POLI176: Text as Data*
Summer II, Tu/Th 11am-1:50pm
Yin Yuan

Class Information

Time: Tu/Th 11am-1:50pm
Zoom Link: https://ucsd.zoom.us/j/95694462805

Instructor Information

Yin Yuan
yiy055@ucsd.edu
Office Hours: Fridays, 11am-1pm

TAs Information

Kengchi Chang
kechang@ucsd.edu
Office Hours: TBD

Kennedy Middleton
kmiddlet@ucsd.edu
Office Hours: TBD

Course Overview

In the digital age, a large part of human knowledge, communications and activities are recorded in the form of text that is publicly accessible in large quantities. Wikipedia, political party's manifestos, politicians' press releases, newspapers, social media posts...these new forms of data open up opportunities to explore new questions about and measure human behaviors in a way not possible before.

This course focuses on implementation of common tools of automated text analysis. We focus on two broad tasks: classification (categorizing texts into pre-specified "bins") and scaling (arranging or scoring texts along a continuum). Within each type of task, the tools are introduced following different stages of inquiry. First, a general question is proposed which directs us to collect texts from particular sources. The collected raw texts are then preprocessed and represented in a form amenable to further statistical analyses. Next, we want to get to know our data better through exploratory analyses. In social science researches, this stage facilitates hypotheses formation. With the help of unsupervised machine learning techniques, we discover interesting ways to organize and categorize our data that inform us of questions and hypotheses that have never occurred to us. Third, inspired by our exploration, we create specific measures using texts often with the help of supervised machine learning techniques. Finally, we use these measures to test hypotheses and make causal claims that explain social phenomena or inform decisions.

This course is application oriented. While we do not delve deeply into the mathematical details behind techniques and algorithms, we do not treat them completely as black box and will help you understand their inner workings on an intuitive level.

Course Structure

Each lecture consists of two parts. In the first half of the lecture we explain how certain text analytical tools work and look at how social scientists have used them to answer interesting questions. The second half of the lecture focuses on implementation of these tool in R. We will pose coding challenges and pause for you to work them out in groups before going through solutions together.

*The development of this course is influenced by Molly Roberts, Justin Grimmer, Brandon Stewart and Arthur Spirling.

Prerequisites

The primary programming language of this course is R. Although a quick refresher of R might be given, you are highly encouraged to familiarize yourself with common data structures in R (vector, matrices, data frames, lists), conditional statement, for loop, function, the apply family, etc. This introduction to R (https://thomasleeper.com/Rcourse/Intro2R/Intro2R.pdf) covers most of the basics. Datacamp (www.datacamp.com) offers "Introduction to R" and "Intermediate R" that together take about 10 hours to complete. You can refer to these resources or other online resources (e.g. Codeacademy, Coursera) if R is completely new to you.

Problem Sets

There will be three problem sets throughout the course (please see course outline below for specific due dates). Each problem set will be posted a week prior to the due date. The problem sets are designed to be a "learning-by-doing" process where you can practice what you learned in class with real world problems.

Note: All deadlines in this course are at 11:59pm on that date.

Final Project and Memos

You will submit a report for your final project (12-15 pages, double-spaced, 12pt, tables and figures included) on **September 5th (Saturday)**.

However, you will not complete the final project in one setting. Throughout the course, you will write three memos (1-2 pages, double-spaced, 12pt) for your final project. In each memo, you will answer to a prompt that guides you in thinking about how the methods discussed in class can help you explore the questions you are interested in.

Grading

The course grade is determined by the following components:

Three Problem Sets	$15\% \times 3 = 45\%$
Three Memos	$5\% \times 3 = 15\%$
Final Project	40%

Grade Scale

Final letter grades will be assigned according to the following scale:

A/A+	93 – 96/97+	C+	77 – 79
A–	90 – 92	C	73 – 76
B+	87 – 89	C–	70 – 72
B	83 – 86	D	60 – 69
B–	80 – 82	F	0 – 59

Please note: students taking Pass/Not Pass as grading option must achieve a C– (i.e. 70) for a Pass.

Logistics and Class Policy

The lectures will be recorded and uploaded after class for the convenience of students in different time zones.

In general, late assignments will be deducted 2 points passing the deadline (with a 30-minute grace period) and additional 2 points for every 24 hours passing the deadline. However, we understand the challenges and struggles amid the pandemic and political events, so please reach out to us if you need an extension (prior to the due date) or if you need any help. Your health, whether physical or mental, always takes priority.

The university's Principles of Community along with all rules and practices regarding Academic Integrity apply in this course. Although you are encouraged to work together in class on coding challenges, I expect you to work on the problem sets and final project INDEPENDENTLY. Please do not share codes or any other form of solutions to the problem sets with other students.

Textbook and course readings

This class does not require a main textbook. Most of the readings are recent publications on the subject matter and occasionally book chapters are picked from multiple textbooks.

For those interested, here is a list of the most recommended textbooks in machine learning and natural language processing. All of them have free electronic versions online. Some electronic copies could be accessed via UCSD library while connected to UCSD VPN.

Machine Learning

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. New York: Springer series in statistics.
- Bishop, C. M. (2006). Pattern recognition and machine learning. springer.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Natural Language Processing

- Speech and Language Processing (3rd ed.). Daniel Jurafsky & James H. Martin. Draft of October 2, 2019. https://web.stanford.edu/~jurafsky/slp3/2.pdf
- Eisenstein, J. (2018). Natural language processing. https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Course Outline

Introducing Text as Data

In this section, we briefly survey the fast evolving field of automated text analysis and shows the potentials and challenges of its application in social sciences. We provide an overview of the techniques you are about to learn and the logic behind how the course is organized.

August 4: Potentials and Limits of Text as Data

- 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), 267-297.

Collect, Select, Preprocess and Represent Texts

In this section, we are going to learn how to collect and select text and think about the challenges that might emerge in this process (e.g. selection bias, generalizability). We will also learn how to transform raw texts into numerical form for statistical analysis and the implication of the choices we make.

August 6: Collect, Select, Preprocess and Represent Texts

- 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), 168-189.
- Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. Social Science Computer Review, 33(6), 712-729.
- Chapter 2, Speech and Language Processing (3rd ed.). Daniel Jurafsky & James H. Martin. Draft of October 2, 2019. https://web.stanford.edu/~jurafsky/slp3/2.pdf

Unsupervised Classification: Exploring What We Want to Measure

In this section, we will learn how to discover interesting ways to categorize and organize our text data. At this point, we still do not know what our data has to offer, and we do not have clear enough hypotheses to know what to measure or what categories we should group our texts into. We use unsupervised methods to help with hypotheses formation. In addition, we will learn how to characterize and summarize the patterns that emerge by finding distinctive words and interpreting topic model outputs.

August 11: Clustering and PCA

- Chapter 10, James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences, 108(7), 2643-2650.

August 13: Distinctive Words

Memo 1 Due

- Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association, 58(302), 275-309.
- 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(4), 372-403.
- Chuang, J., Manning, C. D., & Heer, J. (2012). "Without the clutter of unimportant words" Descriptive keyphrases for text visualization. ACM Transactions on Computer-Human Interaction (TOCHI), 19(3), 1-29.

Optional

- Taddy, M. (2013). Multinomial inverse regression for text analysis. Journal of the American Statistical Association, 108(503), 755-770.

August 18: Topic Models: Theory

- Bogdanov, P., & Mohr, J. W. (2013). Topic models. what they are and why they matter. Poetics, 31, 545-569.
- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4), 1064-1082.

Optional

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Political Analysis, 18(1), 1-35.
- 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), 209-228.

August 20: Topic Models: Interpretation and Application

Problem Set 1 Due
Memo 2 Due

- Nelson, L. K. (2020). Computational grounded theory: A methodological framework. Sociological Methods & Research, 49(1), 3-42.

– DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. Poetics, 41(6), 570-606.

– Lacombe, M. J. (2019). The political weaponization of gun owners: The National Rifle Association's cultivation, dissemination, and use of a group social identity. The Journal of Politics, 81(4), 1342-1356.

Optional

– Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

– Hannah, L. A., & Wallach, H. M. (2014). Summarizing topics: From word lists to phrases. In NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing (pp. 1-5).

– King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. American Political Science Review, 326-343.

Supervised Classification: Categorizing Texts into Known Categories

Given the categories and organizational structures we just discovered, how do we proceed to accurately and efficiently place our texts into the right "bins"? In this section, we will learn how to turn our texts into quantitative measures (either as classifications or proportions of categories) given our hypotheses.

August 25: Naive Bayes, SVM, Readme

- 3.4. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer
- Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. Journal of Information Technology & Politics, 5(1), 33-48.
- Jamal, A. A., Keohane, R. O., Romney, D., & Tingley, D. (2015). Anti-Americanism and anti-interventionism in Arabic Twitter discourses. Perspectives on Politics, 13(1), 55-73.

Optional

– D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. Political analysis, 22(2), 224-242.

– Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229-247.

August 27: Ensemble and Validation (CONTINUE INTO DICTIONARY METHODS)

Problem Set 2 Due
Memo 3 Due

- Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. Journal of Information Technology & Politics, 4(4), 31-46.
- Section 2.2 and 5.1. James, Witten, Hastie and Tibshirani. An Introduction to Statistical Learning Springer.

Optional

- Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).
- Grimmer, J., Westwood, S. J., & Messing, S. (2014). The impression of influence: legislator communication, representation, and democratic accountability. Princeton University Press.
- Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. Political Analysis, 25(4), 413-434.

Arranging Texts in a Continuum: Dictionary and Scaling

Often instead of dichotomous or discrete measures, we would like to arrange our texts in a continuum. Rather than knowing a politician is a conservative or a liberal, we are more interested in quantifying where he/she is located on a left-right scale. Sometimes even the dimensions along which texts are arranged themselves are subjects of interests. Like classification tasks, we can use supervised methods to score or scale a text with a pre-defined "ruler", or we can discover the "rulers" themselves first.

August 27 (CONTINUED): Dictionary

- Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of happiness studies, 11(4), 441-456.
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.

September 1: Scaling: Wordscore and Wordfish

- Lowe, W. (2008). Understanding wordscores. Political Analysis, 356-371.
- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American political science review, 97(2), 311-331.
- Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. American Journal of Political Science, 52(3), 705-722.

Discovering Dimensions

- Spirling, A. (2012). US treaty making with American Indians: Institutional change and relative power, 1784âĂŞ1911. American Journal of Political Science, 56(1), 84-97.

Optional

- Pan, J., & Xu, Y. (2018). China's ideological spectrum. The Journal of Politics, 80(1), 254-273.
- Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 114(25), 6521-6526.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.
- Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. Political Communication, 29(2), 205-231.

Causal Inference, Text Network and Word Embedding

In social sciences, we often do not stop at exploratory and descriptive analyses. Our ultimate goal of discovering and creating measures is to reveal relationships between phenomena we are interested in, especially causal relationships. In other words, we would like to be able to not only predict, but also explain. Moreover, we use language to communicate, and we communicate in a social network. How could we combine the power of text analysis and social network analysis (SNA) to study patterns of communication? Finally, rapid progress in Natural Language Processing (NLP) and deep learning has enabled us to model language in a much more sophisticated way. We will look at how word embeddings represent the meaning of words by capturing the context of their appearances.

September 3: Causal Inference, Text Network and Word Embedding

Problem Set 3 Due

- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.
- Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to make causal inferences using texts. arXiv preprint arXiv:1802.02163.
- Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. Proceedings of the National Academy of Sciences, 113(42), 11823-11828.

Optional

- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- Roberts, M. E., Stewart, B. M., & Nielsen, R. (2015). Matching methods for high-dimensional data with applications to text. Unpublished manuscript.