Emory University
Department of Quantitative Theory and Methods

QTM 340 (Fall 2024)
Text as Data
T/Th 2:30-3:45pm, MCS E301A

Professor: Lauren Klein (lauren.klein@emory.edu)

Land Acknowledgment

Emory University is located on Muscogee (Creek) land. Emory was founded in 1836, during a period of sustained oppression, land dispossession, and forced removal of Muscogee (Creek) and Ani'yunwi'ya (Cherokee) peoples from Georgia and the Southeast. Emory owes an immense debt to the Muscogee, Ani'yunwi'ya and other original peoples, and their descendants, who have cared for and inhabited these lands.

Read the full Land Acknowledgment and History Statement developed by Emory faculty.

Dr. Klein's Office Hours:
Tuesdays 1-2pm and by appointment (schedule here)

Course Description
What does it mean to turn text into data? What are the data-scientific techniques that are commonly employed in order to analyze text? How are they applied in the humanities and social sciences? How are they applied in the world? This course explores these questions by focusing on how popular methods of text analysis, including those involving large language models, can be used to pursue humanistic and social-scientific research questions. Additional methods covered include text classification, clustering, and topic modeling, as well as methods for creating, cleaning, and parsing textual datasets. Along the way, we will also discuss the issues of ethics involved in our increasing reliance on large language models as well as the people whose labor—intellectual, physical, and emotional—that they depend upon.

Introductory courses in computer science and probability and statistics are recommended as perquisites for this course. You will complete all class exercises and homework assignments in Python. I expect you to participate in class discussion and present your final project at the end of the semester. I will also require some short writing assignments.

Required Course Materials
All required readings will be posted on Canvas and/or are available online.

List of Graded Assignments
Your grade for the course will be calculated as follows:

- Attendance and class participation: 10%
- Reading assignments (2): 15%
- Quizzes (3): 25%
- Final project preparation assignments (5): 20%
- Final project: 30%

Description of Graded Assignments

Class Participation

Class participation is often assumed to be a hazy concept, but it actually involves a careful assessment in five distinct areas. Here are short descriptions of each of these areas, adapted from grading criteria developed by Dr. Mark Sample of Davidson College:

- **Preparation:** Reading/reviewing any assigned material before class.
- **Presence:** Being verbally and nonverbally engaged during class.
- **Focus:** Avoiding distractions during class (both in person and online).
- **Asking questions** in class and in office hours, as well as via email when appropriate.
- **Specificity:** Referring to specific ideas from readings and prior class discussions when contributing to class discussion and/or in conversations during office hours.

Reading Assignments

Much of the required reading for this course involves technical papers in the fields of natural language processing (NLP), computational social science, and the digital and quantitative humanities. Because these readings have been carefully selected as models of what work with text as data can accomplish—and more concretely, as models of what work *you* might accomplish in your final project—it is absolutely essential that you read them. In order to assess your completion of the reading assignments, you will be required to submit structured written summaries of two such papers over the course of the semester. You can decide which two papers you decide to summarize, as long as you submit your summary before the start of the class in which the paper is discussed. You will receive a letter grade (A-F) on each summary. I will provide a more detailed assignment sheet after the introductory unit of the course.

Quizzes and Final Project Preparation (FPP) Assignments

Over the course of the semester, you will be completing eight short assessments. The first three, labeled "quizzes" on the syllabus, are designed to allow you to put your newly-learned skills into practice, and must be submitted individually. The final five are designed to lead up to the final project, and may be submitted as a project group. These assessments differ from the quizzes in that they are more open-ended, and assess your ability to conceptualize and implement a complete text analysis workflow (or other final project-related task). All quizzes and FPP assignments must be submitted via Canvas by the start of the day's class. You will receive a letter grade (A-F) for each submission. Some FPP assignments will receive written feedback as well.

Final Project

In addition to the assignments above, you will be completing a final project: a fully-developed application of text analysis techniques to a research question of your own devising. You will be required to present your project to the class and submit a research paper that documents your work. You may work alone or in groups of two or three. You will receive a letter grade (A-F) on the basis of your contribution, as well as written feedback. The final project is due on Tuesday, December 12th, the date of the final exam. (There is no final exam for this course).

Specific information about each assignment will be distributed no later than one week before the due date. You will receive detailed information about the final project well in advance.

If you would like additional feedback on any assignment, or have additional questions, please schedule a meeting with me during my office hours.

Policy on Late/Skipped Assignments
All assignments are mandatory. Should your group submit an assignment after the due date, your grade for that assignment will decrease by a 1/3rd letter grade for each day that it is late (e.g. B becomes B-). Should you fail to submit an assignment entirely, you will receive an F on that assignment. Should you need an extension, please contact us 24 hours *in advance* to negotiate the deadline. Should you experience difficulties working within your group, please also contact us as soon as possible.

Grading Rubric
This chart of grading characteristics, also adapted from criteria developed by Professor Mark Sample, describes the general rubric I employ when evaluating project-based work:

GRADE	CHARACTERISTICS
A	**Exceptional**. The research question is substantive and well-scoped. The motivation for undertaking the project is clearly stated, as are its stakes. The student/group has clearly identified how the project extends and/or otherwise contributes to the existing scholarship on the subject. The student/group has identified (either by selecting or creating) a corpus of significant research potential, and matched their methods of analysis both to the research question and to the corpus. They have employed the fullest possible range of methods that are appropriate to the research question, given the constraints of the particular project. They have analyzed the results of the research to the fullest extent possible, clearly identifying the implications of the research for the existing scholarship and in more general terms. They have considered the limitations of the research as well as possible next steps. The work reflects an *original and in-depth* engagement with the research topic.
B	**Satisfactory.** The research question is well-scoped and the motivation for undertaking the project is clearly stated, although its contributions are less substantive and its stakes are less compelling. The student/group has clearly identified how the project engages with existing scholarship, although they have not made clear how it extends and/or otherwise contributes to existing scholarship. The student/group has identified (either by selecting or creating) a corpus of solid research potential, and matched their methods of analysis both to the research question and to the corpus. They have employed methods that are appropriate to the research question, although they have not pursued all possible methods of analysis, given the constraints of the particular project. They have

	analyzed the results of the analysis sufficiently, identifying the major implications of the research, both for the existing scholarship and in more general terms, but they have not pursued those implications to the fullest extent possible. They have not fully considered the limitations of the research and/or possible next steps. The work reflects a *moderate* engagement with the research topic: satisfactory and certainly solid, but not as original or in-depth as it might be.
C	**Underdeveloped.** The research question is poorly scoped and the motivation for undertaking the project is unclear. The contributions of the research are not articulated or, if they are, remain uncompelling. The student/group has not clearly identified how the project engages with existing scholarship. The selected corpus lacks significant research potential and/or the methods of analysis are poorly matched to the research question and/or to the corpus. Few methods of analysis are employed. The results of the analysis are not sufficiently explored; few implications of the research, either for the existing scholarship or in more general terms, are considered. The individual/group has not considered the limitations of the research and/or possible next steps. The work reflects a *passing* engagement with the research topic: an attempt has been made, but not to a satisfactory degree.
D	**Limited.** The research question is poorly scoped and the motivation for undertaking the project is unclear. The contributions of the research are not articulated. The student/group has not identified how the project engages with existing scholarship. The selected corpus lacks significant research potential, and the methods of analysis are poorly matched to the research question and/or to the corpus. Few methods of analysis are employed; they may be incompletely applied. The results of the analysis are scarcely explored, and no extended implications and/or limitations of the research are considered. The individual/group has not considered possible next steps. The work displays *no evidence of student engagement* with the topic: a cursory attempt has been made, but it remains insufficient and/or incomplete.
F	**No Credit.** The work is missing or consists of one or two unfinished sections.

Process for Calculating Final Grades

At the end of the semester, I will convert each of your letter grades to a 12 point GPA scale (e.g. A = 12, A- = 11, B+ = 10) and weight each of these numbers according to the percentage listed above. On Canvas, the letter grade—NOT the numerical/percentage grade—will reflect your grade in the course.

Policy on Attendance and Punctuality
Prior to the pandemic, I allowed two excused absences, no questions asked, with your grade beginning to be lowered with the third absence. Since then, conditions have changed. *I do not want to pressure you to come to class if you might be sick*. With that said, you are responsible for finding out what was discussed in class on any days that you might miss. I do not provide copies of my lecture notes, although I may post slidedecks to Canvas. In addition, beginning with the third absence, *you must email me to let me know that you will be missing class for health reason*s. Finally, please be respectful to your peers and arrive on time. If you arrive more than 15 minutes late, you will be considered absent for that class.

Contacting your Professor
I can be reached via my Emory email address. I respond to email M-F 9am-5pm, and outside of those hours if my schedule allows. Please allow 24 hours for a response, and 48 hours if your message is sent over the weekend.

Office of Accessibility Services
Office of Accessibility Services works with students who have disabilities to provide reasonable accommodations. In order to receive consideration for reasonable accommodations, you must contact OAS. It is the responsibility of the student to register with OAS. Please note that accommodations are not retroactive and that disability accommodations are not provided until an accommodation letter has been processed. Students registered with OAS who have a letter outlining their academic accommodations, are strongly encouraged to coordinate a meeting time with your professor that will be best for both to discuss a protocol to implement the accommodations as needed throughout the semester. This meeting should occur as early in the semester as possible. Students must renew their accommodation letter every semester they attend classes. Contact the Office of Accessibility Services for more information at (404) 727-9877 or accessibility@emory.edu. Additional information is available at the OAS website at http://equityandinclusion.emory.edu/access/students/index.html.

Writing Center and ESL Program
Tutors in the Emory Writing Center and the ESL Program are available to support Emory College students as they work on any type of writing assignment, at any stage of the composing process. Tutors can assist with a range of projects, from traditional papers and presentations to websites and other multimedia projects. Writing Center and ESL tutors take a similar approach as they work with students on concerns including idea development, structure, use of sources, grammar, and word choice. They do not proofread for students. Instead, they discuss strategies and resources students can use as they write, revise, and edit their own work. Students who are non-native speakers of English are welcome to visit either the Writing Center tutors or the ESL tutors. All other students in the college should see Writing Center tutors. Learn more and make an appointment by visiting the websites of the ESL Program and the Writing Center. Please review tutoring policies before your visit.

Honor Code
The Honor Code applies to all work submitted for courses in Emory College. Students who violate the Honor Code may be subject to a written mark on their record, failure of the course,

suspension, permanent exclusion, or a combination of these and other sanctions. The Honor Code may be reviewed online at: http://catalog.college.emory.edu/academic/policies-regulations/honor-code.html.

If you are unsure as to what constitutes plagiarism, please contact me before submitting your assignment.

A Closing Note on Generative AI Tools

At this point, I assume that you are all at least somewhat familiar with the capabilities and limitations of Generative AI tools (e.g. Claude, ChatGPT, Gemini, Copilot). As an occasional user of these tools myself, as well as your professor, I believe that we should best navigate this new technological landscape together. As such, I ask that you hold the following principles and guidelines in mind throughout this course:

- *Writing is thinking*. *Writing is learning*. Very often, we do not entirely know what we think until we are required to put it into words. This is why writing is hard! While you may find it helpful to use genAI to brainstorm ideas, fine-tune research questions, suggest alternate phrasings, and the like, **you may not use genAI to produce final content for written assignments**. This is not only because it short-changes your learning in the course, but also because you may very well be plagiarizing someone else's words or phrases without you knowing it! (See "Speak, Memory" on Class 13.)
- *GenAI is a terrible source of facts*. This has to do with the models' underlying architecture, which we will in fact learn about in this course! For the purposes of your assignments, and your general edification, what you must know is that **genAI cannot be trusted for facts or truthful information**. All facts, claims, quotations, and other information that you include in your written assignments (including your final projects) must be backed up by citations from scholarly sources.
- *GenAI may at times be helpful for certain technical tasks*. Generally speaking, using genAI for programming and data processing tasks will get you some but not all of the way to your goal. You'll need your own knowledge and experience to adapt/debug any machine-generated code so that you can ensure that it's doing the thing you want it to. Which leads to a different version of the first principle: *coding is thinking and learning too*. **Take advantage of the supportive learning environment of this class to develop your own programming and debugging skills**. You will need them in order to make the most appropriate and effective use of code-related genAI tools in the future.

I will require a formal genAI disclosure statement with most assignments in this course, and I will discuss this with you as the first deadline approaches. I also want to ensure a level playing field for all students in this class. If you figure out how to do something with genAI that may be interesting or helpful to other students, I will ask that you share it with your peers. I will set aside time in class for such conversations/demonstrations as the need arises.

Course Schedule at a Glance

Lecture	Topic	Date	Milestone
1	What can you do with text as data?	Aug 29	
2	What should you do with text as data?	Sep 03	**Due:** Quiz #0
3	Introduction to HuggingFace `transformers`	Sep 05	
4	Introduction to the chatGPT API	Sep 10	
5	*// No class - Professor at Oxford*	Sep 12	**Due:** Quiz #1
6	Turning Words into Numbers	Sep 17	
7	Word Embeddings	Sep 19	
8	*// No class - Professor at Denison*	Sep 24	**Due:** Quiz #2
9	Contextual Embeddings	Sep 26	
10	Working with LLMs	Oct 01	
11	Working with (small) LMs	Oct 03	
12	Transformers I – joint w/ other section	Oct 08	**Due:** Quiz #3
13	Transformers II – joint w/ other section	Oct 10	
14	*Fall break (no classes)*	Oct 15	
15	Creating Your Own Data (libraries and APIs) **(async)**	Oct 17	**Due:** Dataset ideas
16	The People Behind LLM Data – joint w/ other section	Oct 22	
17	The People in LLM Data – joint w/ other section	Oct 24	
18	Cleaning Your Own Data (regex)	Oct 29	**Due:** Final project proposal OR dataset
19	Parsing/Tagging Your Own Data (spaCy)	Oct 31	
20	EDA 1: Clustering	Nov 05	**Due:** Final project proposal OR dataset
21	EDA 2: Topic Modeling	Nov 07	
22	Validation	Nov 12	
23	*// No class - Professor at ASA (conference)*	Nov 14	**Due:** Datasheet
24	Chaining methods	Nov 19	
25	Guest lecture TBD **(async)**	Nov 21	**Due:** Final project first pass

26	Guest lecture TBD **(async)**	Nov 26	
27	*// No class - Thanksgiving*	Nov 28	
28	Final presentations	Dec 03	
29	Final presentations **(async)**	Dec 05	
30	Course Wrap-Up	Dec 10	
X	[Date of Final Exam]	Dec 12	**Due: Final project**

Class-by-Class Schedule

Class schedule subject to change.
Please consult Canvas for the most current class schedule.

Overview and Introduction

1. Thursday, August 29th – What can you do with text as data?

// In class: syllabus overview, transcription exercise, Voyant

2. Tuesday, September 3rd – What *should* you do with text as data?

DUE: Quiz #0, video intro
Before class:

- Spend at least 15 minutes playing AI Dungeon
- Read: Emily M. Bender, Timnit Gebru et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"

// In class: Discussion of readings

Unit 1: Generative AI and How We Got Here

3. Thursday, September 5th – Introduction to HuggingFace transformers

Read before class:

- Visual Storytelling Team and Madumhita Murgia, "Generative AI exists because of the transformer," *The Financial Times* (2023)
- Kenrick Cai, "The $2 Billion Emoji: Hugging Face Wants To Be a Launchpad for a Machine Learning Revolution," *Forbes* (2022)
- Joseph Ferrer, "What is Hugging Face? The AI Community's Open Source Oasis," *Datacamp* (2023)

// In class: Introduction to the HuggingFace `transformers` library

4. Tuesday, September 10th – Introduction to the chatGPT API

Read before class:

- Elizabeth Pain, "How to (seriously) read a scientific paper," *Science* (2016)
- Li Lucy and David Bamman, "Gender and Representation Bias in GPT-3 Generated Stories," *Proceedings of the Third Workshop on Narrative Understanding* (2021). Association for Computational Linguistics.
- Optional: Travis Zack, Eric Lehman, et al., "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study," *The Lancet* (2024)

// In class: Introduction to the chatGPT API (and discussion of how to read technical papers)

5. Thursday, September 12th – NO CLASS MEETING, professor at Oxford

Due: Quiz #1: Querying LLMs via language and code

6. Tuesday, September 17th – Turning words into numbers

Read before class:
- Matt Daniels, "The Language of Hip Hop," *The Pudding* (2017)
- Yiwei Luo, Kristina Gligoric, and Dan Jurafsky, "Othering and Low Status Framing of Immigrant Cuisines in US Restaurant Reviews and Large Language Models," *ICWSM* (2024)
- Optional: Anjalie Field, Chan Young Park, et al., "Controlled Analyses of Social Biases in Wikipedia Bios," *WWW* (2022)

// In class: sklearn's `countVectorizer` and TF-IDF

7. Thursday, September 19th – Word Embeddings
Read before class:
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *PNAS (*2018)
- Lucas Avelar, "An Imagined Geography of Empire," *Digital Humanities* (2024)
- Optional: Sandeep Soni, Lauren Klein, and Jacob Eisenstein, "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers," *Journal of Cultural Analytics* (2021)

// In class: `word2vec`

8. Tuesday, September 24th – NO CLASS MEETING, Professor at Denison
Due: Quiz #2: Putting Word Embeddings to Use

Unit 2: Working with Language Models

9. Thursday, September 26th – Contextual Embeddings
Read before class:
- Li Lucy, Divya Tadimeti, and David Bamman, "Discovering Differences in the Representation of People Using Contextual Semantic Axes," *EMNLP* 2022
- Naitian Zhou, David Jurgens, and David Bamman, "Social Meme-ing: Measuring Linguistic Variation in Memes," *NACCL* 2024

// In class: contextual embeddings; maybe document or semantic embeddings if we have time

10. Tuesday, October 1st – Working with LLMs
- Julia Mendelsohn, Ronan Le Bras, et al., "From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models," *ACL* (2023)
- Melanie Walsh, Anna Preus, and Maria Antoniak, "Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets," *arXiv preprint* (2024)
- Optional: Caleb Ziems, William Held et al., "Can Large Language Models Transform Computational Social Science?" *ACL* (2024)

// In class: zero-shot prompting; maybe fine-tuning or RAG if we have time

11. Thursday, October 3rd – Working with (small) LMs
Read before class:
- Richard Jean So, "Recognition: Literary Distinction and Blackness," from *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction* (Columbia UP, 2021) – Canvas

- Rob Voigt, Nicholas Camp, et al., "Language from police body camera footage shows racial disparities in officer respect," *PNAS* (2017)
- Optional: Harini Suresh, Rajiv Movva, et al., "Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection," *FAccT* (2022)

// In class: LR classification, outliers

12. Tuesday, October 8th – Transformers day 1 – joint w/ Soni section
Due: Quiz #3: Classification and Interpretation
Read before class:

- Daniel Jurafsky and James H. Martin. "Speech and Language Processing." (Ch.9 and 10; 3rd edition) Link, Link
- Review: "Generative AI exists because of the transformer"
- Optional: Jay Alammar, "The Illustrated Transformer"

// In class: under the hood of transformers, day 1

13. Thursday, October 10th – Transformers day 2 – joint w/ Soni section
Read before class:

- Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman, "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4," *EMNLP* (2023).
- Optional: Moin Nadeem, Anna Bethke, and Siva Reddy, StereoSet: Measuring stereotypical bias in pretrained language models," *ACL* (2021)

// In class: under the hood of transformers, day 2

14. Tuesday, October 15th – NO CLASS METING - FALL BREAK

Unit 3: All About Data

15. Thursday, October 17th – Creating Your Own Data – async!
Due: Dataset ideas
Read before class:

- Tess McNulty, "What's On Top of TikTok?" *Public Books* (2023)
- Adina Gitomer, Julia Atienza-Barthelemy, and Brooke Foucault Welles, "Stop Scrolling! Youth Activism and Political Remix on TikTok," *Emerging Research in Online Governance* (2023)

// In class: software libraries and APIs with pyktok ("pick-tock")

16. Tuesday, October 22nd – The People Behind LLM Data – joint w/ Soni section
Read before class:

- Adrienne Williams, Milagros Miceli, and Timnit Gebru, "The Exploited Labor Behind Artificial Intelligence," *NOEMA* (2022)
- Carlos Toxtli, Siddharth Suri, and Saiph Savage, "Quantifying the Invisible Labor in Crowd Work," *CSCW* (2021)
- Optional: Explore Data Workers' Inquiry, Worker Info Exchange

// In class: discussion of readings

17. Thursday, October 24th – The People in LLM Data – joint w/ Soni section
Read before class:

- Suchin Gururangan, Dallas Card, et al., "Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection," *EMNLP* (2022)
- Li Lucy, Suchin Gururagnan, et al., "AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters," *Association for Computational Linguistics* (ACL) 2024.
- Optional: Sachin Kumar, Vidhisha Balachandran et al, "Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey," *EACL* (2023)

// In class: discussion of readings

18. Tuesday, October 29th – Processing Your Own Data
Due: Project proposal OR dataset
Read before class:

- Fei-Fei Li, "A Hypothesis," from *The Worlds I See: Curiosity, Expiration, and Discovery at the Dawn of AI* (Flatiron Books, 2023) – Canvas
- Joy Buolamwini, "Crawling Through Data," from *Unmasking AI* (Random House, 2023) – Canvas
- Optional: Arunesh Mathur, Angelina Wang, et al., "Manipulative tactics are the norm in political emails: Evidence from 300K emails from the 2020 US election cycle," *Big Data & Society* (2023)

// In class: regex

19. Thursday, October 31st – Parsing/Tagging Your Own Data (spaCy)

- Maarten Sap et al., "Connotation Frames of Power and Agency in Modern Films," *ACL* (2017)
- Xiaoyun Gong, Yuxi Lin, Ye Ding, and Lauren Klein, "Gender and Power in Japanese Light Novels," *CHR 2022*
- Optional: Dallas Card, Serina Chang, et al., "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration," *PNAS* (2022)

// In class: spaCy (POS tagging and NER)

Unit 4: Putting it All Together

20. Tuesday, November 5th – EDA 1: Clustering
Due: Project proposal OR dataset
Read before class:

- Matthew Wilkens, "Genre, Computation, and the Varieties of Twentieth-Century US Fiction," *Cultural Analytics* (2016)
- Alexander Hoyle, Rupak Sarkar, et al, "Natural Language Decompositions of Implicit Content Enable Better Text Representations," *EMNLP* (2023)
- Optional: Ben Schmidt, "Clustering," from *Humanities Data Analysis* and/or "Machine Learning at Sea"

// In class: k-means clustering, maybe some other clustering algos if time

21. Thursday, November 7th – EDA 2: Topic Modeling
Read before class:

- Li Lucy, Dorottya Demsky, et al., "Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas US History Textbooks," *AERA Open* (2020)
- Melanie Walsh and Maria Antoniak, "The Goodreads 'Classics': A Computational STudy of Readers, Amazon, and Crowdsourced Amateur Criticism," *Journal of Cultural Analytics* (2021)
- Optional: Sathvika Anand, Quinn Dombrowski, and Xanda Schofield, "DSC #20: Xanda Rescues the Topic Model Disaster" (2023)

// In class: topic modeling w/ `gensim`

22. Tuesday, November 12 – Validation
Read before class:

- Rotem Dror, Gili Baumer, et al., "The Hitchhikers Guide to Testing Statistical Significance in Natural Language Processing," *ACL* (2018)
- Laura K. Nelson, "Situated Knowledges and Partial Perspectives: A Framework for Radical Objectivity in Computational Social Science and Computational Humanities," *New Literary History* (2022)
- Optional: Jason Brownlee, "A Gentle Introduction to the Bootstrap Method" (2019)
- Optional: Jared Wilber, "The Permutation Test: A Visual Explanation of Statistical Testing" (2019)
- Optional: Andres Karjus, "Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence," arXiv preprint (2023)

// In class: validation in NLP (ablation, permutation testing, etc)

23. Thursday November 14 – NO CLASS MEETING, professor at conference
Due: datasheet

24. Tuesday, November 19th – Chaining Methods
Read before class:

- Maria Antoniak, David Mimno, and Karen Levy, "Narrative Paths and Negotiation of Power in Birth Stories," *CSCW* (2019)

25. Thursday, November 21st – Guest lecture TBD – async!
Due: Final Project First Pass
Read before class:

- TBD

26. Tuesday, November 26th – Guest lecture TBD – async!
Read before class:

- TBD

27. Thursday, November 28th – NO CLASS - Thanksgiving

Unit 5: Final Project Presentations and Course Wrap-Up

28. Tuesday, December 3rd – Final presentations

29. Thursday, December 5th – Final presentations – async!

30. Tuesday, December 10 – Course Wrap-Up

FINAL PROJECT DUE THURSDAY, DECEMBER 12th