: Clustering texts

Lab: Clustering methods in R

Week 12: 06 - 10 November

Lecture: Unsupervised methods: introduction to topic models

Lecture: Unsupervised methods: derivatives of LDA; structural topic models

No class 10 November

Week 13: 13 - 17 November

Lab: Topic modeling in `R`, with emphasis on STM

Lecture: Ensemble learning: letting multiple models attack the same problem

Week 14: 20 - 24 November

Lecture: Neural network and deep learning approaches to text

Lab: Using `mxnet` and `word2vec` in `R`

No class 22 - 24 November: Mixed class/lab on 20 November.

Week 15: 27 - 01 December

Lecture: Beyond atomic representations: are they useful, and if so, how?

Lab: Using `word2vec` in R, continued

Week 16: 04 - 06 December

Lab: Project presentations

Lab: Project presentations

REQUIRED INFORMATION

Academic Honesty

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Plagiarism of written assignments – using someone else's words or ideas without proper citation - will not be tolerated. If you are unsure whether your work meets standards of academic honesty, please feel free to discuss your questions or concerns with me.

Disabilities

Students with disabilities (including mental health, chronic or temporary medical conditions) that have been certified by the Office of Student Life Disability Services will be appropriately accommodated and should inform the instructor as soon as possible of their needs. The Office of Student Life Disability Services is located in 098 Baker Hall, 113 W. 12th Avenue; telephone 614-292-3307, slds@osu.edu; http://slds.osu.edu