This version: February 14, 2024. Syllabus contents and order may change, be warned!

Text-as-Data
DS-GA 1015

Spring Semester 2024
Wed 10:00 AM - 11:40 PM Lecture 194 Mercer St Room 204

Marco Morucci
marco.morucci@nyu.edu
Office Hours: Wed 11:40-12:40PM (194 Mercer St Room 204)

Teaching Assistant: Trellace Lawrimore
trellace.lawrimore@nyu.edu
Office Hours: Wednesday 4-5pm, CDS Rm 244
Sign up in advance: tinyurl.com/tmlohs24
Lab/Section: Thu 1.30 PM - 2.20 PM
31 Washington Pl (Silver Ctr) Room 507

Teaching Assistant: Yirong (Brett) Bian
yb970@nyu.edu
Office Hours: Thursday 4-5pm, CDS Rm 763
Lab/Section: Thu 2.45 PM - 3.35 PM
31 Washington Pl (Silver Ctr) Room 514

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. Additionally knowledge of programming in `python` and basic computer science concepts such as algorithms and data structure will be required.

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

The plan is that this course will provide the following on a weekly basis:

- approximately 140 minutes of lecture material (from the instructor)
- a 50 minute section/lab with the TA.

The information and skills that you need to complete your homework assignments and term projects will be provided by the instructor or the TAs.

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 ipyhton notebook with both text 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 8, 2023 with no extensions or exceptions.

Lecture Recordings and Zoom Participation Enrolled students must attend live all content we provide; at the moment, we have no plans to record or broadcast anything.

Tools

Brightspace Our primary course site is hosted on NYU Brightspace. On this site you will find links to the most recent version of the syllabus, updates on course schedule as well as links to slides and book chapters. Your assignments should also be submitted via the assignments tab on our Brightspace site. We will also make important announcements via Brightspace.

Piazza The Piazza site for our course will enable you to ask and answer questions about the course. The teaching staff will monitor Piazza and try to provide answers at least once a day, however it is not guaranteed that we will get to your question within the current day. This applies especially to questions asked outside of regular work hours or during weekends. We **strongly** encourage you to try and answer each other's questions: this will both speed us up and provide you with a way to test your understanding, as the teaching staff will be able to review and correct student answers. We will also use Piazza to answer questions about assignments – make sure you monitor the forum.

- **Note: In order to use Piazza you must sign up.**
  Find our class signup link at: https://piazza.com/nyu/spring2024/dsga1015

Software All the programming work for this course will be done in Python.

Course Policies and Academic Integrity

Late Homeworks You must turn in all homework **on time**. We will inform you how to turn in your homework, as it will not be accepted by email. Every day late results in a grade level drop: an A becomes an $A-$, an $A-$ becomes a $B+$, and so on. If you are sick or have some other **unavoidable** problem that interferes with your ability to submit your homework on time, please let us know **before** the deadline.

Regrade requests should be handled with care: if you believe there has been a grading error on your assignment or exam, you may submit a regrade request by emailing your TA. We will then review your case and, if necessary, update your grade. Note that we also reserve the right to **lower** your grade if, upon review of your assignment, we determine you were graded too generously on any question. You must submit all grade appeals within **one week** of the grade being returned to you.

Final Grades Your final grade will be computed as follows: each homework assignment grade will be converted from points to percentage, then a weighted sum of the percentages will be computed with each assignment weighted as described prior. Your final paper will be assigned a letter grade. A curve will then be applied to the letter grades to decide your final grade. There is no **fixed** *a priori* curve for final grades, but our distribution is fair and in keeping with other introductory courses at NYU. Even though we will not know what the final curve will look like until we have

final grades in hand, we will use the curve to **strictly increase** your final grade, that is, your final grade will be the largest of your curved or uncurved letter grades.

There are no exceptions to the above policies. They are here to help and protect you as students as much as they are for us as instructors.

Academic integrity policy All students are expected to do their own work. Students may discuss assignments with each other, as well as with the course staff. Any discussion with others must be noted on a student's submitted assignment. Excessive collaboration (i.e., beyond discussing the assignment) will be considered a violation of academic integrity. Questions regarding acceptable collaboration should be directed to the class instructor prior to the collaboration. It is a violation of the honor code to copy or derive solutions from other students (or anyone at all), textbooks, previous instances of this course, or other courses covering the same topics. Copying solutions from other students, or from students who previously took a similar course, is also clearly a violation of the honor code. Finally, a good point to keep in mind is that you must be able to explain and/or re-derive anything that you submit. This is particularly important if you should adapt solutions from online sources. Please also refer to the general NYU academic integrity statement. For the purposes of the final paper, these rules are intended to apply to each paper team, rather than to each student individually.

AI policy We live in the age of viable generative AI. Banning these tools is neither realistic, nor desirable (not to mention ironic, in a class about language processing). In fact, learning to use these tools is an emerging skill. Note that AI tools do not always produce correct or accurate results. In addition, it is unwise to rely on them too much. There are situations where you won't have access to these tools, for instance during technical interviews. In addition, there are also skills someone with an advanced degree in Data Science is just expected to have on tap - without AI assistance or looking anything up. To integrate both considerations, you can use generative AI tools to do the assignments in this class, but If you use an AI to guide you in completing an assignment, you have to disclose which parts were generated by the AI.

Academic accommodations Academic accommodations are available for student accessibility. Please contact the Moses Center for Student Accessibility (212-998-4980; mosescsa@nyu.edu) for further information. We recommend that students requesting academic accommodations reach out to the Moses Center **as early as possible in the semester** for assistance.

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:

- Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. 2022
- 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. There will be a set of suggested readings for each lecture. While reading these papers is not required for the course, we still recommend students to read at least one or two per lecture to better familiarize themselves with the topics of the course.

Course Schedule

Lecture 1 (01/24): Course Introduction

Topics

- Introduction to the course
- Class Syllabus and policies

Lecture 2 (01/31): Representing text

Topics

- vector space model of a document
- feature choices/representation
- preprocessing: stemming and stopping
- bag of words (and alternatives)
- sparseness

Readings

- 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 Ithttps://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145

Lecture 3 (02/7): Describing text

Topics

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

Readings

- MRS, Ch 5
- Mozer, Reagan, et al. "Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality." Political Analysis 28.4 (2020): 445-468.

Lecture 4 (02/14): Describing text 2

Homework 1 Out, Due 02/28

Topics

- 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.

Lecture 5 (02/21): Supervised tasks 1

Topics

- 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.

Lecture 6 (02/28): Supervised tasks 2

Homework 2 Out, Due 03/13

- basics/varieties of machine learning
- classification of documents
- evaluation of techniques: precision, recall
- support vector machines

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.
- 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.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. science, 343(6176), 1203-1205.

Lecture 7 (03/6): Unsupervised tasks 1

Topics

- fundamentals of unsupervised learning
- (principal) components and data reduction
- singular value decomposition
- clustering (documents)
- Latent Semantic Analysis/Indexing
- parametric scaling of political speech
- count models: 'wordfish'
- basics of semi-supervised techniques

Readings

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

Lecture 8 (03/13): Unsupervised tasks 2

Homework 3 Out, Due 04/03

Topics

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

Readings

- 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.

Lecture 9 (03/27): Introduction to Deep Learning

Topics

- Deep Neural Networks
- Training and inference in DNNs
- Recurrent Neural Networks for text data

Readings

- TBD

Lecture 10 (04/03): Deep learning and text representation

Topics

- Deep Text representation: encoders and decoders
- Obtaining and using text encodings
- Fine-tuning encoders

Readings

- TBD

Lecture 11 (04/10): Supervised and unsupervised text tasks with deep learning (Tentative) Homework 4 Out, Due 04/24

Topics

- Text classification with Deep models
- Sentence prediction
- Topic modeling and clustering

Readings

- TBD

12 (04/17): From RNNs to Attention

Topics

- The attention mechanism
- Large Language Models
- Emergent features of LLMs: what we know and what we don't

Readings

- TBD

Lecture 13 (04/24): Working with Black-box LLMs

Topics

- LLMs for supervised tasks
- Instruction fine-tuning and prompt generation
- Small LMs and model fine-tuning via subspace projections

Readings

- TBD

Lecture 14 (05/14): More on practical LLMs, Course Review

Topics

- TBD

Readings

- TBD