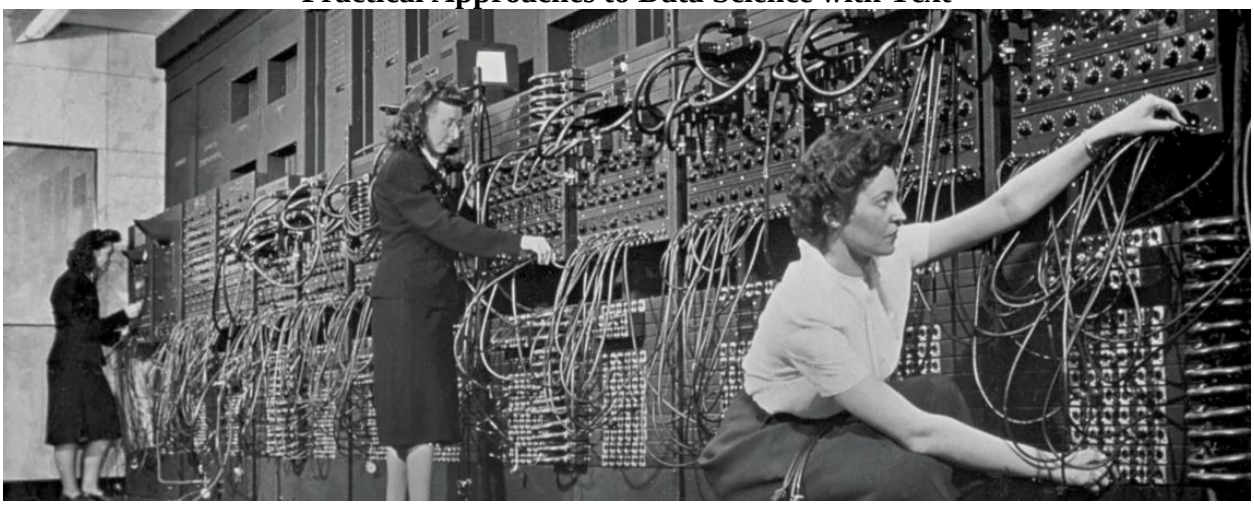
Emory University
Department of Quantitative Theory & Methods

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

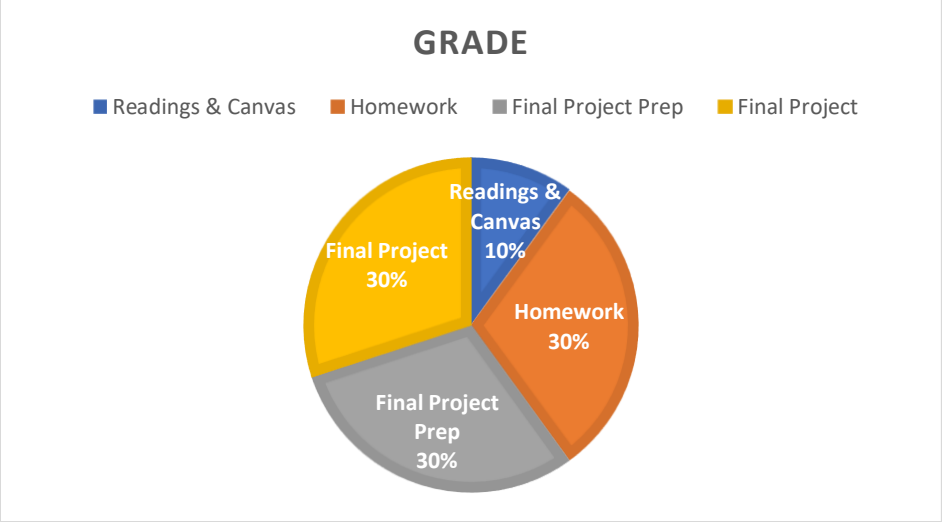
Your success in this class is important to me. We all need accommodations because we all learn differently. If there are aspects of this course that prevent you from learning or exclude you, let me know as soon as possible. Together we'll develop strategies to meet both your needs and the requirements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

GitHub

TBD

HW 0 Due: Install Anaconda & GitHub

9/1 – ASYNCHRONOUS

Platforms and People

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

Notebook: Intro to Jupyter

Canvas: discussion of Irani

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

Canvas: discussion of Anderson and Daniels

9/8 – ASYNCHRONOUS

APIs

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

Notebook: APIs (ex: Genius and Twitter)

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"

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

In class: Text parsing and regex with your song lyrics

Unit 2: Operationalizing Text as Data

9/15 – ASYNCRONOUS

Sentiment analysis (and dictionaries more generally)

Read: Ethan Reed, "Measured Unrest in the Poetry of the Black Arts Movement"

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

In class: Sentiment analysis and discussion of context

9/17 – SYNCHRONOUS

sklearn: countvectorizer

Read: TBD on Turning Words into Numbers

Data TBD

9/22 – ASYNCHRONOUS

Intro of final project

HW 2 Due: TBD

9/24 – SYNCHRONOUS

Word counts, tf-idf

Read: Charlie Smart, "The Differences in How CNN, MSNBC, & Fox Cover the News"

9/29 – ASYNCRONOUS

Natural Language Processing (NER, POS tagging, etc)

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

Read: Ishan Misra et al., "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels"

10/1 – SYNCHRONOUS

Topic modeling

Read: Cameron Blevins, "Topic Modeling Martha Ballard's Diary"

Read: Lisa Rhody, "Topic Modeling and Figurative Language"

10/6 – ASYNCHRONOUS

Word vectors

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

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

10/8 – SYNCHRONOUS

BERT (Bidirectional Encoder Representations from Transformers)

TBD

Unit 3: (More) Modeling Textual Data

10/13 – ASYNCHRONOUS

Pandas / Final Project Brainstorming Session

HW 3 Due: Analyzing the Colored Conventions Project Corpus

10/15 – SYNCHRONOUS

Another Look at Data

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

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

10/20 – ASYNCHRONOUS

Language Models

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

10/22 – SYNCHRONOUS

Similarity

Read: Patrick Juola, "How a Computer Program Helped Show J.K. Rowling Wrote A Cuckoo's Calling"

(Optional) more technical version: Patrick Juola, "Rowling and 'Galbraith': An Authorial Analysis"

FPP 1 Due: Datasheet OR Data Biography

10/27 – ASYNCHRONOUS

Classification

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

10/29 – SYNCHRONOUS

Neural Networks

TBD

11/3 – ASYNCHRONOUS

Clustering

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

Read: Alexis Madrigal, "How Netflix Reverse Engineered Hollywood"

FPP 2 Due: Project Proposal

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"

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

11/10 – ASYNCHRONOUS

Validation, day 1

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

FPP 3 Due: Annotated bibliography

11/12 – SYNCHRONOUS

Validation, day 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"

11/17 – SYNCHRONOUS

Project presentations

11/19 – SYNCHRONOUS

Project presentations

11/24 – SYNCHRONOUS

Course wrap-up and assessment

FINAL PROJECTS DUE DECEMBER 16TH, 5:30PM

In the spirit of the Honor Code, I acknowledge Lauren F. Klein, who wrote and designed the first version of this syllabus. She drew on the syllabi of Jinho Choi, Alison Parrish, David Mimno, David Bamman, Ryan Cordell, and Ben Schmidt, as well as input from Heather Froehlich, Ted Underwood, Jacob Eisenstein, and Jim Case.