Emory University
Department of Quantitative Theory & Methods

QTM 340 (Fall 2019)
Practical Approaches to Data Science with Text
T/Th 11:30am-12:45pm, North Decatur Building 155

Professor: Lauren Klein (lauren.klein@emory.edu)
Office Hours: Tuesdays 1-2pm, Callaway N310 (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 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. Contemporary topics including data ethics, data justice, and issues with "humans in the loop" are also discussed. Introductory courses in computer science and probability and statistics are recommended as perquisites for QTM 340. All class exercises and homework assignments are done in Python. Students are expected to participate in class discussion and present their final projects at the end of the semester. Some short writing assignments are also required.

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

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

- Reading Assignments and Canvas Discussions: 10%
- 3 homework assignments: 30%
- 3 final project preparation assignments: 30%
- 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 classroom discussions—and what you, in particular, have to contribute—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, as well as the occasional quiz.

Canvas Discussions

In effort to stimulate classroom 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 weeks during which you will be required to find and share at least one relevant data science project (broadly conceived) that involves text, and are responsible for providing a short (i.e. 250 word) description of the project on Canvas, highlighting what makes it relevant to the class. Due by midnight on the night BEFORE the class meets.

You will receive a √+, √, or √- on the basis of your contribution. Students seeking additional feedback on their post should schedule a meeting with the professor during office hours.

Homework and Final Project Preparation Assignments
Over the course of the semester, you will be completing 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 in 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 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 assignment will be distributed no later than two weeks before the due date.

Attendance, Punctuality, and Late/Skipped Assignments
You are allowed three excused absences, no questions asked. However, you are responsible for finding out what was discussed in the course on any days that you miss; I do not provide copies of lecture notes, but Jupyter notebooks will be made available on GitHub after each course meeting.

Beginning with the fourth absence, your overall course grade will be lowered by a half letter grade (e.g. B to B-) for each unexcused absence.

Please be respectful to your fellow students and arrive on time. If you arrive more than 15 minutes late, you will be considered absent for that class. If you absolutely must miss a class meeting, please contact me at least 24 hours in advance in order to make alternate arrangements.

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

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.

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.

Class-by-Class Schedule

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

Introduction and Overview

8/29 – What does it mean to be practical?

In class: Syllabus overview, intro/transcription exercise

9/3 – 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

9/5 – Platforms and People

Read: Scott Weingart, “The Route of a Text Message”
Read: Lilly Irani, “Justice for ‘Data Janitors’”

HW 0 Due: Install Anaconda/Jupyter

In class: Intro to Jupyter and discussion of the platforms and people that make data science with text possible.

9/10 – Web Scraping

Read: Astead Herndon et al.,, “What Do Rally Playlists Say About the Candidates?”
Read: David Zentgraf, “What Every Programmer Absolutely, Positively Needs to Know about Encodings and Character Sets to Work with Text”

In class: Web scraping and HTML parsing using Beautiful Soup

9/12 – APIs

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

In class: APIs (ex: Genius and Twitter)

9/17 – Text parsing / regular expressions

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

9/19 – 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

Unit 2: Operationalizing Text as Data

9/24 – Intro to Colored Conventions Project Corpus

Read: P. Gabrielle Foreman, Sarah Patterson, and Jim Casey, "Introduction to the Colored Conventions Movement" and "CCP Principles"
Watch: "The Colored Conventions Project in Three Videos" (watch the three videos)

9/26 – Topic modeling

Read: Cameron Blevins, "Topic Modeling Martha Ballard's Diary"
Read: Lisa Rhody, "Topic Modeling and Figurative Language"

10/1 – Intro of final project

HW 2 Due: Contributing back to the Colored Conventions Project

10/3 – Word counts, n-grams

Read: Patrick Juola, "How a Computer Program Helped Show J.K. Rowling Wrote A Cuckoo's Calling" (and more technical version)
Read: Charlie Smart, "The Differences in How CNN, MSNBC, & Fox Cover the News"

10/8 –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"

10/10 – Word vectors

Read: Sarah Connell, "Word Embedding Models are the New Topic Models"
Read: Ben Schmidt, "Gendered Language in Teacher Reviews"

[FALL BREAK]

Unit 3: (More) Modeling Textual Data

10/17 – Final Project Brainstorming Session

HW 3 Due: Analyzing the Colored Conventions Project Corpus

10/22 – Another Look at Data

Read: Heather Krasue, "Data Biographies: Getting to Know Your Data"
Read: Timnit Gebru et al., "Datasheets for Datasets"

10/24 – Language Models

Read: David Smith and Ryan Cordell, "Mass Digitization" and "What is Text, Probably?"
Read: Richard Jean So, "All Models are Wrong"

10/29 – 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

10/31 – Classification

Read: Terra Blevins et al., "Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression"
Read: Hoyt Long and Richard Jean So, "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning"

11/5 – Clustering

Read: Matt Daniels, "The Language of Hip Hop"
Read: Alexis Madrigal, "How Netflix Reverse Engineered Hollywood"

11/7 – NO CLASS, Professor at conference

FPP 2 Due: Project Proposal

Unit 4: Arguing with Textual Data

11/12 – 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"

11/14 – Testing (statistical) hypotheses

Read: Nan Z. Da, "The Computational Case Against Computational Literary Studies"
Read: Ted Dunning, "Surprise and Coincidence" (original paper)
Watch: Jonathan Stray, "Solve Every Statistics Problem with One Weird Trick"

11/19 – Challenging Tests and Testing

Adam Kilgarriff, "Language is Never, Ever, Ever, Random"
David Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis"

FPP 3 Due: Annotated bibliography

11/21 – Validation by whom, for whom?

Read: Matthew Salganik, "Validation," from *Bit by Bit: Social Research in the Digital Age*
Read: Safiya Noble, "Introduction" and "Searching for Black Girls" from *Algorithms of Oppression: How Search Engines Reinforce Racism*

11/26 – NO CLASS, Thanksgiving

[THANKSGIVING BREAK]

12/3 – Project presentations

12/5 – Project presentations

12/10 – Course wrap-up and assessment

FINAL PROJECTS DUE DECEMBER 17TH, 5:30PM

In the spirit of the Honor Code, I acknowledge 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 Casey, which have informed the schedule of this course.