YData: Humanities Data Mining

S&DS 176 / S&DS 576, Spring 2021
Tuesdays & Thursdays, 1:00–2:15 EST

Instructors: Dr. Catherine DeRose & Dr. Douglas Duhaime
Contact information: catherine.derose@yale.edu & douglas.duhaime@yale.edu
Undergraduate Learning Assistants (ULAs): Alan George & Patrycja Gorska

Course Description

What new modes of inquiry become available when we transform novels into bags of words and images into pixels? What is lost in the process? This course explores how we can use computational methods to pursue questions in the humanities, while also looking at how humanistic methods can inform the work of algorithms in research and society at large.

We begin this course with a series of questions at the intersections of the humanities and quantitative analysis: What is data? How can we turn texts into data? To explore these questions from both theoretical and technical perspectives, each course week will be divided into discussion and lab sessions. Discussion sessions will introduce concepts and humanities-based case studies that will ground the hands-on technical work we'll do in the labs. Over the course of the semester, we'll survey some of the most popular methods in modern data science—classification, vectorization, and visualization—to see what kinds of questions we can ask and answer. We'll conclude the semester with Open Lab sessions during which you will leverage the skills covered in this course to create your own data science projects with cultural heritage data.

Learning Objectives

Through this course, you will:

- Develop an understanding of data—its history, contexts, and uses—in the (digital) humanities
- Learn how to prepare datasets of varying scales
- Identify appropriate methods for studying the available data and research questions at hand
- Gain experience programming data-specific tasks in Python and using out-of-the-box visualization software
- Design your own data-centric projects from the ground up
- Establish a foundation for continuing with courses in digital humanities and data science

Assignments & Grading

20% PROBLEM SETS (5 total, each worth 4%)
Problem sets are intended to help reinforce some of the programming concepts covered in class. As you work to complete problem sets, you should feel encouraged to use the following resources in case you need guidance or help:

- Office Hours & Email: You can contact the course instructors to ask for guidance on any of the Problem sets.
- Online resources: Websites like StackOverflow.com are terrific places to ask questions about programming.
- Classmates: Working with classmates on problem sets is a great way to increase understanding of programming concepts and tasks.

Problem sets will be graded based on their completion. A submission is considered complete if it includes either the output from the assigned task(s), or the error message that was received when trying to accomplish the tasks, with your thoughts on what might be contributing to the error and how it could be corrected (even if you couldn't get it to work!). Your submissions will help us identify concepts that might need to be revisited during lab sessions.

Problem sets be submitted to our Canvas site before class starts.

10% PROJECT REVIEWS (2 total, each worth 5%)

Project reviews consist of a 2-page response to a digital humanities project we'll be discussing in class. In addition to being case studies that might provide inspiration or insight for your own projects later in the term, these reviews are designed to help breakdown the various components that go into a computational project in the humanities.

For each review, we'll ask you to consider:

- What is the data that is being used?
- What method(s) does the project employ?
- What kinds of questions does that method help you ask of the data?
- What is the visualization trying to communicate, and why is or isn't it effective?
- Who makes up the project team (how many people, what are their areas of expertise)?

Short writing assignments should be submitted to our Canvas site before class starts.

20% PARTICIPATION

One primary goal of the class is to increase your familiarity and comfort with computational methods. Each week, we will introduce different digital humanities tools and approaches that you will be asked to try out. You will *not* be graded on your mastery of those tools or approaches, but rather on your engagement with them, your willingness to try them and experiment.

Digital humanities is a collaborative area of research that brings together scholars from across disciplines to drive the field forward. We learn from each other. We engage with each other's ideas, and we help to problem solve when we encounter a technical or methodological issue. Participation in this seminar includes coming to class prepared and contributing to class discussions by asking questions, sharing ideas, listening attentively, and offering support.

We understand that talking in class is not always easy. If you are uncomfortable or unsure of how to enter class discussions, email us or talk to us after class and we can discuss strategies and alternative ways to engage with observations raised in class. The participation grade will be jointly determined by your contributions to classroom discussion and your active attendance.

20% MIDTERM: LITERATURE REVIEW

The midterm assignment will consist of a 2-3-page literature review that surveys 4-5 papers or projects that use a particular computational method. The goal of this assignment is to help you develop some expertise in the ways a particular method has been leveraged in humanities scholarship, which could inform the design of your final project.

Graduate students in the course should submit, in addition to this literature review, a dataset they could use to create their own study using the selected technical method.

The literature review should be submitted to our Canvas site before class starts.

5% PRESENTATION

As a capstone to the seminar, you will be asked to give a 5-minute presentation on the project you've designed—what was your guiding research question, how did you approach it using a technique covered in class, and what did you find out. Along with receiving feedback on your project, you'll gain experience presenting digital humanities research to a disciplinarily diverse audience. We encourage you to use slides or demo your project live.

25% FINAL PROJECT

The final project will consist of a 3-4-page paper that uses one or more of the techniques covered in class to advance a historically-inflected argument. This project will be your opportunity to use the techniques covered in class to craft an original, data-driven argument about a historical dataset. You are encouraged, but not required, to use the scholarship and/or dataset you generated for the course midterm in the creation of this project. You are also encouraged to make liberal use of data visualizations and data tables (these do not count toward the page length) in your final project.

The final project should be submitted to our Canvas site before midnight on the due date.

Attendance Policy

Since your participation is crucial to the success of the class, regular, on-time attendance is required and will count toward your participation grade. That said, we recognize that there may be times when you need to miss a class, so we have built-in two excused absences—we do not need to be contacted in advance, nor do we need a reason for why you're missing class. If you require missing a class beyond two days, please contact us 24 hours in advance or your participation grade will be reduced for that day.

Late Work

Problem Sets, Project Reviews, and the Midterm are due by the start of class time on the day they're due. Credit for late work will be reduced by half a letter grade for each 24-hour window after the due date.

The Presentation and Final Project will not be accepted late, unless the instructors are contacted a week in advance or a Dean's Excuse is presented for an extension.

Academic Integrity

All writing for this class should be your own work, unless otherwise cited (any recognizable citation scheme may be used so long as it is consistent). For help with citations, please email us or reach out to the Poorvu Center for Teaching and Learning: https://poorvucenter.yale.edu/. For Yale College's academic integrity statement and policy, visit: http://catalog.yale.edu/undergraduate-regulations/regulations/academic-dishonesty/.

When it comes to coding, it is often good practice to refer to and reuse code you find from trusted places online. Since one goal for this class is to increase your experience with Python, we encourage you to start any given exercise by writing as much of the code yourself as you can before turning to the internet. If you then get stuck, Stack Overflow tends to be a reliable source to look for help (and, of course, you can reach out to us and the ULAs).

Texts and Software

All readings for this class are open access or available online through Yale University Library. Along with readings, we will analyze existing projects that typify different digital humanities approaches. Links to the readings and projects (along with lab tutorials, datasets, and recommended resources) can be found on our course GitHub site: https://github.com/YaleDHLab/humanities-data-mining

For the lab sessions each week, you are expected to work from a laptop or desktop computer (Mac or PC) over which you have administrative control. If that won't be possible, please let us know as soon as possible so that we can find an alternative. We will be working primarily with Google Colab notebooks in the class, meaning you generally won't have to install software. No prior programming knowledge is required; we will introduce all concepts and tools starting from the beginning. All of the software we'll be using is free.

Virtual Office Hours

We would be very glad to meet with you via Zoom outside of class time to discuss ideas or questions, or to provide additional technical help or recommendations on follow-up readings or tutorials. Our scheduled office hours are on:

- Fridays from 10:00–11:00 am EST (Catherine and Doug)
- Sundays from 2:00–4:00 pm EST (Alan)

- Sundays from 3:00–5:00 pm EST (Patrycja)
- By appointment (Catherine and Doug)

For the Zoom, use our regular class link. The above times are subject to change if we find that a different drop-in time would work better with your schedules. We are also available via email for quick questions or to set up an appointment.

In addition to our office hours, you can also find support at the Digital Humanities Lab's Office Hours, where DHLab staff and graduate student consultants discuss ideas, troubleshoot issues, and share online resources and tutorials.

Schedule

* *content may shift in response to students' interests*

Week	Day	In-Class Topic	Homework Due by Start of Class
Week 1	Tuesday Feb 2	Introduction: Computational Digital Humanities	Readings: • Michael Witmore, "Text: A Massively Addressable Object" • Ted Underwood, "Seven ways humanists are using computers to Understand Text"
	Thursday Feb 4	Lab – Intro to Python, 1	
Week 2	Tuesday Feb 9	What Counts as "Data"	Readings: • Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities" • Johanna Drucker, "Why Distant Reading Isn't" Coding: Problem Set 1
	Thursday Feb 11	Lab – Intro to Python, 2	
Week 3	Tuesday Feb 16	Introduction to Data Visualization	Readings: • Catherine D'Ignazio and Lauren Klein, "Feminist Data Visualization" Coding: Problem Set 2

	Thursday Feb 18	Lab – Tableau	Sign up for free Tableau Public accounts and download software
Week 4	Tuesday Feb 23	Humanities Text Analysis 1 – Intro	Readings: • Richard Jean So, "All Models are Wrong" • Jean Baptiste-Michel et al. "Quantitative Analysis of Culture Using Millions of Digitized Books" Writing: Project Review 1
	Thursday Feb 25	Lab – Named Entity Recognition	
Week 5	Tuesday Mar 2	Humanities Text Analysis 2 – Vector Space Models (Clustering & Classification)	Readings: • Patrick Juola, "How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling" • Franco Moretti, "The Slaughterhouse of Literature" in *Distant Reading*
	Thursday Mar 4	Lab – Supervised Classification with Scikit-Learn	
Week 6	Tuesday Mar 9	SPRING BREAK DAY – no class	
	Thursday Mar 11	Lab – Mid-Semester Recap	Coding: Problem Set 3
Week 7	Tuesday Mar 16	Humanities Text Analysis 3 – Topic Modeling	Readings: • Cameron Blevins, "Topic Modeling Martha Ballard's Diary" Writing: Midterm Literature Review
	Thursday Mar 18	Lab – Topic Modeling	
Week 8	Tuesday Mar 23	Text & Image Analysis: Neural Networks	Readings: • Gideon Lewis-Kraus, "The Great A.I. Awakening" • Sarah Connell, "Word Embeddings are the New Topic Models"

			Coding: Problem Set 4
	Thursday Mar 25	Lab – Word Embeddings	
Week 9	Tuesday Mar 30	Computer Vision 1 – Color & Art	Readings: • Tim Wallace and Krishna Karra, "The True Colors of America's Political Spectrum are Gray and Green" Coding: Problem Set 5
	Thursday Apr 1	Lab – Color Extraction	
Week 10	Tuesday Apr 6	Computer Vision 2 – Visual Similarity	Readings: • Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images" Writing: Project Review 2
	Thursday Apr 8	SPRING BREAK DAY – no class	
Week 11	Tuesday Apr 13	Lab – Image Similarity	
	Thursday Apr 15	Computer Vision 3 – Video	Readings: • Taylor Arnold and Lauren Tilton "Distant Viewing: Analyzing Large Visual Corpora"
Week 12	Tuesday Apr 20	Lab – Distant Viewing Lab	
	Thursday Apr 22	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	
Week 13	Tuesday Apr 27	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	

	Thursday Apr 29	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	
Week 14	Tuesday May 4	Presentations	
	Thursday May 6	Presentations & Next Steps	
Finals Week	Friday May 14	No class	Final Project