Distant-Reading the Long Nineteenth Century

Ted Underwood
Tuesday 1:00 – 2:50 pm
English Building 113

Spring 2015
ENGL 533
tunder@illinois.edu

It's odd to start one syllabus by quoting another, but I don't think I can improve on this, from Matthew Wilkens' 2012 course on "Digital Humanities":

> Contemporary criticism has a problem. We long ago gave up the idea that our task was to appreciate and explain a handful of great texts, replacing that goal with a much more important and ambitious one: to understand cultural production as a whole by way of the aesthetic objects it creates. But we have continued to practice our craft as if the methods developed in pursuit of the old project were the only ones suited to the new task.

I think that explains why Franco Moretti's phrase "distant reading" has gained traction, slowly, over the last fifteen years. Close reading is a great thing. But if we were building the discipline of literary study from scratch today, with our current goals in mind, we'd probably also teach students how to find patterns in large digital libraries. Those libraries are creating opportunities for literary-historical research that would be obvious if we hadn't already specialized in a different scale of analysis.

For instance, just to consider a few questions scholars have started to explore in relation to nineteenth-century fiction: why are novels set in some places more often than others? Does the rise of consumer capitalism change the narrative role of money? How do representations of gender change from one period, or genre, or audience, to another? We're not going to answer all those questions in this course, but you will come away understanding how to make headway on questions like those, where a persuasive answer may require thousands of volumes as evidence. I hope you'll also be empowered to pose similar questions of your own, in the nineteenth century or elsewhere.

How much technical knowledge do literary scholars need in order to undertake this kind of research? We don't need to abandon our existing strengths to become super-proficient programmers. But we do need to be confident that no aspects of our sources are hidden from us. We need to be able to go behind a glossy public website and get our hands on the underlying stuff (on digital texts, or bibliographic information, or whatever else we're discussing).

In this course, we'll use a language called R to "get our hands on the underlying stuff" — aka, manipulate data. While R is a programming language, you don't have to master programming in order to use it; it allows you to take steps one at a time. We're going to proceed slowly, but I think by the end of the course you'll be able to do meaningful research with R on a large sample of texts between 1750 and 1922.

The course will assume no previous knowledge of computers or statistics. If you've got a laptop, bring it to class; if not, we'll work something out.

Texts

Matthew Jockers, *Text Analysis with R for Students of Literature.*

Paul Teetor, *R Cookbook.*

Most other readings will be available on the web, or through the course website. We also have access to about 160,000 volumes of literary writing between 1700 and 1922, in HathiTrust Digital Library. (In practice, we'll work with smaller subsets of that.)

Requirements

Most graduate seminars in the humanities involve a lot of reading and then a long paper at the end. In this course, your work will be spread more evenly across the semester. There is a final project, but it's not intended to have the same weight as a seminar paper.

Homework (40% of grade)

There will be ten data analysis assignments. They're due at 9 a.m. Tuesday; you can e-mail me your code. Collaboration on homework is explicitly encouraged. If you're *all* stumped, you can ask me for help.

Proposal for individual project (10% of grade)

This is a three-page paper proposing a literary-historical experiment.

Final individual project (40% of grade)

The experiment itself. You'll identify a research question, define a plan of analysis and a relevant dataset; obtain the data (with help from me); write code to analyze it, and finally interpret your results. But let me underline the word "experiment." Ambiguous or inconclusive results are absolutely fine; think of this as preliminary exploration, not as a "paper." We'll present short versions on the last day of class; final versions should run 10-12 pp. plus code.

Seminar participation (10% of grade).

My sources

In designing this course I consulted syllabi by Rachel Buurma, James A. Evans, Andrew Goldstone, Lauren Klein, Alan Liu, Andrew Piper, Benjamin Schmidt, and Matthew Wilkens. The influence of Goldstone's syllabus for "Literary Data" (Spring 2015) was particularly pervasive: I borrowed a lot from it.

Tuesday Jan 20	**How well do we understand literary history?** Moretti, "The Slaughterhouse of Literature." Bode, "Literary Studies in the Digital Age" from *Reading by Numbers* (2014). *In class*: Install Rstudio. What's a "vector"?
Jan 27	**Getting at the underlying stuff.** *Homework*: Chapter 1 of Jockers, *Text Analysis with R*, and the first section (2.1) of Chapter 2. Teetor, 2.1-2, 2.5-10, 2.14. Also, a practice problem set (indexing vectors, assignment, expressions). *In class*: We'll work through the practice problem set, and start to grapple with *Moby Dick*.
Feb 3	**What (if anything) is "style"?** Burrows, *Computation into Criticism* (excerpt). Allison, et. al., "Quantitative Formalism," Stanford Lit Lab Pamphlet 1. *Homework:* Jockers, Chapter 2. Teetor, 5.1-4. 7.1-4. First Actual Problem Set (most frequent words). *In class:* Named vectors, relative frequencies. Conditionals and loops.
Feb 10	**How can we meaningfully compare texts?** Schmidt, "Comparing Corpuses by Word Use." Mueller (I suspect), "Comparing Word Form Counts." Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence" (*skim*, be amazed that we understand parts of it). *Homework:* Teetor, Chapter 6 and Chapter 9, "Introduction." Jockers, Chapter 4. Second Problem Set. *In class:* Functions. We'll use Dunning's log-likelihood to identify words that characterize Austen or Melville.

Feb 17	**Using correlations to explore plot.** Schmidt, "Typical TV episodes" and "Fundamental plot arcs." Underwood, "Plot arcs in the novel." *Homework:* Jockers, Chapter 5. Teetor, "Introduction" to Chapter 5. Third Problem Set (Correlations). *In class:* We'll identify seductive correlations in the plots of 19c novels and then ask ourselves, are they meaningful? "Multiple comparisons."
Feb 24	**Wait … this is formalist. What about history?** Liu, "Where is Cultural Criticism in the Digital Humanities?" Klein, "The Image of Absence." *Homework:* Jockers, Chapters 7 and 8. Fourth Problem Set (build a KWIC index that finds references to slavery). *In class:* Discuss theoretical and technical problems with our index.
March 3	**What about History, Part Two: Second Time as Farce.** Underwood, Long, and So, "Cents and Sensibility." Wilkens, "Geographic Imagination of Civil War-Era American Fiction." Krippendorff, excerpts from *Content Analysis*. *Homework:* Jockers, Chapter 9. Fifth Problem Set (build a KWIC index that finds references to whatever you want. You are now dangerous.) *In class:* Introduce dplyr.
Mar 10	**Wait … how is this not sociology?** English, "Everywhere and Nowhere: The Sociology of Literature." Radway, from *Reading the Romance*. "Television Violence: A Coding Scheme," Mustonen and Pulkinnen. *Homework:* Propose a scheme for content analysis on 19c novels. Think small. Like, who kisses who in novels? When do novels talk about blood? In R, type: `vignette("introduction","dplyr")`. *In class:* We compare schemes, and create a collective plan and data format.

Mar 17	**A medium-to-low-tech content analysis.** "The Gerbner Violence Profile: A Public Debate" from Krippendorff and Bock, *The Content Analysis Reader*. Teetor, 9.introduction – 9.9 Homework: Sixth Problem Set (carry out content analysis by the weekend, and share that data, so we can collectively merge our results). Read a few chapters of a novel of your choice for context. In class: We'll use dplyr to find patterns in our merged data.
Mar 19	***Three-page proposal for individual project due.***
Mar 31	**Clustering, exploratory analysis, unsupervised learning.** Teetor, 13.4, 13.6. Jockers, Chapter 11. *Homework:* Seventh Problem. Clustering poets. Also, choose two volumes of poetry we're using and read four poems from each. *In class:* The "curse of dimensionality." Install MALLET.
April 7	**What's a "model"? Supervised learning.** McCarty: "Knowing …: Modeling in Literary Studies." Leo Breiman: "Statistical Modeling: The Two Cultures." *Homework:* Eighth Problem. Authorship attribution experiment. *In class:* We model everything we can think of. Genre, audience, etc.
April 14	**Topic modeling.** Blei, "Probabilistic Topic Models." Underwood, "Topic Modeling Made Just Simple Enough" (don't trust everything in this; it was three years ago). DiMaggio, Nag, and Blei, "Exploiting Affinities." *Homework:* Ninth Problem. Topic-model long-nineteenth-century fiction, everyone using a different list of stopwords. *In class:* Attempt to interpret our topic models.

April 21	**More topic modeling.** Heuser and Le-Khac, "A Quantitative Literary History…The Semantic Cohort Method" (Stanford Lit Lab Pamphlet Series). Liu, "The Meaning of the Digital Humanities." Schmidt, "Words Alone: Dismantling Topic Models in the Humanities." *Homework:* Tenth Problem (create visualizations of change-across-time based on a topic model of poetry, or fiction). Identify a volume whose relationship to a particular topic in your model seems to pose an interesting question. *In class:* We'll discuss the interpretability of topic models, and choose a short literary text to be read in more depth for next time.
April 28	**Literary networks.** Elson, Dames, and McKeown, "Extracting Social Networks from Literary Fiction." Long and So, "Network Analysis and the Sociology of Modernism." *Homework:* Read the text we collectively identified last time; come to class prepared to talk about the relationship between close and distant readings.
May 5	Last day. Presentation of individual projects.
May 11	***Final projects due.***