This version: January 8, 2021. Syllabus contents and order may change, be warned!

Text-as-Data
DS-GA 1015

Spring Semester 2021
Tuesdays 11:00 AM - 12:40 PM "Lecture" 19W4, 101

Prof. Arthur Spirling
`arthur.spirling@nyu.edu`
Office Hours: Wednesday, 1030–1130AM
Zoom: `https://nyu.zoom.us/j/6678696568`

Teaching Assistant: Ms Lucia Motolinia
`lucia.motolinia@nyu.edu`
Office Hours: Friday, 10–11AM
Lab/Section: Thursdays, 2-2:50PM

Prerequisites

At the very least, students should have a first class in statistics and/or inference under their belt before taking this course. In particular, basic knowledge of calculus, probability, densities, distributions, statistical tests, hypothesis testing, the linear model, maximum likelihood and generalized linear models is assumed. The core language and software environment of this course is `R`. If you not familiar with `R`, you will struggle with the assigned exercises. Please check with the instructor if you unclear as to whether you are qualified for this 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 analysis of text from a social science perspective, with a special focus on politics. The course is applied in nature, and while we will give some theoretical treatment of the topics at hand, the primary aim to help students understand the types of questions we can ask with text, and how to go about answering them. With that in mind, we first explain how texts may be modeled as quantitative entities and discuss how they might be 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 student conduct their own text as data research projects and this class provides the foundations on which more focussed, technical research can be built.

While many of the techniques we discuss have their origins in computer science or statistics, this is *not* a CS class: we will spend relatively little time on traditional Natural Language Processing issues (such as machine translation, optical character recognition, parts of speech tagging etc). Other offerings in the university cover those matters more than adequately. Similarly, this class will not much deal with *obtaining* text data: again, there are excellent classes elsewhere dealing with e.g. web-scraping.

Structure

As of January 8, 2021, the plan is that this course will provide the following on a weekly basis:

- approximately 70 minutes of pre-recorded lecture material (from the professor)
- approximately 30 minutes of "in-person" flipped material (from the professor) in room 101 of 19W4th Street. This will be recorded and livecast.
- a remote 50 minute section/lab with the TA, which will be recorded.

Enrolled students must attend live or watch remotely (live or later) all content we provide. The information and skills that you need to complete your homework assignments and term projects will be provided by the Professor or the TA. Note that **the pre-recorded lecture/flipped session is subject to demand: if turnout is low, we will simply revert to 'live' (but recorded) lecturing.** During the first week of this class, everything will be **remote** (nothing in person).

Sections: your TA will hold section Thu 2.00 PM - 2.50 PM, remotely. The Zoom link is `https://nyu.zoom.us/j/2990706756?pwd=OGdpUU5ack1kdEhZUjI3ektaZUx0QT09`. The TA's materials will be stored on the NYU Classes site.

Assessment

There are no written exams in the class, and your grade will be based on a combination of:

- **Homeworks (50%):** There will be (at least) three homeworks, all of which will involve modeling and coding of text data, and some theoretical work. Intellectual honesty is important at NYU: you may confer with colleagues, but all work on the homework must be your own. If you copy work or allow another student to copy your work, the homework will be graded zero and your case will be passed to appropriate authorities in the university. Homeworks must come in electronically as a *single RMarkdown PDF/HTML document of both answers and code.*
- **Final Paper (50%):** There will be a final written paper of not longer than 10 double spaced pages of text, which explores an original research project or idea. This may be substantive or technical in nature. You are encouraged to work in teams of up to two people on this paper. The deadline for the paper will be May 12, 2021 with no extensions or exceptions.

Software

We will be using `R`, a statistical package. You can download and install `R` for free, from here:

`https://cran.r-project.org/`

To write and edit `R` code, you can use any software with which you are familiar and/or enjoy using. We suggest `R Studio`, which is free:

`https://www.rstudio.com/products/RStudio/`

Textbooks and Reading

There are no required textbooks for the course. We will draw from some of the following (and other places!), and will make efforts to provide the readings online where appropriate:

- Klaus Krippendorff. Content Analysis: An Introduction to Its Methodology. Third Edition. Sage. 2013.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- Daniel Jurafsky and James H. Martin Speech and Language Processing, 2nd Edition. Prentice Hall. 2008
- Christopher Bishop. Pattern Recognition and Machine Learning, Springer. 2006.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer. 2009.
- Kevin Murphy Machine Learning: A Probabilistic Perspective. 1st Edition. MIT Press. 2012.

Because the class is focussed on answering substantive questions with the techniques on offer, many of the readings are applied in nature.

COURSE SCHEDULE

1 Feb 2: Introduction and Overview (REMOTE ONLY THIS WEEK)

This class is great: take it. Note that this week, we will be entirely remote: there will be **no flipped session**, and the professor will use his zoom link to lecture.

2 Feb 9: Representing Text

- vector space model of a document
- feature choices/representation

- pre-processing: stemming and stopping
- bag of words (and alternatives)
- sparseness

Reading

- MRS ch 6 "Scoring, term weighting and the vector space model"
- Denny, Matthew and Arthur Spirling, 2017. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145

3 Feb 16: Descriptive Inference I

- word distributions: Zipf's Law/Heap's Law
- co-occurrence, collocations and phrasemes
- key words in context
- dis(similarity) measures and testing for differences

Reading

- MRS, Ch 5

4 Feb 23: Descriptive Inference II

- lexical diversity
- sophistication/readability/complexity
- linguistic style and author attribution
- sampling distributions for estimates

Reading

- Benoit, K., Laver, M. and Mikhaylov, S. 2009. Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. American Journal of Political Science, 53: 495–513.
- A Spirling. 2016. Democratization and Linguistic Complexity, Journal of Politics.
- F Mosteller and D Wallace. 1963. Inference in an Authorship Problem, Journal of the American Statistical Association, Volume 58, Issue 302, 275–309.

- R Peng and N Hengartner. 2002. Quantitative Analysis of Literary Styles, The American Statistician, Volume 56.
- Benoit, K., Munger, K. and Spirling, A. 2017. Measuring and Explaining Political Sophistication Through Textual Complexity`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3062061`
- Hengel, Erin, 2017. Publishing while female Are women held to higher standards? Evidence from peer review `http://www.erinhengel.com/research/publishing_female.pdf`
- Huang, L., Perry, P., & Spirling, A. (2020). A General Model of Author "Style" with Application to the UK House of Commons, 1935–2018. Political Analysis.

5 Mar 2: Supervised Techniques I

- dictionary based approaches
- sentiment (and other) dictionaries, LIWC
- Goldman-Sachs case study
- event extraction
- lie detection

Reading

- Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (pages 1–27 only).
- Gary King and Will Lowe. 2003. An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. International Organization, 57, pp 617–642.
- Michael Laver and John Garry. 2000. Estimating Policy Positions from Political Texts. American Journal of Political Science Vol. 44, No. 3, pp. 619-634
- Yla R. Tausczik and James W. Pennebaker. 2009. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. March 2010 vol. 29 no. 1 24-54.

6 Mar 9: Supervised Techniques II

- classification of documents
- evaluation of techniques: precision, recall
- crowdsourcing
- Naive Bayes Classification, estimating proportions
- ideological scales with 'wordscores'

Reading

- MRS. "Text classification and Naive Bayes".
- Benoit, Kenneth, Conway, Drew, Lauderdale, Benjamin E., Laver, Michael and Mikhaylov, Slava. 2015. Crowd-sourced text analysis: reproducible and agile production of political data. American Political Science Review.
- Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data American Political Science Review 97(2)
- W Lowe. 2008. Understanding Wordscores, Political Analysis, 16 (4): 356-371.
- D Hopkins and G King. 2010. A Method of Automated Nonparametric Content Analysis for Social Science American Journal of Political Science, Vol. 54, No. 1, January 2010, 229–247.

7 Mar 16: Supervised Techniques IIIA

- basics/varieties of machine learning
- support vector machines

Reading

- Pedro Domingos. 2012. A Few Useful Things to Know About Machine Learning. Communications of the ACM CACM, Volume 55 Issue 10.Pages 78–87
- Daniel Diermeier, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2012. Language and Ideology in Congress British Journal of Political Science, 42, 31–55.
- V D'Orazio, S Landis, G Palmer, P Schrodt. 2014. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines Political Analysis 22 (2): 224-242.

8 Mar 23: From Supervised to Unsupervised

Supervised Techniques IIIB

- k-NN models
- random forests/tree techniques
- ensembles

Reading

- Siroky, David S. 2009. Navigating Random Forests and related advances in algorithmic modeling. Statist. Surv. 3, 147–163.
- Hillard, D. Purpura, S. Wilkerson, J. 2008. Computer Assisted Topic Classification for Mixed Methods Social Science Research. Journal of Information Technology and Politics 4:4.

Unsupervised Techniques I

- fundamentals of unsupervised learning
- (principal) components and data reduction
- singular value decomposition

Reading

- W Venables and B Ripley. 1999. Modern Applied Statistics with S. 4th Ed. Ch 11.

9 Mar 30: Unsupervised Techniques II

- clustering (documents)
- Latent Semantic Analysis/Indexing
- parametric scaling of political speech
- count models: 'wordfish'
- basics of semi-supervised techniques

Reading

- Justin Grimmer and Gary King. 2010. General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences. Vol 108, No 7. 2643–2650.
- Thomas K Landauer , Peter W. Foltz , Darrell Laham. 1998. An introduction to latent semantic analysis. Discourse Processes Vol. 25, Iss. 2–3.
- Simon Jackman 2000. Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo American Journal of Political Science, Vol. 44, No. 2, 375–404
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. American Journal of Political Science 52(3): 705-722.

10 Apr 6: Unsupervised Techniques III

- plate notation/graphical model
- basics of Bayesian methods
- Latent Dirichlet Allocation and Topic Modeling
- Variational Inference
- model selection/choosing k

Reading

- DM Blei, AY Ng and MI Jordan, 2003. Latent Dirichlet Allocation, Journal of machine Learning research 3, 993-1022.
- DM Blei and MI Jordan, 2006. Variational inference for Dirichlet process mixtures, Bayesian Analysis, Volume 1, Number 1, 121–143.
- H Wallach, I Murray, R Salakhutdinov and D Mimno. 2009. Evaluation Methods for Topic Models ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning, 1105–1112
- Grimmer, J. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases, Political Analysis, 18 (1): 1–35.

Apr 13: Work on Final Project, no lecture (will set up consulting time)

11 Apr 20: Unsupervised Techniques IV

- Correlated Topic Model
- Dynamic Topic Model
- Structural Topic Model
- Embeddings: Word2Vec (theory/overview)

Reading

- DM Blei and John D Lafferty, 2007. A Correlated Topic Model of Science. The Annals of Applied Statistics, Vol. 1, No. 1, 17–35.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. and Radev, D. R. (2010), How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science, 54: 209–228.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G. (2014), Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58: 1064–1082
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119. 2013.
- Meyer, David. How Exactly Does Word2Vec Work? `http://www.1-4-5.net/~dmm/ml/how_does_word2vec_work.pdf`
- Maja Rudolph, Francisco Ruiz, Susan Athey, and David Blei. 2017. Structured embedding models for grouped data. arXiv 1709.10367. `https://arxiv.org/abs/1709.10367`

- Motolinia, Lucia. “Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico.” American Political Science Review (2020): 1-17.

12 Apr 27: Special Topics I

- modeling debate and discourse
- networks of communication
- bursts and memes

Reading

- Eggers, A. C. and Spirling, A. 2014. Ministerial Responsiveness in Westminster Systems: Institutional Choices and House of Commons Debate, 1832-1915. American Journal of Political Science, 58: 873–887.
- J. Kleinberg. Bursty and Hierarchical Structure in Streams Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2002.
- Leskovec, Jure and Backstrom, Lars and Kleinberg, Jon. 2009. Meme-tracking and the Dynamics of the News Cycle. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Eggers, A.C and Spirling, A The Shadow Cabinet in Westminster. Systems. Modeling Opposition Agenda Setting in the House of Commons, 1832-1915. British Journal of Political Science, forthcoming.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. Proceedings of WWW.

13 May 4: Special Topics II

- word embeddings for applied social science research
- Word2Vec vs Glove
- Turing tests: intrinsic v extrinsic assessment
- embeddings for time series problems

Reading

- Spirling and Rodriguez. “Word Embeddings: What works, what doesn’t, and how to tell the difference for applied research” Working paper, available `https://github.com/ArthurSpirling/EmbeddingsPaper/blob/master/Paper/Embeddings_SpirlingRodriguez.pdf`here

- Rheault, Ludovic, and Christopher Cochrane. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." Political Analysis (2019): 1-22.
- Rodman, Emma. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." Political Analysis (2019): 1-25.
- Khodak, Mikhail, et al. "A la carte embedding: Cheap but effective induction of semantic feature vectors." arXiv preprint arXiv:1805.05388 (2018).

14 Wed, May 12: Final Projects Due