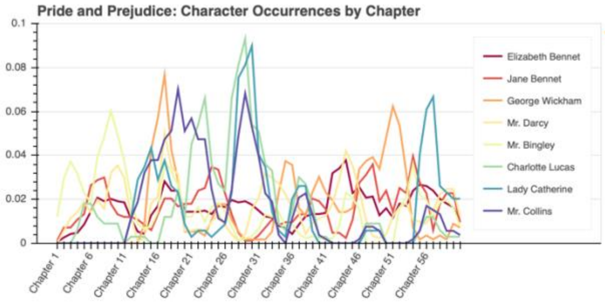
Literary Text Mining; or, Introduction to Computational Literary Studies

English 184E
Autumn 2023, Stanford University, 5 units
Time: Tuesday and Thursday at 3-4:20 PM in 380-381T

Instructor: **Alex Sherman** (he/him/his)
PhD Candidate in English
ajsherm@stanford.edu
Office hours: 1-2 in the Literary Lab 460-401 (across from the elevators)

Course description

He merely closed his eyes when she reached out to touch his forehead



Computational literary studies is a decades-old field within the digital humanities (DH) that uses quantitative methods to answer perennial literary questions about texts' meaning, significance, history, form, politics, and so on. This course introduces students to computational studies by focusing on a foundational method: counting words. How can counting words, deliberately limiting our interpretations to the quantitative, help us frame and answer interesting questions about texts—whether a single text, a few, or thousands?

We will investigate these questions by studying Jane Austen's fiction alongside thousands of Austen fanfiction stories using RStudio and the R programming langauge. As such, the course also serves as a brief introduction to Austen and fanfiction scholarship and to programming in R. No prior Austen or R experience is required.

This course is the first in a sequence with a course taught by Professor Mark Algee-Hewitt in winter of 2023-4 which teaches more advanced NLP methods.

Learning outcomes

By the end of the quarter, you will have learned about:
-Translating among literary critical research questions, quantitative operationalizing, and computational findings
-The theory and motivations behind computational literary studies
-Basic syntax in the R programming language
-Using RStudio to load texts, count words, and make visualizations
-The basics of Jane Austen's oeuvre and the Austen fanfic community
-How to collaborate on computational literary research projects

Image credits: Ameriel; Cassandra Austen; Paige McKenzie

Tech requirements

You will need a computer running Windows, MacOS, or Linux. iPadOS will not work; ChromeOS will require installing Linux. That computer should have Microsoft Excel installed, which you should be able to access through your Stanford Office 365 account. Google Sheets is not Excel. Prior to the second day of class, you will follow a guide to install RStudio on your computer. Finally, you will find posted course materials and upload your work via Canvas.

You will need to bring a computer to class. If you run a desktop or iPad, please explore checking out a laptop from Lathrop for our class sessions.

Assignments

Reading Austen and Austen fanfic (strongly suggested)
Purpose: To inform class discussion, problem set extensions, and the final project.
Description: While not required, reading one or two Austen novels—especially *Pride and Prejudice*, the most popular source for fanfic—along with lots of fanfic stories will greatly help you in other aspects of the class. It is very hard to know what kinds of research questions and findings are interesting without some reading of a corpus.
Grading: NA

Weekly discussion assignments
Purpose: To help prepare us for weekly discussions.
Description: Before each week's discussion day, there is a short assignment detailed in the syllabus below. Due date: Upload to Canvas before each week's discussion (except week 1).
Grading: 10% of final grade. While I may comment on them afterwards, they are graded entirely on completion. Don't stress about these! Use them as your space and time to think for yourself.

Weekly problem sets
Purpose: To help you both practice coding in R and using computation to address literary topics we care about. These are the bedrock of the course.
Description: After each week's lab, I will assign a problem set that includes rote practice; a specific programming task in R; a general computational question; and a one paragraph reflection on how that question points to some literary critical argument, further question, topic, etc.
Due date: I informally encourage you to upload your work to Canvas before each week's discussion, so you can read over the lab handbook in advance, but they are formally due before each week's lab (except weeks 1 and 10).
Grading: 40%. See the longer rubric in each problem set.

Two problem set extensions
Purpose: Practice making longer literary arguments based on computational analyses to prepare for the final project.
Description: Extend a week's topic by investigating your own research question using the week's computational methods. Write 450-550 words explaining your findings, include a table or figure, and make some literary critical argument based on your results. See the assignment sheet for details.
Due date: I encourage you to do these soon after submitting the problem set, as the feedback will help you with the final project, but they are both formally due at the final exam period.

Grading: 15%. See the assignment sheet.

Presentation
Purpose: An initial goal to help start work on the final project and a chance to get feedback from classmates.
Description: As a group, present the corpus and research questions you're considering for the final project, along with some initial results or attempts.
Due date: Last day of class.
Grading: 5%. See the assignment sheet. Don't stress over these being polished—the point is to share work in progress so that you can get meaningful help from classmates.

Final project group submission
Purpose: A central part of the digital humanities is collaboration. Besides an opportunity to fully investigate a question you have about Austen and/or Austen fanfiction—synthesizing all we've learned in the course—the final project also helps practice project management, splitting computational work, and co-authoring writing.
Description: Frame a research question and corpus; explain the computational methods you used to study it; present your computational findings; and make a synthetic argument about the literary critical significance of your results. Write 1500-2000 words, including tables and figures. See the assignment sheet for details.
Due date: End of the final exam period for the course.
Grading: 15%. See the assignment sheet.

Final project individual reflection
Purpose: To let you reflect on the final project apart from the group submission.
Description: Explain your own thought process and work on the group project, and evaluate what you learned and what you might do differently in 500-750 words.
Due date: End of the final exam period for the course.
Grading: 5%. See the assignment sheet.

Weekly quizzes
Purpose: To hold you accountable for preparing before class and to help retain the terms and syntax of R.
Description: At the start of class, you will have five minutes to handwrite an answer to one question on the week's readings and one on the Lab.
Due date: Tuesdays of weeks 2-10, plus Thursday of week 1.
Grading: 10%. Two questions, each graded correct or incorrect.

Extra credit with extra problem set extensions
Purpose: To encourage you to further practice the skills for the problem set extensions.
Description: You can do up to 3 more problem set extensions. See above for more details.
Due date: End of the final exam period for the course.
Grading: Extra 2.5% per extra extension.

Logistics

Communication: Official course announcements and feedback on assignments will go out via Canvas and email. I will also set up a course Slack channel to enable students to help each other and so that

I can provide quick, informal help with assignments (e.g., "I've tried everything but this code isn't working.") While I may cross-post announcements to the Slack, please do not use the Slack for communicating course business: if you have questions about an absence, grading, extension, and so on, email me, don't use Slack.

Attendance: Since course sections will be highly interactive and include quizzes, it is important that you attend class. I also recognize that unexpected circumstances may arise. Please respect your classmates and instructors by staying home if you feel sick.

All students may receive two excused absences if you email me before class, without any penalty or make-up assignments. If you need more absences, please email me to arrange accommodations, which will usually involve additional writing assignments. If you are feeling well enough, we can also arrange for you to Zoom into class given 24 hours advance notice. Beyond these exceptions, absences will negatively impact your participation grade.

Submitting assignments, late submissions, and extensions: All assignments must be submitted via Canvas. Some assignment guides may also direct you to post them as a Discussion on Canvas so that the whole class can discuss them.

Problem sets submitted less than 24 hours late lose half their points; after that, I post answers, and they lose all credit. All other late assignments lose a half letter grade for each day it is overdue (e.g., an A paper turned in two days late becomes a B paper). If you have major extenuating circumstances, please email me and we can figure something out.

Office hours: I encourage you to come to office hours often, especially at the beginning of the quarter to introduce ourselves. I have set office hours but am very happy to make appointments.

Schedule

UNIT 1	Why and How to Count
Week 1	Intro to DH and to R
	Discussion: DH and learning what we don't already know; discussion and collaboration norms
	Lab: Data and Text
Week 2	Jane Austen, Janeites, and Metadata
	Discussion: History of Jane Austen and Austen Fanfiction
	Lab: Storing and manipulating metadata
Week 3	Operationalizing

Discussion and brainstorming: What would we want to measure about Jane Austen and fanfiction?

Lab: Functions

UNIT 2 Counting Words

Week 4 The Power of the Word

Discussion: What difference does a word make?

Lab: Regular expressions in R

Week 5 The Affordances of Mere Length

Discussion: How do you determine the length of a story, computationally or otherwise? What do and don't we count in measuring length? How does that length change other aspects of the story?

Lab: Counting Words (in Many Texts)

Week 6 Telltale Words

Discussion: What are some words that we strongly associate with certain Austen texts or with Austen texts in general? What about fanfiction? How can comparing distributions of words help us understand whole texts?

Lab: Document Term Matrices; Comparing Documents and Corpora

Week 7 Arranging and Ordering Words

Discussion: What are the problems with just ignoring word proximity and order, both locally and globally? What effects do Austen and fanfic writers achieve by putting words in certain places, and how can we measure those?

Lab: Dispersion, Windows, PMI with Windows

UNIT 3 Collaboration

Week 8 Showing Your Work

Discussion: How can we present our arguments in forms that are vivid, comprehensible, and thought-provoking? What are the risks of neat visual reductions of data?

Lab: Base Plot and ggplot

Week 9 Defining a Project

Discussion: A distinctive thing about humanities research is the flexible, back and forth between arguments and the "evidence" considered, i.e., the texts and corpora we choose to analyze. What kinds of corpora make sense for different research questions, and vice versa? What are good practices for working together on DH projects?

Structured Group Work Time in Class

Week 10 Doing a Project

Unstructured Group Work Time in Class

Project Presentations

Access and Accommodations

Stanford is committed to providing equal educational opportunities for disabled students. Disabled students are a valued and essential part of the Stanford community. We welcome you to our class.

If you experience disability, please register with the Office of Accessible Education (OAE). Professional staff will evaluate your needs, support appropriate and reasonable accommodations, and prepare an Academic Accommodation Letter for faculty. To get started, or to re-initiate services, please visit **oae.stanford.edu**.

If you already have an Academic Accommodation Letter, I invite you to share your letter with me. Academic Accommodation Letters should be shared at the earliest possible opportunity so I may partner with you and OAE to identify any barriers to access and inclusion that might be encountered in your experience of this course.

Honor code

The Stanford University Honor Code is a part of this course.

It is Stanford's statement on academic integrity first written by Stanford students in 1921. It articulates university expectations of students and faculty in establishing and maintaining the highest standards in academic work. It is agreed to by every student who enrolls and by every instructor who accepts appointment at Stanford.

The Honor Code states:

1) The Honor Code is an undertaking of the students, individually and collectively

- that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

- that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

2) The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

3) While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Penalties for violation of the Honor Code can be serious (e.g., suspension, and even expulsion).