PPOL 628 – Text As Data: Computational Linguistics for Social Scientists

Matthew Denny   Monday 3rd February, 2020

Contact and Office Hours

- **Instructor name:** Matthew Denny (pronouns: he/him). Feel free to call me Matt, or Dr./Professor Denny if you prefer.
- **Email:** md1723@georgetown.edu or matthewjdenny@gmail.com.
- **Course Teaching Assistant:** Ted Ellsworth, email: tedellsw@terpmail.umd.edu
- **Office Hours:** TBD but most likely for one hour before class on Tuesdays.
- **Instructor Website:** `http://www.mjdenny.com`.
- **Course Website:** `https://github.com/matthewjdenny/PPOL_628_Text_As_Data`.

Course Overview

The social sciences have experienced an explosion of interest in the quantitative analysis of textual data over the past decade. This interest has not been misplaced, as text as data studies have revealed important social phenomena across a wide range of disciplines, and at a previously inconceivable scale. Yet, as with any emerging field, the quantitative analysis of text data presents many poorly documented pitfalls, as well as significant technical and theoretical challenges. This course seeks to arm its participants with the theoretical background, practical experience, and technical capacity to pursue cutting edge social science research using text data. This course is designed to cover the key technical aspects of conducting research with text data: from data collection and preprocessing, through to description and inferential analysis. We will cover various techniques from computational linguistics such as parts-of-speech tagging and sentiment analysis, term-category associations, supervised learning with text, topic modelling, and word embeddings, to name a few. Students are expected to have some experience with R programming, and some background in statistical analysis for the social sciences.

Prerequisites

I expect that students are comfortable using R, have some basic grounding in statistics, and have worked on at least one significant project[1] where they had to manipulate and analyze data at some point in the past few years. More specifically:

- I expect that students can load in and manipulate data in R. If you are a proficient Python user then I leave it up to you to decide if you feel comfortable working in R. I also expect that students have at least some basic familiarity with concepts like conditional statements (if/then) and looping (for and while loops).

[1] e.g. writing a paper for submission to an academic journal, working as an RA, a project at your previous employer, etc.

- I expect that students are familiar with basic concepts in statistics such as the normal, uniform, multinomial distribution, and what it means to sample from a distribution; linear and logistic regression and the interpretation of parameter estimates from these models; basic statistical/mathematical concepts such as mean, variance, expected value, logarithmic and exponential functions.
- I expect that students have some familiarity with social science research even they are not in a social science field.

Evaluation

It is important to me that students: put effort into this course, develop an understanding of the materials we cover in class, and demonstrate that they can apply the techniques we cover in their own work. As such, your grade in this class will be determined by the following four components: class participation, weekly homework assignments, a final project, and a final exam. If you put in the effort, do the readings, and pick a project you are excited about, you will get a good grade in this course.

Participation (25 %)

I expect students to do the readings each week, to pay attention in class, to participate in discussing the articles we read, and to try out any example code we go over for themselves. I understand that talking in class is not always ideally suited to all students, so if you find talking in class challenging or unpleasant, shoot me an email and we can work out other ways for you to participate in class. Part of your class participation grade will be to contribute at least one question to the weekly reading questions document. I will email out a link to this document each week. I expect your

Homework (25 %)

I believe it is critically important that students apply the concepts and tools we learn about in each class on their own. Your first assignment in this class will be to find or collect a corpus of documents that you will analyze for your homework assignments and final project (this assignment is described in more detail in the Schedule section). Each week you will responsible for applying and/or extending the example code we went over in the previous week to your own data, and providing a short writeup of what you found.

- Please use a consistent format for the name of your writeup along the lines of: **firstname_lastname_week_xx** Please send him your homework exported as a PDF file so that there will be no issues in reading/opening it across operating systems. So for the second week, I would send **matt_denny_week_2.pdf** as my homework submission.
- Please keep your writeup to 1-2 pages! When we have covered a method that makes sense to display the results of as a plot, please include a plot. In general, I would like one paragraph about what your plan was for the task for that week (e.g. your theory of how you should preprocess your data), and one paragraph about what you actually did and your results (e.g. how many documents you ended up collecting, some descriptive statistics of the covariates you collected, etc.). You may include the code you used for the homework as supplemental materials if you would like feedback but please let Ted know you plan to do this in your email and be as specific as you can about any questions you have on your code.
- Grading of homework. I am asking our TA to grade your homework based on the following rubric. Note that we will drop your lowest grade on the homework over the semester.
  - **Full Credit:** Homework was turned in on time, the student made an effort to apply the relevant methods, and the student provided a writeup that at least provided some basic discussion of what they found.

  - **Half Credit:** The student turned in their assignment late without getting in touch with Ted or I beforehand with a request for an extension, and/or the student did not apply the relevant methods to their data, and/or the student provided effectively no discussion of what they found.
  - **No Credit:** Homework was not turned in.

Final Project (25 %)

Building off of one of the analyses you perform as part of your homework for the class, you will be asked to write an 8-10 page report fleshing out this analysis to learn something substantively interesting from your data. This report can be part of a larger project you are working on, including collaborative work, but I expect you to do all of the analyses and writeup yourself. What I am looking for in this report is pretty simple: explore an aspect of your data that you find interesting or important and use at least one of the techniques we go over in this class to actually learn something from your data. You will be given more detailed guidelines for this project as the semester progresses.

Note that for both the homework assignments and the final project, you will be asked to collect your own corpus (collection of documents) for this class. You may use a dataset you find online, one your adviser already has, one you have already collected, or one you collect for this class. Your dataset must contain:

- 100+ documents (come see me if you really want to use a corpus containing fewer documents).
- Your dataset should be 100+ pages (30,000+ words) in total. This is important because some of the algorithms we use in this class require a reasonably large amount of text to produce actually interpretable results.
- Your dataset should have one categorical variable and one continuous variable per document for at least 100 documents. You can hand-code these for a subset of documents if you are working with a larger dataset and there is not an easy way to gather these covariates automatically/programatically.

Final Exam (25%)

This class will have a one-hour final exam. I know, this sounds terrible, but I promise it will not be a big deal if you do the readings and come to class. I will ask you to answer 2-3 broad understanding questions drawn from the topics we cover in class. I will give you a list of 10-20 candidate questions several weeks before the final and will pick the exam questions directly from this list. My goal is that you distil your thoughts on the topics we cover in class while studying for the final to make them stick with you after the course. There will not be any math or programming in these questions. Here is an example question that will be on the list:

- In this course we have learned about a number of different choices a researcher must make while preprocessing their text data (e.g. whether or not to stem, whether to remove stopwords, whether to lowercase, etc.). Provide an argument for why you made two of these choices when working with the data you used for your homework assignments/final project.

Academic Integrity

This course stresses individual work. While I encourage you to meet up with your classmates to discuss the readings and help each other with the homework assignments, I expect you to do all of the work you turn in for this course. This means don't plagiarize, and don't turn in somebody else's work as your own (note, you are more than free to adapt the code from class examples/labs). If you don't do your own work, you are wasting your time, and my time. I will report this type of behavior to the university. If you are worried that you might run afoul of the academic integrity standards I and Georgetown University lay out for you, just come talk to me!

Resources

Below are a few resources you will likely find helpful for this course, and for further exploration of text as data methods. I like the following textbooks, which are available for free online. We have assigned readings from all of them, but there is lots more good stuff in these that we will not cover in the course.

- James et al. (2017) An Introduction to Statistical Learning. http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
- Manning, C. D., & Schtze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. http://nlp.stanford.edu/fsnlp/
- Jacob Eisenstein. (2018). Natural Language Processing. https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes-june-1.pdf

If you need to brush up on your R skills, I recommend Hadley Wickham's R for Data Science book, which is available for free on line at: https://r4ds.had.co.nz/. I have also written a number of R tutorials which available on my website http://www.mjdenny.com, and in particular you might find this one (http://www.mjdenny.com/R_Tutorial.html) useful to get a sense of how I teach about R.

Schedule

PLEASE BRING A LAPTOP WITH YOU TO ALL CLASS MEETINGS, INCLUDING THE FIRST ONE. ALL READINGS FOR A CLASS SHOULD BE COMPLETED BEFORE THAT CLASS.

1. **01/14/20 – Introductions and logistics.**
    - We will go over the syllabus, introduce our selves, and I will be happy to answer any questions you have.
    - We will beging to discuss potential datasets (or corpora) that you might use in your homework and final projects.
    - The rest of class will be spent going over an R refresher an introduction to string processing in R.
    - Lab: string manipulation and regular expressions.
2. **01/21/20 – An overview of the field and collecting your own data.**
    - We will discuss some classic readings in the field and the types of research questions text data can be useful for answering, and the ways researchers have applied text analysis in the past.
    - I will expect you to have several ideas for a corpus you intend to use for this class, and to have done some preliminary research into how you might obtain these candidate corpora.
    - We will go over the basics of web scraping for those of you who are interested in collecting your own data.
    - Read Chapter 1 from: http://brenocon.com/phdthesis/thesis.pdf
    - Read: Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267297. https://doi.org/10.1093/pan/mps028

- Read: Wilkerson, J. D., & Casas, A. (2016). Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annual Review of Political Science, 118. http://www.annualreviews.org/doi/10.1146/annurev-polisci-052615-025542
- Lab: basic web scraping and forming a quanteda corpus object.
- Homework (due before next class): Complete initial data collection as described in the **Final Project** section above. Hand in a 1-2 page writeup where describe what dataset you intend to use, how you collected it, and give a brief overview of what type of documents they are, and the range, values., etc for your categorical and continuous variables. As with all homework assignments going forward, feel free to ask questions in your writeup if you would like feedback from the TA, and treat this as an opportunity to collect and document your thoughts.

3. **01/28/20 – Text preprocessing (data collection due).**

   - You will be expected to have at least a preliminary version of your dataset collected before class and available as a quanteda corpus object.
   - We will cover the basics of text preprocessing, and discuss the consequences of preprocessing choices and how to select and evaluate a preprocessing specification.
   - In our lab we will cover various text preprocessing options and the PreText R package for assessing the robustness of a given set of preprocessing choices.
   - Read Chapter 4, Section 2 from: Manning, C. D., & Schtze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. http://nlp.stanford.edu/fsnlp/
   - Read: Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), 168189. https://doi.org/doi:10.1017/pan.2017.44
   - Lab: basic text preprocessing and and PreText.
   - Homework (Due before next class): Hand in a 1-2 page writeup where you go over the substantive questions you are interested in answering with your dataset, and your initial thoughts on how you think you ought to preprocess your data. To get you started, why are your data interesting? What thing do you hope to learn from the analysis of your data, and why is this important/how does it tie in with existing research? How well suited do you feel your text are to actually telling you what you want to know from your data? Do you foresee any issues? It is totally fine if your data are not ideal for answering your research question, it is just a good idea to spell out your conception of these potential complications ahead of time.

4. **02/04/20 – TENTATIVELY NO CLASS (Travel Conflict)**

   - I will be travelling for work, so no class this week.
   - No readings this week.
   - Please turn in the homework assignment described in the previous week by when we would have had class this week.

5. **02/11/20 – Basic NLP: Parts of speech.**

   - We will discuss parts of speech, what they are useful for, and how they constitute phrases.
   - In our lab we will cover using a Part of Speech tagger in R, and how to work with the results.
   - Read Chapter 3 from: Manning, C. D., & Schtze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. http://nlp.stanford.edu/fsnlp/
   - Read: Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(01). https://doi.org/10.1017/S1351324900000048

- Read: Handler, A., Denny, M. J., Wallach, H., & OConnor, B. (2016). Bag of What? Simple Noun Phrase Extraction for Text Analysis. In Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing. `https://brenocon.com/handler2016phrases.pdf.em`
- Lab: parts of speech tagging and phrase extraction.
- Homework (Due before next class): Apply a part of speech tagger and/or phrasemachine to your documents. Hand in a 1-2 page document summarizing what you found. Some questions to get you started on your writeup: What is the distribution over PoS tags in your corpus? Are any parts of speech particularly interesting? How does your vocabulary size change when using phrasemachine with 1, 2, and 3-grams vs. all 1, 2, and 3-grams? Pull out a specific Part of speech tag and look at how tokens of that type (e.g. verbs) are used in your documents, does anything stick out to you (e.g. differences across your categorical variable)?

6. **02/18/20 – NO CLASS (Holiday)**
7. **02/25/20 – Dictionary-based methods, basic sentiment analysis.**
   - This week we will discuss the application of human-curated dictionaries in social science analyses and provide examples in the domain of sentiment analysis.
   - Read: Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication, 29(2), 205231. `https://doi.org/10.1080/10584609.2012.671234`
   - Read: Guo et al. (2016) Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. `http://chrisjvargo.com/wp-content/uploads/2016/04/Journalism-Mass-Communication-Quarterly-2016-Guo-107769` `pdf`
   - Lab: applying (sentiment) dictionaries.
   - Homework (Due before next class): Apply one of the generic sentiment dictionaries we go over in class to your data and analyze the results. Hand in a 1-2 page writeup where you go over your findings. To get you started, I would like to see you perform and interpret a regression analysis of your covariates and sentiment term counts, ratios of positive/negative, etc. Do you see any interesting patterns? Do your results make sense? If sentiment analysis does not really make sense in your context, you could also try creating your own custom dictionary (a good start on a potential part of your final project).
8. **03/03/20 – Corpus Description, Word counts, TF-IDF.**
   - Our primary focus this week will be on quantitatively describing and exploring a corpus.
   - Read: OConnor, B. (2014). MITEXTEXPLORER: Linked brushing and mutual information for exploratory text data analysis. In ACL 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces. `http://aclweb.org/anthology/W14-3101`.
   - Read: Lewis (1992). Feature selection and feature extraction for text categorization. `https://www.aclweb.org/anthology/H92-1041.pdf`
   - Read: White (2016). Bag of Works Retrieval: TF-IDF Weighting of Co-cited Works. `http://ceur-ws.org/Vol-1567/paper7.pdf`
   - Lab: corpus description, statistically unusual terms.
   - Homework (Due before next class): Apply the corpus description methods discussed in class to your own data. Hand in a 1-2 page writeup of your findings. To get you started, here are some questions: What terms show up as having the highest TF-IDF weights? Are you surprised by

these terms/is there an interesting story to tell about them (also may be the case that there is not)? If you preprocess your data without removing stopwords, how does the ranking of those terms change when you use simple TF weighting vs. IDF vs. TFIDF weighting?

9. **03/10/20 – NO CLASS (Holiday)**

10. **03/17/20 – Term-category associations.**

    - We will discuss statistical approaches to determining term-category associations, and how to apply these methods.
    - Read: Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16, 372403. https://doi.org/10.1093/pan/mpn018
    - Read: Denny (2016). Revisiting Fightin Words: Feature Selection Using an Informed Dirichlet Model. `https://github.com/matthewjdenny/PPOL_628_Text_As_Data/blob/master/Readings/Denny_Fightin_Words.pdf`
    - Read: Denny (2019). Adding the Magic Words: The Importance of Legal Details in Successful Legislation. `https://github.com/matthewjdenny/PPOL_628_Text_As_Data/blob/master/Readings/Denny_Magic_Words.pdf`
    - Lab: Contingency tables and Fightin' words.
    - Homework (Due before next class): Try out the fightin words on ACMI feature selection method on your data. Hand in a 1-2 page document describing what you found. Some questions to get you started? Would you theoretically expect there to be terms that strongly distinguish between classes? Did your empirical results match up to your expectations? How does ACMI scoring compare to stopword lists, your intuition, etc. Does it make sense in the case of your corpus/categorical covariates?

11. **03/24/20 – Text Reuse**

    - We will discuss concepts related to text reuse and plagiarism detection and their application in the social sciences.
    - I will present some of my research in this area and introduce some software tools.
    - Read: Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. American Journal of Political Science, 59(4), 943956. `https://doi.org/10.1111/ajps.12175`
    - Read: Casas, A., Denny, M. J., & Wilkerson, J. (2019). More Effective Than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process. American Journal of Political Science. `https://ssrn.com/abstract=3098325`
    - Read: Denny (2016) Assessing Editing Patterns Across Document Versions `https://github.com/matthewjdenny/PPOL_628_Text_As_Data/blob/master/Readings/Denny_Capturing_Edit_Structur pdf`
    - Lab: text reuse measures.
    - Homework (Due before next class): Apply the document similarity/text reuse methods we go over in the lab to your data. Hand in a 1-2 page document describing what you found. Some questions to get you started: Is your corpus one where you expect any direct text reuse between documents? If so, what did you find and does it match up with you expectations? If not, are there any features of document similarity that were particularly relevant to your theoretical question? How does document similarity relate to your covariate data?

12. **03/31/20 – Supervised Learning with Text.**

- We will voer the basics of supervised document classification.
- Read Chapter 4 through to the end of part 4.3 from: James et al. (2017) An Introduction to Statistical Learning. `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing` `pdf`
- Read: Jukra et al. RTextTools: A Supervised Learning Package for Text Classification `https://research.vu.nl/ws/portalfiles/portal/828993/310585.pdf`
- Read Chapter 4 from: Jacob Eisenstein. (2018). Natural Language Processing. `https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes-june-1.pdf`
- Lab: basics of supervised learning with text in R.
- Homework (Due before next class): Apply a supervised learning method to your data (e.g. use categorical covariate, hand coding to for a training set, then predict classes for some subset of your documents). Hand in a 1-2 page document describing what you found. Some questions to get you started: How does a simple dictionary method compare to a full supervised learning run for your classification task? What features are particularly important predictors? How well did the model perform in and out of sample?

13. **04/07/20 – Introduction to Topic Models.**
    - We will cover the intuition, the basics of the math, and the basics of implementing a topic model.
    - Time permitting, we will get into a discussion of evaluating topic model fit.
    - Read: `https://medium.com/@pratikbarhate/latent-dirichlet-allocation-for-beginners-a-hig`
    - Make a moderate attempt to read: Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3, 9931022. `http://dl.acm.org/citation.cfm?id=944937`
    - Read: Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09, (4), 18. `https://doi.org/10.1145/1553374.1553515`
    - Lab: fitting and evaluating a topic model.
    - Homework (Due before next class): Run a basic topic model on your documents and interpret/evaluate the output. Hand in a 1-2 page document describing what you found. Some questions to get you started: How many topics did you select and why? How does changing the number of topics affect your results? How do your results change as you remove/include stopwords? Are the top terms interpretable as meaningful topics? Where there some topics you expected but did not see, or some you saw but did not expect? Are the topics generally topical or keying on some other feature of the documents? Use KWIC to look at some of the top terms, do they have the meaning you expected? How to your results change qualitatively if you use a different initialization?

14. **04/14/20 – Assessing Topic Models and the Structural Topic Model.**
    - We will being by wrapping up our discussion of topic model fit and validation, then move on toa brief discussion and tutorial on using structural topic models.
    - Read: Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209228.
    - Read: Roberts, M. E. et al. (2014). Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4), 10641082. `https://doi.org/10.1111/ajps.12103`
    - Lab: fitting a structural topic model.

- Homework (Due before next class): Apply a structural topic model to your data. Follow (at least some of) the approach outlined by Quinn et al. to validate your findings. Hand in a 1-2 page document describing what you found. Some questions to get you started: Did you find any interesting relationships between your covariates and topics? Were these what you expected? What did you learn from your validation efforts? Did your understanding of your results change?

15. **04/21/20 – Word Embeddings.**
    - This class will focus more on using pre-trained word embeddings than training a new embedding model.
    - We will also discuss the limitations of the word embedding approach and some potential applications in the social sciences.
    - Read: https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf
    - Read: https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00008
    - Lab: applying word embeddings.
    - Homework (Due before next class): Apply a pre-trained set of embeddings to your data. Examine the relationships between documents (closeness) and terms. Hand in a 1-2 page document describing what you found. Some questions to get you started: How do document similarities as calculated via averaging word embeddings compare to the other methods we went over in class? Explore the embeddings of some particular words of interest, what do you see? Are there any words that are "close" to words in a dictionary you created that you would consider adding to that dictionary?

16. **04/28/20 – Make up class/special topics.**
    - We are likely to fall behind at some point during the semester, so this will serve as a make-up class. If we manage to stay on track through the entire semester, I can answer questions about the final exam, your final projects, and/or go over some special topics from my own research.

17. **05/02/20 – Final Exam (7:00-9:00 pm).**
    - If I have correctly interpreted the final exam schedule, we are slotted for a final exam at this time. I intend that you only work on the final exam for one hour, so we will probably start around 7:15 and end at 8:15.

18. **05/08/20 – Final Project Reports Due.**
    - No extensions on this as I will only have a few days to grade these and submit your grades.