LITERARY DATA: SOME APPROACHES

http://www.rci.rutgers.edu/~ag978/litdata

Thursdays, 1:10–4:10 p.m. in Murray 305 (Plangere Seminar Room)
Professor Andrew Goldstone (andrew.goldstone@rutgers.edu)
Office hours: Mondays, 2:00–4:00 p.m. in Murray 019

COURSE DESCRIPTION

In the last ten years, the strange quasi-disciplinary formation known as DH or Digital Humanities has renewed the struggle over methods in literary studies. Analyses of digitized texts using computer-assisted techniques promise to transform the kinds of evidence, the methods of interpretation, and the modes of argument which matter to literary scholarship. Data is now a subject of energetic debate in literary studies: what constitutes literary data, and how should it be analyzed and interpreted? How might aggregation and quantification produce new knowledge in literary scholarship? What methods are most appropriate for grappling with the enormous, and enormously messy, world of digitized literary texts and data about literature?

This course pursues two aims in parallel: to engage with the history and current practice of literary data analysis, and to introduce the foundational skills of literary data analysis in the R programming language. Class time will be divided between seminar and practical instruction. The seminar discussions trace theoretical debates about literary data from structuralism and scientific bibliography, to experiments in computational stylistics, to contemporary scholarly controversies in and around DH. The practicum surveys the fundamentals of programming and data manipulation, with an introduction to selected numerical techniques and data visualizations. Short homework exercises supplement the in-class instruction, with an emphasis on handling actual literary data of various kinds.

There are two major assignments. A short position paper on a theoretical question about literary data and DH is due at midterm. The final assignment is to plan, carry out, and report on a small-scale project in literary data analysis. This project is to be undertaken in small groups; the report will detail methods and interpretations together with code and data.

No special technical expertise of any kind is expected; instruction begins from first principles. However, the work of programming does require willingness to experiment, patience in the face of frustration, and the nerve to ask for help as often as needed.

Bring your own laptop to class, if you have one; loaner laptops will also be available for in-class workshops. MacOS X and Linux are the preferred operating systems for work in the course, but Windows will be accommodated as well.
LEARNING GOALS

1. Engage critically, through discussion and formal writing, with contemporary debates about computational methods in literary study.

2. Understand contemporary discussions of the digital humanities in the context of arguments about data in the human sciences in the last fifty years.

3. Develop basic competence in analyzing data using the R environment, including obtaining, reformatting, tabulating, and visualizing multiple forms of data.

4. Understand the fundamentals of computation by mastering the most important constructs in the R programming language.

5. Gain practical experience in analyzing literary data for the purpose of answering research questions in literary scholarship through a collaborative data-analysis research project.

This course fulfills the department’s B distribution requirement.

REQUIREMENTS AND GRADING

10% SEMINAR PARTICIPATION IN CLASS

Seminar discussion: your thoughtful participation is required. I do not expect everyone’s manner of thoughtful participation to be identical.

Practicum: I expect your best effort, not instant mastery. I also expect that you will seek my help or the help of classmates when you are confused or stuck.

10% PRACTICUM HOMEWORK

Programming homework sets are due each week at 6 a.m. on Thursday.

30% SHORT PAPER

A conference-length paper (7–8 pp.) on some of the theoretical questions raised in the first half of the course, engaging with several scholarly essays (at least one of which must be from the syllabus).

50% FINAL PROJECT

A collaborative project making use of the techniques discussed in the course. The ideal project identifies a research question, obtains a relevant dataset, analyzes the data, and develops a written argument on the basis of the analysis. Given the new skills the course teaches and the difficulties of data-gathering, however, the project may also consist of a proposal for such a research project.
together with a pilot study representing the progress the group was able to make. Working computer code, data, and a report are submitted together. I will work closely with each group.

**STUDENTS WITH DISABILITIES**

All reasonable accommodation will be given to students with disabilities. Students who may require accommodation should speak with me at the start of the semester. You may also contact the Office of Disability Services (disabilityservices.rutgers.edu; 848-445-6800).

**SCHEDULE**

In addition to the assigned readings from Jockers, *Text Analysis*, this schedule also notes suggested sections from Teetor, *R Cookbook*, which is useful for review. Two additional, optional texts with relevant reference material on the R language are Spector, *Data Manipulation with R* and Navarro, *Learning Statistics with R*. Finally, two more advanced texts on R available online are Wickham, *Advanced R* and Chambers, *Software for Data Analysis*; by the end of the course, the former will be approachable (the latter is exhaustive).

Many readings are available online; in digital versions of this syllabus, click a title to go to its bibliography entry, where, in most cases, you’ll find URLs.

Homework is listed under the date it is due.

**JANUARY 22. INTRODUCTION. WHAT IS LITERARY DATA?**

Rosenberg, “Data before the Fact.”
Optional: “What Is Data in Literary Studies?”

*Practicum.* Introduction to programming: transforming inputs to outputs; data and data types; expressions, statements, assignment. Documenting code and literate programming. Some hands-on time for R and RStudio setup.

**JANUARY 29. THE DATA OF THE HUMAN SCIENCES.**

Lévi-Strauss, “The Structural Study of Myth.”


*Practicum.* Boolean logic and logical indexing. String data and first text functions.
FEBRUARY 5. STYLE AND STYLISTICS.

Moretti, “Style, Inc.”
Trumpener, “Paratext and Genre System.”


*Practicum.* Conditionals and loops.

FEBRUARY 12. MORE COUNTING. NO SEMINAR MEETING.

Hourlong group tutorial meetings are on Monday the 9th at noon, Monday the 16th, at 1 p.m., and Wednesday the 18th at 10 a.m. in Murray 302.


FEBRUARY 19. SOCIOLOGY OF FORMS.

Moretti, *Graphs, Maps, Trees.*
Jockers, “Metadata.”
Hoover, “Quantitative Analysis and Literary Studies.”
Griswold, “Number Magic in Nigeria.”
Michel et al., “Quantitative Analysis of Culture Using Millions of Digitized Books.”


*Practicum.* Data structures wrap-up. Introduction to regular expressions.

FEBRUARY 26. THE FORMATION OF DH.

Hockey, “The History of Humanities Computing.”
McCarty, “Knowing...: Modeling in Literary Studies.”
McPherson, “Why Are the Digital Humanities So White?”
Kirschenbaum, “What Is 'Digital Humanities,' and Why Are They Saying Such Terrible Things about It?”


MARCH 5. CLASS CANCELED.

Missed seminar guest: Natalia Cecire (University of Sussex).

Homework. Jockers, Text Analysis, chap 8 (KWIC, function definition). Function practice: abstracting what we’ve done so far. (Teetor, R Cookbook, 2.12.)

MARCH 12. READING LITERARY DATA.

Ramsay, Reading Machines, chap. 1 (and optionally chap. 5).
Underwood, “Theorizing Research Practices We Forgot to Theorize Twenty Years Ago.”
Cecire, “Ways of Not Reading Gertrude Stein.”
(Optionally: Clement, “A thing not beginning and not ending.”)

Homework. Work through vignette("introduction","dplyr"). (Related, but not about dplyr: Teetor, R Cookbook, 5.27–33 and all of chap. 6. Expect confusion the first time.)


(MARCH 19. SPRING RECESS.)

(MARCH 25.) SHORT PAPER DUE.

MARCH 26. DESIGN, VISUAL METHOD, GRAMMATICAL VISUALIZATION.

Seminar guest: Francesca Giannetti, Digital Humanities Librarian.

Klein, “The Image of Absence.”
Lunenfeld et al., Digital Humanities, 40–45.
Manovich, “What Is Visualisation?”
Healy and Moody, “Data Visualization in Sociology” (optional).

Homework (due March 30). Reshaping and aggregating practice; visualization introduced.. Optionally: read ahead in the sections of Wickham, ggplot2, assigned for next week.

Practicum. Visualization; descriptive statistics; their relationship.

APRIL 2. MARKUP AND HYPERTEXT.

McGann, Radiant Textuality, 88–97, 106–35 (optional), 137–53.
Folsom, “Database as Genre.”
Download the sample TEI files and follow the browsing guide.
Homework (due April 6). Plots, rebooted. Wickham, ggplot2, chaps. 2–3 (qplot and the ggplot grammar). Optional: Jockers, Text Analysis, chaps. 6–7 (lexical variety; hapax richness).

Practicum. HTML and XML. XML in R.


Blei, “Probabilistic Topic Models.”
Grimmer and Stewart, “Text as Data.”
Goldstone and Underwood, “The Quiet Transformations of Literary Studies.”
Goldstone and Underwood, “Quiet Transformations.”

Homework (due April 11). Jockers, Text Analysis, chap. 10 (XML). Additionally: more XML practice; HTML and web scraping.

Practicum. What in the name of Michel Foucault is topic modeling?


Schmidt, “Words Alone.”
DiMaggio, Nag, and Blei, “Exploiting Affinities.”
Krippendorff, Content Analysis, 24–35.

Homework (due April 16). Exploring a topic model. Additionally, read Jockers, Text Analysis, 13.1–8 (generating a model), but skip the exercises.

Practicum. Even more modeling possibilities (LSA to LDA and what lies beyond).

April 23. Network Analysis and Tie Meaning.

DeWitt, “Advances in the Visualization of Data.”
Finn, Becoming Yourself.
Optional: Houston, “Toward a Computational Analysis of Victorian Poetics.”
Optional: Elson, Dames, and McKeown, “Extracting Social Networks from Literary Fiction.”

Practicum. Basic network visualization in R.

April 30. Literary Geography.

Bourdieu, Rules of Art (excerpt).

Homework. Long project abstract due.
Practicum. Informal project presentations and workshop.

MAY 25. FINAL PROJECTS DUE.

READINGS


Jockers, Matthew L. *Text Analysis with R for Students of Literature*. New York: Springer, 2014. ISBN: 9783319031637. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1007/978-3-319-03164-4. This textbook is available to you in PDF through the library. You are likely to find it easier to work from in paper. Springer sells both an ordinary hardcover and a print-on-demand softcover.


Navarro, Dan. Learning Statistics with R. http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/. Not required. This statistics textbook, free online, includes a good introductory discussion of R.


ACKNOWLEDGMENTS

I have drawn on contributions by the following people, either in discussion with me or through publicly-available course syllabuses or both: David Bamman, Natalia Cecire, Anne DeWitt, Lauren Klein, John Kucich, Meredith McGill, Benjamin Schmidt, Scott Selisker, Cosma Shalizi, Ted Underwood, and Matthew Wilkens.

This syllabus is available for duplication or modification for other courses and non-commercial uses under a Creative Commons Attribution-NonCommercial 4.0 International License. Acknowledgment with attribution is requested.