The Stone and the Shell

Using large digital libraries to advance literary history

Seven ways humanists are using computers to understand text.

Posted on **June 4, 2015** by **tedunderwood**

[This is an updated version of a blog post I wrote three years ago, which organized introductory resources for a workshop. Getting ready for another workshop this summer, I glanced back at the old post and realized it's out of date, because we've collectively covered a lot of ground in three years. Here's an overhaul.]

Why are humanists using computers to understand text at all?
Part of the point of the phrase "digital humanities" is to claim information technology as something that belongs *in* the humanities — not an invader from some other field. And it's true, humanistic interpretation has always had a technological dimension: we organized writing with commonplace books and concordances before we took up keyword search [Nowviskie, 2004; Stallybrass, 2007].

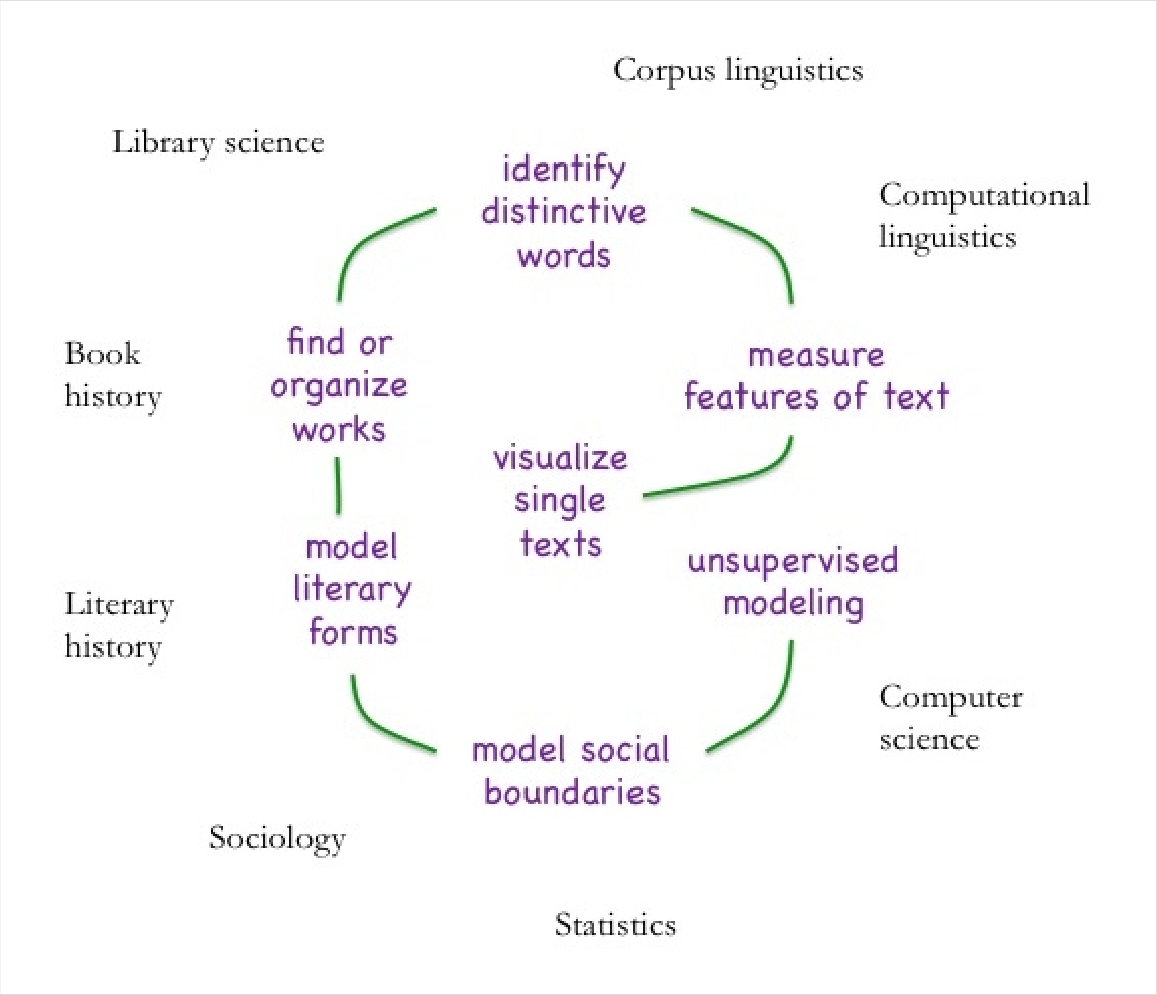
But framing new research opportunities as a specifically humanistic movement called "DH" has the downside of obscuring a bigger picture. Computational methods are transforming the social and natural sciences as much as the humanities, and they're doing so partly by creating new conversations between disciplines. One of the main ways computers are changing the textual humanities is by mediating new connections to social science. The statistical models that help sociologists understand social stratification and social change haven't in the past contributed much to the humanities, because it's been difficult to connect quantitative models to the richer, looser sort of evidence provided by written documents. But that barrier is dissolving. As new methods make it easier to represent unstructured text in a statistical model, a lot of fascinating questions are opening up for social scientists and humanists alike [O'Connor et. al. 2011].

In short, computational analysis of text is not a specific new technology or a subfield of digital humanities; it's a wide en several different disciplines. Humanists often nd digital tools that will automate familiar tasks. Tha ls you could use to create a concordance or a w re involved forms of text analysis do start to resemble er no obligation to dabble in social science.

But I should also warn you tl thing called "text analysis" or "distant reading" n about methods, and if you get drawn into the conve ry a lot of things that aren't packaged yet as tools

What can we actually do?

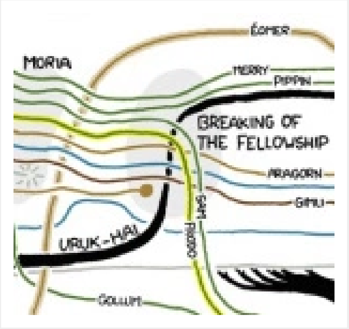
The image below is a map of a few things you might do with text (inspired by, though different from, Alan Liu's map of "digital humanities"). The idea is to give you a loose sense of how different activities are related to different disciplinary traditions. We'll start in the center, and spiral out; this is just a way to organize discussion, and isn't necessarily meant to suggest a sequential work flow.



1) Visualize single texts.

Text analysis is sometimes represented as part of a “new modesty” in the humanities [Williams]. Generally, that’s a bizarre notion. Most of the methods described in this post aim to reveal patterns hidden from individual readers — not a particularly modest project. But there are a few forms of analysis that might count as surface readings, because they visualize textual patterns that are open to direct inspection.

For instance, people love cartoons by Randall Munroe that visualize the plots of familiar movies by showing which characters are together at different points in the narrative.



— Detail from an xkcd cartoon.

These cartoons reveal little we didn’t know. They’re fun to explore in part because the narratives being represented *are* familiar: we get to rediscover familiar material in a graphical medium that makes it easy to zoom back and forth between macroscopic patterns and details. Network graphs that connect characters are fun to explore for a similar reason. It’s still a matter of debate what (if anything) they reveal; it’s important to keep in mind that fictional networks can behave very differently from real-world social networks [Elson, et al., 2010]. But people tend to find them interesting.

A concordance also, in a sense, tells us nothing we couldn’t learn by reading on our own. But critics nevertheless find them useful. If you want to make a concordance for a single work (or for that matter a whole library), AntConc is a good tool.

Visualization strategies themselves are a topic that could deserve a whole separate discussion.

2) Choose features to represent texts.
A scholar undertaking computational analysis of text needs to answer two questions. First, how are you going to represent texts? Second, what are you going to do with that representation once you’ve got it? Most what follows will focus on the second question, because there are a lot of equally good answers to the first one — and your answer to the first question doesn’t necessarily constrain what you do next.

In practice, texts are often represented simply by counting the various words they contain (they are treated as so-called "bags of words"). Because this representation of text is radically different from readers' sequential experience of language, people tend to be surprised that it works. But the goal of computational analysis is not, after all, to reproduce the modes of understanding readers have already achieved. If we're trying to reveal large-scale patterns that *wouldn't* be evident in ordinary reading, it may not actually be necessary to retrace the syntactic patterns that organize readers' understanding of specific passages. And it turns out that a lot of large-scale questions are registered at the level of word choice: authorship, theme, genre, intended audience, and so on. The popularity of Google's Ngram Viewer shows that people often find word frequencies interesting in their own right.

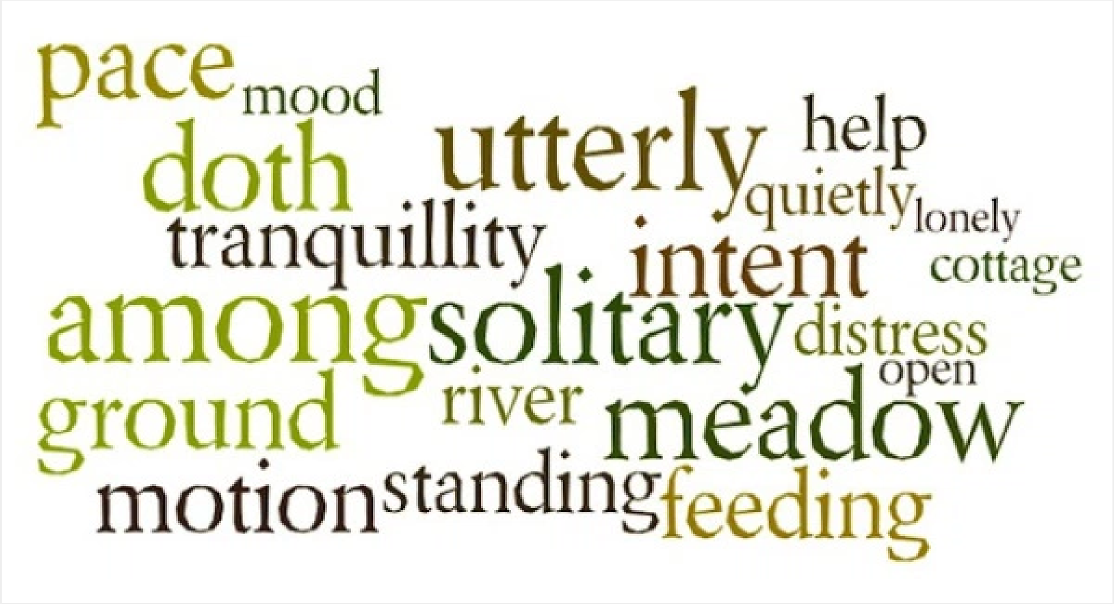
But there are lots of other ways to represent text. You can count two-word phrases, or measure white space if you like. Qualitative information that can't be counted can be represented as a "categorical variable." It's also possible to consider syntax, if you need to. Computational linguists are getting pretty good at parsing sentences; many of their insights have been packaged accessibly in projects like the Natural Language Toolkit. And there will certainly be research questions — involving, for instance, the concept of character — that require syntactic analysis. But they tend not to be questions that are appropriate for people just starting out.

3) Identify distinctive vocabulary.

It can be pretty easy, on the other hand, to produce useful insights on the level of diction. These are claims of a kind that literary scholars have long made: *The Norton Anthology of English Literature* proves that William Wordsworth emblematizes Romantic alienation, for instance, by saying that "the words 'solitary,' 'by one self,' 'alone' sound through his poems" [Greenblatt et. al., 16].

Of course, literary scholars have also learned to be wary of these claims. I guess Wordsworth does write "alone" a lot: but does he really do so more than other writers? "Alone" is a common word. How do we distinguish real insights about diction from specious cherry-picking?

Corpus linguists have developed a number of ways to identify locutions that are really overrepresented in one sample of writing relative to others. One of the most widely used is Dunning's log-likelihood: Ben Schmidt has explained why it works, and it's easily accessible online through Voyant or downloaded in the AntConc application already mentioned. So if you have a sample of one author's writing (say Wordsworth), and a reference corpus against which to contrast it (say, a collection of other poetry), it's really pretty straightforward to identify terms that typify Wordsworth relative to the other sample. (There are also other ways to measure overrepresentation; Adam Kilgarriff recommends a Mann-Whitney test.) And in fact there's pretty good evidence that "solitary" is among the words that distinguish Wordsworth from other poets.

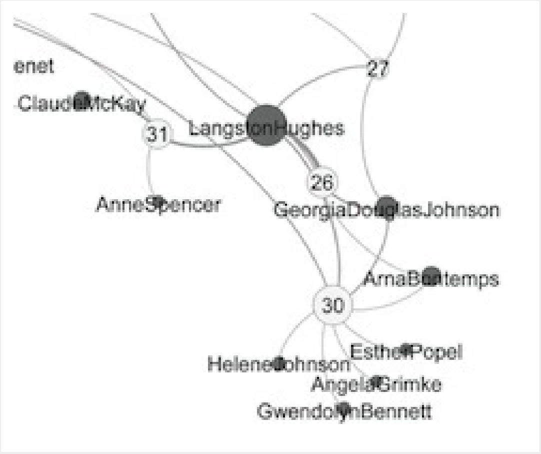


— Words that are consistently more common in works by William Wordsworth than in other poets from 1780 to 1850. I've used Wordle's graphics, but the words have been selected by a Mann-Whitney test, which measures overrepresentation relative to a context — not by Wordle's own (context-free) method.

It's also easy to turn results like this into a word cloud — if you want to. People make fun of word clouds, with some justice; they're eye-catching but don't give you a lot of information. I use them in blog posts, because eye-catching, but I wouldn't in an article.

4) Find or organize works.
This rubric is shorthand for the enormous number of different ways we might use information technology to organize collections of written material or orient ourselves in discursive space. Humanists already do this all the time, of course: we rely very heavily on web search, as well as keyword searching in library catalogs and full-text databases.

But our current array of strategies may not necessarily reveal all the things we want to find. This will be obvious to historians, who work extensively with unpublished material. But it's true even for printed books: works of poetry or fiction published before 1960, for instance, are often not tagged as "poetry" or "fiction."



— A detail from Fig 7 in So and Long, “Network Analysis and the Sociology of Modernism.”

Even if we believed that the task of simply finding things had been solved, we would still need ways to map or organize these collections. One interesting thread of research over the last few years has involved mapping the concrete social connections that organize literary production. Natalie Houston has mapped connections between Victorian poets and publishing houses; Hoyt Long and Richard Jean So have shown how writers are related by publication in the same journals [Houston 2014; So and Long 2013].

There are of course hundreds of other ways humanists might want to organize their material. Maps are often used to visualize references to places, or places of publication. Another obvious approach is to group works by some measure of textual similarity.

There aren’t purpose-built tools to support much of this work. There are tools for building visualizations, but often the larger part of the problem is finding, or constructing, the metadata you need.

5) Model literary forms or genres.
Throughout the rest of this post I’ll be talking about “modeling”; underselling the centrality of that concept seems to me the main oversight in the 2012 post I’m fixing.

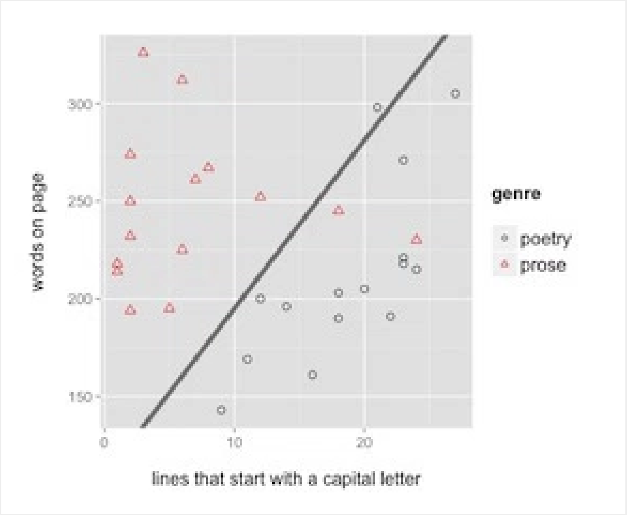
— A model treehouse, by Austin and Zak — CC-NC-SA.

A model is a simplified representation of something, and in principle models can be built out of words, balsa wood, or anything you like. In practice, in the social sciences, *statistical* models are often equations that describe the probability of an association between variables. Often the “response variable” is the thing you’re trying to understand (literary form, voting behavior, or what have you), and the “predictor variables” are things you suspect might help explain or predict it.

This isn’t the only way to approach text analysis; historically, humanists have tended to begin instead by first choosing some aspect of text to measure, and then launching an argument about the significance of the thing they measured. I’ve done that myself, and it can work. But social scientists prefer to tackle problems the other way around: first identify a concept that you’re trying to understand, and then try to model it. There’s something to be said for their bizarrely systematic approach.

Building a model can help humanists in a number of ways. Classically, social scientists model concepts in order to understand them better. If you’re trying to understand the difference between two genres or forms, building a model could help identify the features that distinguish them.

Scholars can also frame models of entirely new genres, as Andrew Piper does in a recent essay on the “conversional novel.”



— A very simple, imaginary statistical model that distinguishes pages of poetry from pages of prose.

In other cases, the point of modeling will not actually be to describe or *explain* the concept being modeled, but very simply to *recognize* it at scale. I found that I needed to build predictive models simply to find the fiction, poetry, and drama in a collection of 850,000 volumes.

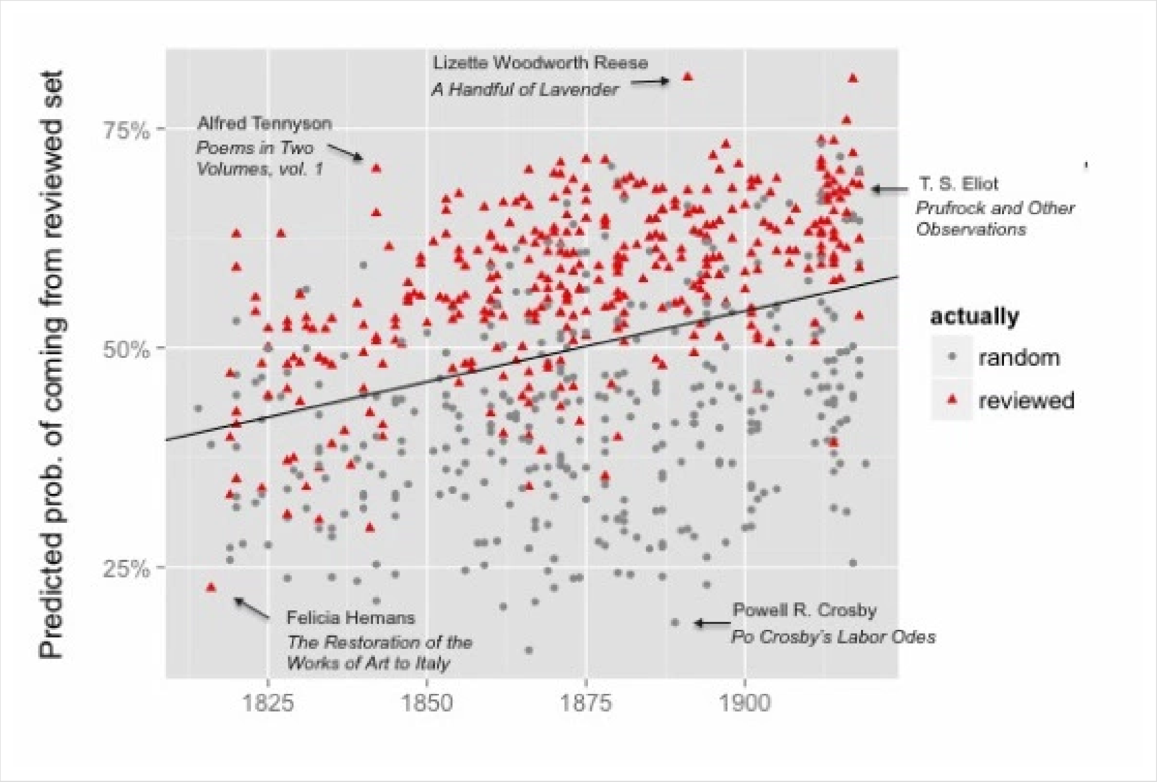
The tension between modeling-to-explain and modeling-to-predict has been discussed at length in other disciplines [Shmueli, 2010]. But statistical models haven't been used extensively in historical research yet, and humanists may well find ways to use them that aren't common in other disciplines. For instance, once we have a model of a phenomenon, we may want to ask questions about the diachronic stability of the pattern we're modeling. (Does a model trained to recognize this genre in one decade make equally good predictions about the next?)

There are lots of software packages that can help you infer models of your data. But assessing the validity and appropriateness of a model is a trickier business. It's important to fully understand the methods we're borrowing, and that's likely to require a bit of background reading. One might start by understanding the assumptions implicit in simple linear models, and work up to the more complex models produced by machine learning algorithms [Sculley and Pasanek 2008]. In particular, it's important to learn something about the problem of "overfitting." Part of the reason statistical models are becoming more useful in the humanities is that new methods make it possible to use hundreds or thousands of variables, which in turn makes it possible to represent unstructured text (those bags of words tend to contain a lot of variables). But large numbers of variables raise the risk of "overfitting" your data, and you'll need to know how to avoid that.

6) Model social boundaries.

There's no reason why statistical models of text need to be restricted to questions of genre and form. Texts are also involved in all kinds of social transactions, and those social contexts are often legible in the text itself.

For instance, Jordan Sellers and I have recently been studying the history of literary distinction by training models to distinguish poetry reviewed in elite periodicals from a random selection of volumes drawn from a digital library. There are a lot of things we might learn by doing this, but the top-line result is that the implicit standards distinguishing elite poetic discourse turn out to be relatively stable across a century.



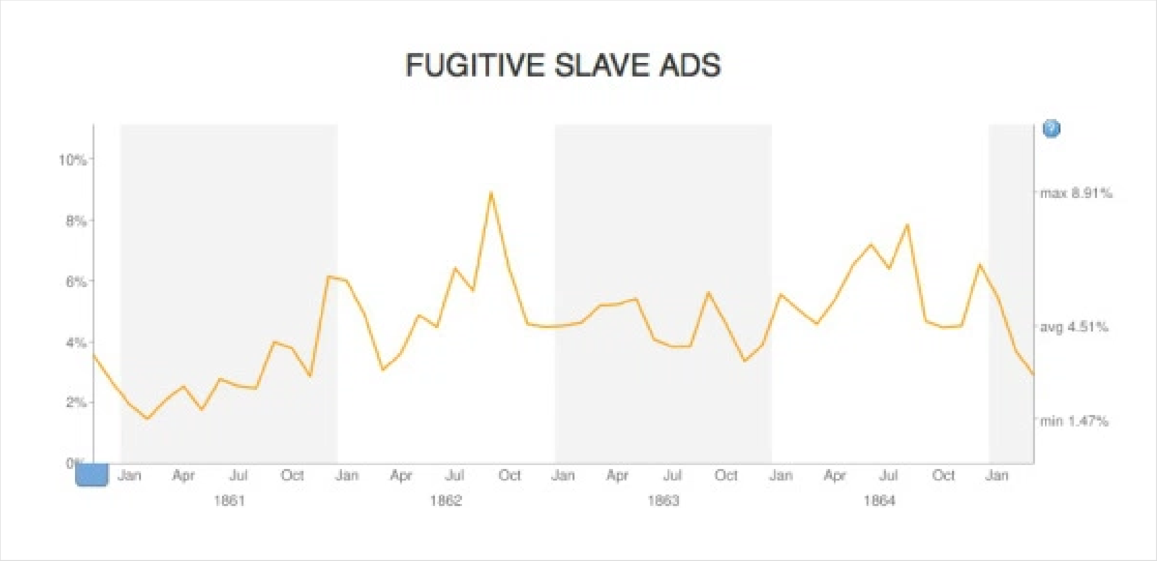
Similar questions could be framed about political or legal history.

7) Unsupervised modeling.

The models we've discussed so far are *supervised* in the sense that they have an explicit goal. You already know (say) which novels got reviewed in prominent periodicals, and which didn't; you're training a model in order to discover whether there are any patterns in the texts themselves that might help us explain this social boundary, or trace its history.

But advances in machine learning have also made it possible to train *unsupervised* models. Here you start with an unlabeled collection of texts; you ask a learning algorithm to organize the collection by finding clusters or patterns of some loosely specified kind. You don't necessarily know what patterns will emerge.

If this sounds epistemologically risky, you're not wrong. Since the hermeneutic circle doesn't allow us to get something for nothing, unsupervised modeling does inevitably involve a lot of (explicit) assumptions. It can nevertheless be extremely useful as an exploratory heuristic, and sometimes as a foundation for argument. A family of unsupervised algorithms called "topic modeling" have attracted a lot of attention in the last few years, from both social scientists and humanists. Robert K. Nelson has used topic modeling, for instance, to identify patterns of publication in a Civil-War-era newspaper from Richmond.



But I'm putting unsupervised models at the end of this list because they may almost be too seductive. Topic modeling is perfectly designed for workshops and demonstrations, since you don't have to start with a specific research question. A group of people with different interests can just pour a collection of texts into the computer, gather round, and see what patterns emerge. Generally, interesting patterns do emerge: topic modeling can be a powerful tool for discovery. But it would be a mistake to take this workflow as paradigmatic for text analysis. Usually researchers begin with specific research questions, and for that reason I suspect we're often going to prefer supervised models.

* * *

In short, there are a lot of new things humanists can do with text, ranging from new versions of things we've always done (make literary arguments about diction), to modeling experiments that take us fairly deep into the methodological terrain of the social sciences. Some of these projects can be crystallized in a push-button "tool," but some of the more ambitious projects require a little familiarity with a data-analysis environment like Rstudio, or even a programming language like Python, and more importantly with the assumptions underpinning quantitative social science. For that reason, I don't expect these methods to become universally diffused in the humanities any time soon. In principle, everything above is accessible for undergraduates, with a semester or two of preparation — but it's not preparation of a kind that English or History majors are guaranteed to have.

Generally I leave blog posts undisturbed after posting them, to document what happened when. But things are changing rapidly, and it's a lot of work to completely overhaul a survey post like this every few years, so in this one case I may keep tinkering and adding stuff as time passes. I'll flag my edits with a date in square brackets.

* * *

SELECTED BIBLIOGRAPHY

Elson, D. K., N. Dames, and K. R. McKeown. "Extracting Social Networks from Literary Fiction." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010. 138-147.

Greenblatt, Stephen, et. al., *Norton Anthology of English Literature* 8th Edition, vol 2 (New York: WW Norton, 2006.

Houston, Natalie. "Towards a Computational Analysis of Victorian Poetics." *Victorian Studies* 56.3 (Spring 2014): 498-510.

Nowviskie, Bethany. "Speculative Computing: Instruments for Interpretive Scholarship." Ph.D dissertation, University of Virginia, 2004.

O'Connor, Brendan, David Bamman, and Noah Smith, "Computational Text Analysis for Social Science: Model Assumptions and Complexity," NIPS Workshop on Computational Social Science, December 2011.

Piper, Andrew. "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel." *New Literary History* 46.1 (2015).

Sculley, D., and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23.4 (2008): 409-24.

Shmueli, Galit. "To Explain or to Predict?" *Statistical Science* 25.3 (2010).

So, Richard Jean, and Hoyt Long, "Network Analysis and the Sociology of Modernism," *boundary 2* 40.2 (2013).

Stallybrass, Peter. "Against Thinking." *PMLA* 122.5 (2007): 1580-1587.

Williams, Jeffrey. "The New Modesty in Literary Criticism." *Chronicle of Higher Education* January 5, 2015.