YData: Humanities Data Mining

S&DS 176 / S&DS 576, Spring 2022
Tuesday & Thursday 1:00-2:15pm

Instructor: Peter Leonard
Email: peter.leonard@yale.edu
Office: Sterling Memorial Library 176D (Franke Family Digital Humanities Lab)

Course Description

What new modes of inquiry become available when we transform narrative into counts of words, and art into pixels? What is lost – and gained – in this process? *Humanities Data Mining* explores how scholars are using computation and algorithmic analysis to pursue humanities research questions, while also looking at how humanistic perspectives can inflect the role of algorithms in the academy – and society at large.

We start with a question that underlies most work in the digital humanities: what counts as data? How can we transform literature and painting into something computationally addressable? To explore these questions from theoretical and ad technical perspectives, each week will include both a discussion and a lab section. Discussion sessions introduce concepts and humanities case studies which 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 we'll use the skills we covered in the course to create data science projects of our own with humanities data.

Humanities Data Mining was created in 2021 by Douglas Duhaime and Catherine DeRose, working in the Yale Digital Humanities Lab. This is the second year the course has been offered.

Learning Objectives

1. Develop an understanding of data in the humanities: its history, contexts and uses.
2. Learn how to prepare datasets of varying scales.
3. Identify appropriate methods for studying the available data and research questions at hand.
4. Gan experience with data-specific tasks in Python, and with data visualization software such as Tableau.
5. Design your own data-centric projects from the ground up.
6. Establish a foundation for continuing with courses in digital humanities and data science.

Assignments and Grading

20% Problem Sets (5 total; 4% each)

Problem sets 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: Come to office hours to meet the instructor and TFs and ask questions in person.
- Online Resources: Google, and specialized communities such as Stack Overflow, are great places to search for 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 should be submitted to the Canvas site before class starts.*

10% Project Reviews (2 total; 5% each)

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.

Talking in class is not easy for everyone. If you are uncomfortable or unsure of how to enter class discussions, email me or talk to me 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 (and dataset for grad students)

The midterm assignment will consist of a 2-3-page literature review which 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 (and dataset, if appropriate) should be submitted to the 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. I 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 in your final project – although these do not count toward the page length.
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. Please contact me 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 you contact me 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). Consult Yale College's academic integrity statement if you have questions. For help with citations, reach out to the Poorvu Center for Teaching and Learning.

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 me and the ULAs).

Course Materials 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 GitHub:

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 me

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.

Office Hours

Scheduled office hours are:

- Tuesdays 10:00-11:00am (Peter Leonard)
  - https://schedule.yale.edu/appointments/peterleonard
- Mondays 2pm-3pm (TF Luna)
  - Link in Canvas
- Mondays 4pm-5pm (TF Sid)
  - Link in Canvas
- Sundays 3pm-5pm
  - (ULAs Kayla & Linh)

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. DHLab Office Hours for Spring 2022 are scheduled to be Monday through Thursday, 2pm to 3pm.

Week	Day	Topic	Due for class
Week 1	Tuesday January 25	Introduction: Computational Digital Humanities	Michael Witmore, "Text: A Massively Addressable Object" Ted Underwood, "Seven ways humanists are using computers to understand text"
	Thursday January 27	Lab: Intro to Python Part 1	
Week 2	Tuesday February 1	What counts as data?	Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities" Johanna Drucker, "Why Distant Reading Isn't" Coding: Problem Set 1
	Thursday February 3	Lab: Intro to Python Part 2	
Week 3	Tuesday February 8	Introduction to Data Visualization	Catherine D'Ignazio and Lauren Klein, "Feminist Data Visualization"

			Coding: Problem Set 2
	Thursday February 10	Lab: Tableau	Sign up for free Tableau Public accounts and download software to your laptop.
Week 4	Tuesday February 15	Humanities Text Analysis: Introduction	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 February 17	Lab: Named Entity Recognition	
Week 5	Tuesday February 22	Humanities Text Analysis: Vector Space Models; Clustering & Classification	Patrick Juola, "How a Computer Program Helped Show J.K. Rowling write *A Cuckoo's Calling*" [sic] Franco Moretti, "The Slaughterhouse of Literature" in *Distant Reading*
	Thursday February 24	Lab: Supervised Classification with Scikit-learn	
Week 6	Tuesday March 1	Humanities Text Analysis: Topic Modeling	Cameron Blevins, "Topic Modeling Martha Ballard's Diary" Coding: Problem Set 3
	Thursday March 3	Lab: Topic Modeling	
Week 7	Tuesday March 8	Text & Image Analysis: Neural Networks	Gideon Lewis-Kraus, "The Great A.I. Awakening" Jonathan Fitzgerald, "Word Embeddings are the New Topic Models" Coding: Problem Set 4
	Thursday March 10	Lab: Word Embeddings	
Week 8	Tuesday March 15	Computer Vision: Color & Art	Tim Wallace and Krishna Karra, "The True Colors of America's Political Spectrum are Gray and Green" Coding: Problem Set 5
	Thursday March 17	Lab: Color Extraction	

Week 9		Spring Recess	
Week 10	Tuesday March 29	Computer Vision: Image Similarity	Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images" Writing: Midterm Literature Review
	Thursday March 31	Lab: Image Similarity	
Week 11	Tuesday April 5	Computer Vision: Video	Taylor Arnold and Lauren Tilton "Distant Viewing: Analyzing Large Visual Corpora"
	Thursday April 7	Lab: Distant Viewing Lab	
Week 12	Tuesday April 12	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	
	Thursday April 14	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	Writing: Project Review 2
Week 13	Tuesday April 19	Open Lab – opportunity to revisit concepts and tools from earlier in the semester and time to work on final projects	
	Thursday April 21	Presentations	
Week 14	Tuesday April 26	Presentations & Next Steps	
	Thursday April 28	No class	Final Project