Emory University
Department of Quantitative Theory and Methods

QTM 340 (Fall 2021)
Practical Approaches to Data Science with Text
T/Th 1:00-2:15pm, Rich Building 211

Professor: Lauren Klein (lauren.klein@emory.edu)
Grader: Mack Hutsell (mack.hutsell@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 2:30-3:30pm, Callaway N310 or via Zoom
Wednesdays 1:15-2:15pm, Zoom only
Book here: https://lkle.in/officehours

Course Description
What does it mean to turn text into data? What are the data science 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 existing methods of text analysis can be used in new and creative ways. These methods include text parsing, natural language processing, language models, and vector space models, as well as statistical approaches including cluster analysis and supervised and unsupervised learning. We will also discuss contemporary topics including data ethics, data justice, and issues with "humans in the loop."

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:

- Reading assignments and canvas discussions: 10%
- Homeworks: 20%
- Quizzes: 25%
- Final project preparation assignments: 15%
- Final project: 30%

Description of Graded Assignments
Reading Assignments
You will be reading a wide range of texts—some written clearly, some more dense; some short, some long. Because these texts will inform our in-class discussions—and what you, in particular, have to say about them—it is absolutely essential that you stay on top of the reading assignments and complete them before the start of each class. Reading assignments are assessed through classroom participation (more on this below), as well as the occasional canvas discussion post (more on this below too). At the end of the semester, you will receive a letter grade (A-F) that reflects your level of engagement with the reading assignments.

Canvas Discussions
In effort to stimulate in-class discussion, as well as to allow you to introduce new material into the course, we will be using the Canvas "Discussion" feature throughout the course. During the second week of the course, you will select two class meetings during which you will be required to find and share at least one relevant data science project that involves text, where "relevant" is very broadly conceived, and will be responsible for providing a short (i.e. 250 word) description of the project on Canvas, highlighting what makes it relevant to the class. Discussion posts are due by midnight on the night BEFORE the class meets, so that I have time to read them and incorporate them into the day's discussion. You will receive a whole letter grade (A, B, C, D, F) upon the completion of each canvas post.

Homework Assignments
Many of the skills-building exercises in this course can be completed on your own. In effort to capture some of the successes of last year's Zoom teaching as well as to preserve more class time for discussion, I have decided to assign some skills-related notebooks as homework. Your task in these cases is to work through the assigned notebook on your own, and submit your completed notebook via Canvas by the start of the day's class. You will receive a whole letter grade (A, B, C, D, F) upon the completion of each homework notebook.

Quizzes and Final Project Preparation (FPP) Assignments
Over the course of the semester, you will be completing nine short assessments. The first five, 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 four ("FPP" assignments on the syllabus) are designed to lead up to the final project, and may be submitted as a project group. These assessments differ from the homeworks in that they are more open-ended and assess your ability to conceptualize and implement a complete text analysis workflow or 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) on the basis of your contribution. Designated FPP assignments will receive written feedback as well.

Final Project
In addition to the assignments described 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.

Specific information about each major assignment will be distributed no later than two weeks before the due date.

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

Attendance, Punctuality, and Late/Skipped Assignments
In ordinary years, I allow three excused absences, no questions asked, with your grade beginning to be lowered with the fourth absence. However, due to the ongoing coronavirus pandemic, I do not want to pressure you to come to class if you might be sick. Therefore, this year, I will allow unlimited absences in this course.

With that said, you are responsible for finding out what was discussed in the course on any days that you miss. I do not provide copies of my lecture notes, although I do post all in-class notebooks online. In addition, beginning with the fourth absence, you must email me to let me know that you will be missing class for health reasons.

Finally, please be respectful to your fellow classmates and arrive on time. If you arrive more than 15 minutes late, you will be considered absent for that class meeting.

All assignments are mandatory. Should you 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 and consequently, a lower grade for the course. Should you need an extension, please contact me *in advance* to discuss your situation.

Grading Process
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.

Grading Rubrics

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.

Homework, Quiz, and FPP Grading

The rubrics for individual homework, quiz, and FPP assignments are created by me and shared with the grader. If you have questions about these rubrics or would like to see them, please contact me.

Final Project Grading

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.

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 Note about COVID
By the time this course begins, it will have been 18 months since the coronavirus pandemic began. I am operating under the assumption that we have all experienced our share of hardships in different ways, and that we are all—again, in different ways and to different degrees--

exhausted. My goal is to make this course meaningful and informative, and certainly not to contribute to our collective (and ongoing) difficulties. If you are having trouble related to the course in any way, *please let me know*. By the same token, we will all need to adapt as the situation continues to evolve. I ask for your flexibility and understanding should the format of the course or assignments need to change in response to new or unexpected developments.

Contacting your Professor and Grader

Both your professor and your grader can be reached via their Emory email addresses. We respond to email M-F 9am-5pm, and outside of those hours only if our schedules allow. Please allow 24 hours for a response, and 48 hours if your message is sent over the weekend.

In addition, I (Dr. Klein) can be reached via my research group's Slack. I've created a channel for QTM 340, which you can join via this link: https://lkle.in/qtm340slack. For quick questions and troubleshooting, Slack may be better than email.

Finally, for questions about a grade that you've received, please contact me (Dr. Klein) and NOT the grader. I will contact the grader on your behalf.

Class-by-Class Schedule

Class schedule subject to change.
Please consult Canvas for the most current class schedule.
A current version of all code for the course can also be found on GitHub.

Overview and Introduction

August 26 – What does it mean to be practical?
// In class: syllabus overview, transcription exercise

August 31 – What can you do with text?
Read: Li-Young Lee, "Persimmons"
// In class: close reading and Voyant exercise
DUE: Homework #0, video intro

September 2 – What can you do with text as data?
Read: Farhad Manjoo, "How Do You Know a Human Wrote This?"; spend at least 30 minutes playing AI Dungeon
DUE: Homework #1, introduce yourself with GPT-2
// In class: demo JupyterHub / Jupyter Notebook

September 7 – What *should* you do with text as data?
Read: Michael Whitmore, "Text: A Massively Addressable Object"; Emily M. Bender and Timnit Gebru et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"
// In class: discussion and more intro to Jupyter / Python

Unit 1: Turning Text into Data

September 9 – Platforms and People
Read: Lilly Irani, "Justice for 'Data Janitors'; Andrew Norman Wilson, *Workers Leaving the Googleplex* (watch video); Astrid Smith and Bridget Whearty, "All the Work You Do Not See" (PDF)
DUE: Homework #2, Intro to Jupyter
// In class: text to data exercise

September 14 – Web scraping
Read: Astead Herndon et al., "What Do Rally Playlists Say About the Candidates?"; Hanah Anderson and Matt Daniels, "Film Dialogue"
// In class: Web scraping and HTML parsing using Beautiful Soup

September 16 – APIs
Read: Xavier Adam, "An Illustrated Introduction to APIs" and "API Whispering 101"
DUE: Quiz #1: Scraping song lyrics from Genius.com
// In class: APIs

September 21 – Text parsing / regular expressions
Read: Joel Spolsky, "The Absolute Minimum Every Software Developer Absolutely, Positivitely Must Know About Unicode and Character Sets (No Excuses!); and optional, more technical version of the previous post; and also optional (but interesting!), Miriam Sweeney and Kelsea Whaley, "Technically White: Emoji Skin-Tone Modifiers as American Technoculture"
DUE: Quiz #2: Scraping song lyrics using the Genius API
// In class: Text parsing and regex with your song lyrics

Unit 2: Introductory Data Science with Text

September 23 -- Sentiment Analysis
Read: Ethan Reed, "Measured Unrest in the Poetry of the Black Arts Movement"; "Poems with Pattern and VADER, Part 1: Quincy Troupe"; "Poems with Pattern and VADER, Part 2: Nikki Giovanni"; Sujay Khandekar et al., "Opico: A Study of Emoji-first Communication in a Mobile Social App"
// In class: Sentiment analysis

September 28 – Natural Language Processing 101, day 1
Read: Patrick Juola, "How a Computer Program Helped Show J.K. Rowling Wrote A Cuckoo's Calling"; Milo Beckman, "These are the Phrases Each GOP Candidate Uses Most"
DUE: Quiz #3: Sentiment analysis of your song lyrics
// In class: word counts, n-grams, lexicons

September 30 – Natural Language Processing 101, day 2
Read: Maarten Sap et al., "Connotation Frames of Power and Agency in Modern Films"; Maria Antoniak et al, "Narrative Paths and Negotiation of Power in Birth Stories"
DUE: Homework #3, Named entity recognition (NER), part-of-speech (POS) tagging
// In class: Guest lecture, Maria Antoniak, Cornell

October 5 – Turning Words into Numbers
Read: Daniel Jurafsky and James H. Martin, "Vector Semantics & Embeddings": SECTIONS 6-6.3, from *Speech and Language Processing*
DUE: Quiz #4: Counting and collocating with emoji
// In class: Intro to scikit-learn

October 7 – (Textual) Information Retrieval
Read: Matt Daniels, "The Language of Hip Hop"; Daniel Jurafsky and James H. Martin, "Vector Semantics & Embeddings": SECTIONS 6.5-6.6, from *Speech and Language Processing*; and optional, Lauren Klein, "Dimensions of Scale"
// In class: TF-IDF and PMI/PPMI; intro of final project

[FALL BREAK]

Unit 3: Modeling Text as Data I

October 14 – Topic Modeling
Read: Lucy Li and David Bamman, "Gender and Representation Bias in GPT-3 Generated Stories"; Richard Jean So, "Consecration: The Canon and Racial Inequality," from *Redlining Culture* (PDF)
// In class: topic modeling

October 19 – Word Embeddings
Read: Lauren Klein and Sandeep Soni, "How Words Lead to Justice"; Laura K. Nelson, "Leveraging the Alignment Between Machine Learning and Intersectionality" (PDF)
DUE: Homework #4, word embedding notebook
// In class: Guest lecture, Dr. Sandeep Soni, UC Berkeley

October 21 – Data, Pandas, Projects, day 1
Read: Ben Schmidt, "Gendered Language in Teacher Reviews" (explore website); Anelise Hanson Shrout, "(Re)Humanizing Data: Digitally Navigating the Bellevue Almshouse"; (and optional) Jessica Marie Johnson, "Markup Bodies" (PDF)
DUE: Quiz #5, exploratory research exercise
// In class: pandas; project brainstorming session

October 26 – Data, Pandas, Projects, day 2
Read: Timnit Gebru et al., "Datasheets for Datasets"; Catherine D'Ignazio and Lauren Klein, "The Numbers Don't Speak for Themselves"
DUE: Final project prep (FPP) #1, formal project brainstorm
// In class: pandas ii; discussion of data and its limits

October 28 – Data, Pandas, Projects, day 3
Read: Melanie Walsh, "The Challenges and Possibilities of Social Media Data" (PDF); Colored Conventions Project, "Introduction to CCP Corpus"; COVID Black, Homegoing (explore website)
// in class: pandas iii; more project discussion, more data discussion

Unit 4: Modeling Text as Data II

November 2 – Classification, day 1
Read: Dan Sinykin and Edwin Roland, "Against Conglomeration: Nonprofit Publishing and American Literature after 1980"
DUE: Homework #5, classification pt 1
// In class: classification, pt 2

November 4 – Classification, day 2
Read: Terra Blevins et al., "Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression"
DUE: FPP #2, datasheet OR project proposal
// In class: Guest lecture, perhaps

November 9 – Clustering (and maybe more on neural networks tbd)
Read: Ben Schmidt, "Genre, Manifolds, and AI"; Matthew Wilkens, "Genre, Computation, and the Varieties of 20th Century U.S. Fiction" (PDF)
// In class: clustering

November 11 – BERT
Read: Ted Underwood, "How Predictable Is Fiction?"; Lucy Li and David Bamman, "Characterizing English Variation across Social Media Communities with BERT"
DUE: FPP #3, datasheet OR project proposal
// In class: sentiment analysis with BERT and next sentence prediction

November 16 – Arguing with models
Read: Dong Nguyen et al., "How we do things with words: Analyzing text as social and cultural data"
// In class: Guest lecture, Lucy Li, UC Berkeley

November 18 – Arguing with models
Read: Richard Jean So, "All Models are Wrong" (PDF); Safiya Noble, "Introduction" and "Searching for Black Girls" from *Algorithms of Oppression: How Search Engines Reinforce Racism* (PDF)
DUE: FPP #4: Final Project First Pass
// In class: discussion of models and their limits

[THANKSGIVING BREAK]

Unit 5: Final Projects and Course Wrap-Up

November 30 – Project Presentations

December 2 – Project Presentations

December 7 – Course wrap-up and assessment

December 9 -- Final project due

This syllabus draws from previous iterations of QTM 340 taught by myself and Dan Sinykin. It also incorporates materials and resources developed by Melanie Walsh, Jinho Choi, Alison Parrish, David Mimno, David Bamman, Ryan Cordell, *and* Ben Schmidt, *as well as suggestions from Heather Froehlich, Ted Underwood, Jacob Eisenstein, Jim Casey, Taylor Arnold, Lauren Tilton, Lisa Rhody, Eileen Clancy, and the Colored Conventions Project Team.*