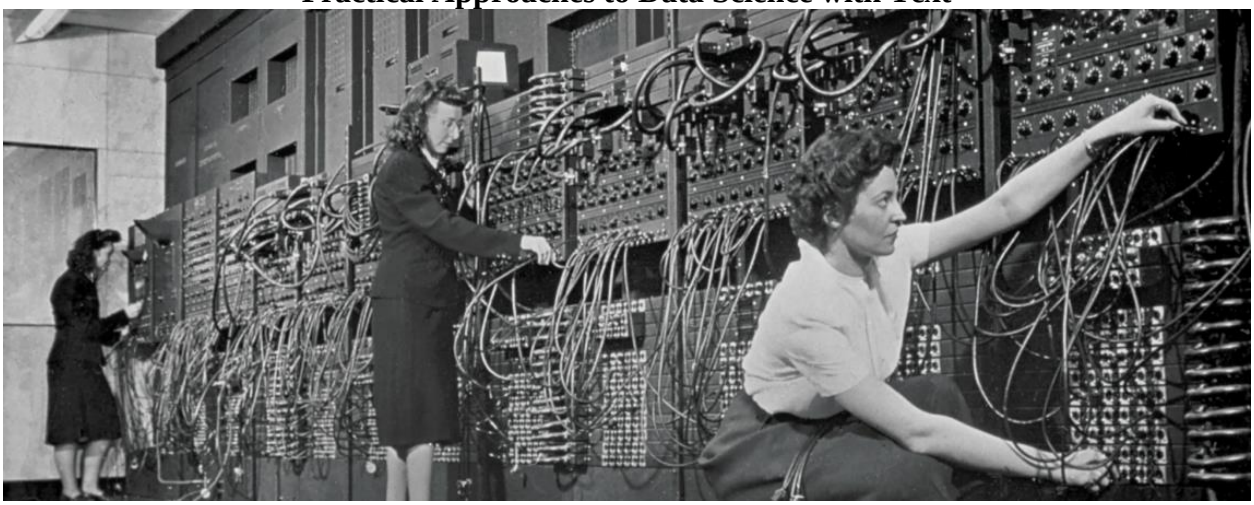
Emory University
Department of Quantitative Theory & Methods

QTM 340 (Fall 2020)
Practical Approaches to Data Science with Text



WHEN: T/Th 4:20pm-5:35pm

WHERE: ONLINE

WHO: Professor Dan Sinykin (daniel.sinykin@emory.edu)

Office Hours: Tuesdays 3-4pm, Zoom (and by appointment)

Prerequisites

QTM 210 or CS 171

Course Description

What does it mean to turn text into data? What are the data science techniques commonly employed 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 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 prerequisites. You will complete 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 require some short writing assignments.

Required Course Materials

All required readings are available online as links in this document and/or posted on Canvas.

Teaching and Learning during the Pandemic

This semester is unusual in that there is a pandemic. This class is being remotely taught. My goal is for all students to receive a high-quality experience to the extent possible. To that end, during the summer I participated in Emory University's workshops on online teaching methods. I cannot guarantee an experience that is identical to pre-pandemic semesters, but my goal is to treat all students equitably, to ensure grading is clear, consistent, and fair, and to teach the most exciting and engaging course possible.

Communication is important. I commit to responding to emails within 48 hours, and my intention to respond faster than that most of the time. I will be slower on weekends. If your situation changes regarding health, housing, or in any other regard with respect to your ability to participate in the class, please contact the appropriate Emory student support organization first and then me as soon as feasible. It is easier for me to address your needs if I know about them as soon as they arise. This does not mean I can successfully respond to every request, but I emphasize that my goal is to treat you all equitably and do what I can to help you succeed.

Office of Accessibility Services

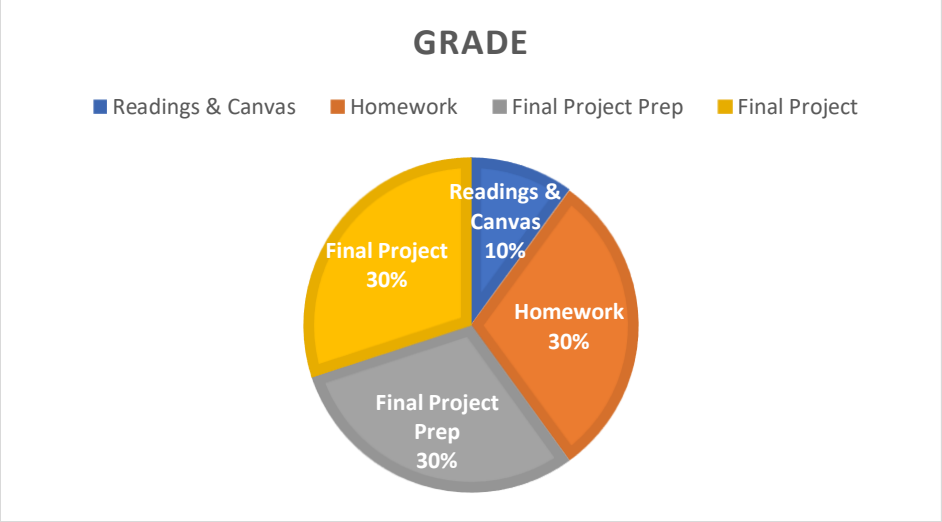
Your success in this class is important to me. We all need accommodations because we all learn differently. If there are aspects of this course that prevent you from learning or exclude you, let me know as soon as possible. Together we'll develop strategies to meet both your needs and the requirements of the course.

I encourage you to visit the Office of Accessibility Services to determine how you could improve your learning as well. If you need official accommodations, you have a right to have these met. 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.

List of Graded Assignments

Your grade for the course will be calculated as follows:

- Reading Assignments and Canvas Discussions: 30%
- 3 homework assignments: 20%
- 3 final project preparation assignments: 20%
- Final project: 30%



Description of Graded Assignments

Reading Assignments

You will read a wide range of texts—some written clearly, some more dense; some short, some long. Because these texts will inform our discussions—and what you, in particular, have to contribute—it is essential that you complete the reading before the start of each class. I assess reading assignments through participation and the occasional quiz.

Canvas Discussions

To stimulate discussion, and to invite you to introduce new material, we will use the Canvas Discussion feature throughout the course. During the second week, you will select two weeks when you will find and share at least one relevant data science project (broadly conceived) that involves text, and are responsible for providing a 250 word description of the project on Canvas, highlighting what makes it relevant to the class. Due eight hours before class time. You will receive a √+, √, or √- on the basis of your contribution.

Homework and Final Project Preparation Assignments

You will complete six small assignments. The first three are designed to enable you to put your newly-learned skills into practice, and must be submitted individually. The second three are designed to lead up to the final project, and may be submitted by your project group. All assignments must be submitted via Canvas by the beginning of class. You will receive a √+, √, or √- on the basis of your contribution. Designated homework and final project preparation assignments will receive written feedback.

Final Project

You will complete a final project: a fully-developed application of text analysis techniques to a research question of your own devising. I will ask you 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 on the basis of your contribution, and written feedback.

I will distribute information about each assignment no later than two weeks before the due date.

Attendance, Punctuality, and Late/Skipped Assignments

You are welcome to take three excused absences, no questions asked. But you are responsible for finding out what we discussed on days that you miss; I do not provide copies of lecture notes, but do make Jupyter notebooks available on GitHub after each course meeting. If you become sick, I will be flexible about attendance. If you are living in a time zone that conflicts with our synchronous sessions, contact me, and we will make an arrangement. Beginning with the fourth absence, your overall course grade will be lowered by a half letter grade (e.g. B to B-). Our class sessions on Zoom will be audio visually recorded for students to refer back to, and for enrolled students who are unable to attend live. Be respectful to your fellow students and arrive on time to synchronous sessions. If you arrive more than 10 minutes late, you will be considered absent. If you must miss class, contact me at least 24 hours in advance to make alternate arrangements.

All assignments are mandatory. Should you miss the due date, you are still welcome to submit the assignment for a grade that will decrease by a half 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, ask in advance.

Final Project Grading

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

GRADE	CHARACTERISTICS
A	**Exceptional**. The work is focused and its methods are sound. It clearly conveys the rationale behind its methodological choices as well as the stakes of its research question. The work demonstrates awareness of its implications and/or limitations, and it incorporates outside research when appropriate. The work reflects *in-depth* engagement with the topic.
B	**Satisfactory.** The work is reasonably focused and its methods are sound. It conveys the rationale behind its methodological choices as well as the stakes of its research question, but they are not fully developed. The work demonstrates some awareness of its implications and/or limitations. Fewer connections are made to outside research. The work reflects *moderate* engagement with the topic.
C	**Underdeveloped.** The work is mostly description or summary, without a consideration of the stakes of the research question. It does not consider the implications and/or limitations of the argument or methods, and few to no connections are made to outside research. The work reflects *passing* engagement with the topic.
D	**Limited.** The work is unfocused or incomplete, and displays *no evidence of student engagement* with the topic.
F	**No Credit.** The work is missing or consists of one or two disconnected paragraphs/charts/etc.

Writing Center Support

The Emory Writing Center staff of undergraduate tutors and graduate fellows is available remotely this fall to support Emory College students as they work on any type of writing assignment in any field: sciences, social sciences, or humanities. Tutors can assist with a range of projects, from traditional papers and presentations to websites and other multimedia projects. 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. Tutors also support the literacy needs of English Language Learners; several tutors are ELL Specialists, who have received additional training. The Writing Center opens for fall on August 31st, with hours throughout the day to accommodate students in various time zones. Learn more and make an appointment at writingcenter.emory.edu. Please note that you need to make (and cancel) appointments at least 3 hours in advance to accommodate our remote staff. Please review our tutoring policies, including our updated policies and procedures for online appointments, on our website 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.

Class-by-Class Schedule

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

Introduction and Overview

8/20 – SYNCHRONOUS

What does it mean to be practical?

In class: syllabus overview, intro/transcription exercise

8/25 – SYNCHRONOUS

What can you do with text?

Read: Li-Young Lee, "Persimmons"

Read: Michael Whitmore, "Text: A Massively Addressable Object"

In class: close reading and Voyant exercise

Unit 1: Turning Text into Data

8/27 – SYNCHRONOUS

JupyterHub and GPT-3

HW0: Initiate JupyterHub

Read: Farhad Manjoo, "How Do You Know a Human Wrote This?"

Spend: at least 30 minutes playing AI Dungeon

Canvas: Use GPT-2 to write your post. Go here. Delete the given text. Write the first 30-50 words of your post, then use tab and choose selections to let GPT-2 finish it.

9/1 – ASYNCHRONOUS

Platforms and People

Read: Lilly Irani, "Justice for 'Data Janitors'"

Notebook: Intro to Jupyter

9/3 – SYNCHRONOUS

Web Scraping

Read: Astead Herndon et al.,, "What Do Rally Playlists Say About the Candidates?"

Read: Hanah Anderson and Matt Daniels, "Film Dialogue"

Notebook: Web scraping and HTML parsing using Beautiful Soup

9/8 – ASYNCHRONOUS

APIs

Read: Xavier Adam, "An Illustrated Introduction to APIs" and "API Whispering 101"

Notebook: APIs

9/10 – SYNCHRONOUS

Text parsing / regular expressions

Read: David Zentgraf, "What Every Programmer Absolutely, Positively Needs to Know about Encodings and Character Sets to Work with Text"

Read: Scott Weingart, "The Route of a Text Message"

Notebook: Text parsing and regex

HW 1 Due: Scrape the lyrics of one candidate's campaign playlist from Genius.com

Unit 2: Introductory Data Science with Text

9/15 – ASYNCHRONOUS

Sentiment Analysis

Read: 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"

Read: Catherine D'Ignazio and Lauren Klein, "The Numbers Don't Speak for Themselves"

Notebook: song lyric corpus | sentiment analysis

9/17 – SYNCHRONOUS

Natural Language Processing 101 (NER, POS tagging)

Read: Lauren Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings"

Notebook: NER, POS tagging

9/22 – ASYNCHRONOUS

Turning Words into Numbers

Read: Daniel Jurafsky & James H. Martin, "Vector Semantics & Embeddings": SECTIONS 6-6.3

OPTIONAL: Aurelie Herbelot, "Distributional Semantics: A Light Introduction"

Notebook: intro to sklearn

HW 2 Due: Sentiment Analysis

9/24 – SYNCHRONOUS

tf-idf

Read: Matt Daniels, "The Language of Hip Hop"

Read: Daniel Jurafsky & James H. Martin, "Vector Semantics & Embeddings": SECTIONS 6.5-6.6

OPTIONAL: Milo Beckman, "These are the Phrases Each GOP Candidate Uses Most"

Notebook: word counts, tf-idf

Intro of final project

9/29 – ASYNCRONOUS

Topic Modeling

Read: Lauren F. Klein, "Dimensions of Scale"

Notebook: topic modeling

10/1 – SYNCHRONOUS

Word Embedding Models

Read: Ben Schmidt, "Gendered Language in Teacher Reviews"

Read: Ben Schmidt, "Rejecting the Gender Binary"

OPTIONAL: Daniel Jurafsky & James H. Martin, "Vector Semantics & Embeddings": SECTIONS 6.8-6.11

OPTIONAL: Sarah Connell, "Word Embedding Models are the New Topic Models"

Notebook: word2vec

10/6 – ASYNCHRONOUS

Pandas

Read: Anelise Hanson Shrout, "(Re)Humanizing Data: Digitally Navigating the Bellevue Almshouse"

Notebook: pandas, pt 1

HW3 Due: Experimental Design

Final Project Brainstorming Sheet Due

10/8 – SYNCHRONOUS

Data

Read: Heather Krasue, "Data Biographies: Getting to Know Your Data"

Read: Timnit Gebru et al., "Datasheets for Datasets"

OPTIONAL: Sarah Allison, "Other people's data: Humanities edition"

Unit 3: Intermediate Data Science with Text

10/13 – ASYNCHRONOUS

Modeling, pt 1

Read: David Smith and Ryan Cordell, "Mass Digitization" and "What is Text, Probably?"

Notebook: pandas, pt 2

10/15 – SYNCHRONOUS

Modeling, pt 2

Read: Safiya Noble, "Introduction" and "Searching for Black Girls" from *Algorithms of Oppression: How Search Engines Reinforce Racism*

Read: Richard Jean So, "All Models are Wrong"

Notebook: pandas, pt 3

10/20 – ASYNCHRONOUS

Classification, pt 1

Read: Dan Sinykin and Edwin Roland, "Against Conglomeration: Nonprofit Publishing and American Literature after 1980"

Notebook: classification, pt 1

HW4 Due: Datasheet

10/22 – SYNCHRONOUS

Classification, pt 2

Read: Terra Blevins et al., "Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression"

Notebook: classification, pt 2

10/27 – ASYNCHRONOUS

Clustering

Read: Matthew Wilkens, "Genre, Computation, and the Varieties of 20th Century U.S. Fiction"

Notebook: clustering

10/29 – SYNCHRONOUS

BERT

Read: Ted Underwood, "How Predictable Is Fiction?"

Notebook: sentiment analysis with BERT | next sentence prediction

HW5 Due: Project Proposal

11/3 – ASYNCHRONOUS

NO CLASS: ELECTION DAY

Unit 4: Arguing with Textual Data

11/5 – SYNCHRONOUS

Making arguments

Read: Dong Nguyen et al., "How we do things with words: Analyzing text as social and cultural data"

OPTIONAL: Ted Underwood, David Bamman, and Sabrina Lee, "The Transformation of Gender in English Language Fiction"

11/10 – ASYNCHRONOUS

Validation

Read: Matthew Salganik, "Validation," from *Bit by Bit: Social Research in the Digital Age*

HW6 Due: Initial Test

OPTIONAL: On replication

Susan Dominus, "When the Revolution Came for Amy Cuddy"

Christine R. Harris, et al., "Two Failures to Replicate High-Performance-Goal Priming Effects"

Andrew Goldstone, "Of Literary Standards and Logistic Regression: A Reproduction"

11/12 – SYNCHRONOUS

Conferences

11/17 – SYNCHRONOUS

Project presentations

11/19 – SYNCHRONOUS

Project presentations

11/24 – SYNCHRONOUS

Course wrap-up and assessment

FINAL PROJECTS DUE FRIDAY, 12/4, 5:30PM

Lauren F. Klein wrote and designed the first version of this syllabus, inspired by the courses of Jinho Choi, Alison Parrish, David Mimno, David Bamman, Ryan Cordell, and Ben Schmidt, as well as suggestions and other input from Heather Froehlich, Ted Underwood, Jacob Eisenstein, Jim Casey, Taylor Arnold, Lauren Tilton, Lisa Rhody, Eileen Clancy, and the Colored Conventions Project Team. I supplemented the syllabus with inspiration and a great deal of language and code from Melanie Walsh's online textbook Introduction to Cultural Analytics & Python.