CA Journal of Cultural Analytics

September 17, 2019

Is there a text in my data? (Part 1): on counting words

Michael Gavin[a]

[a]*University of South Carolina*

ARTICLE INFO

Article DOI: 10.22148/001c.11830

Journal ISSN: 2371-4549

ABSTRACT

This essay is the first in a two-part series. This first installment invites readers to consider a few very basic questions: what does it mean to count words in a text? What happens to the text, and to our understanding of it, when we decompose it into a series of word counts? What relation exists between the textual domain and its numerical image? Or, to restate this question with a nod to literary critic stanley fish, "is there a text in my data?" following one document through a series of typical transformations -- first into a simple list of words and their frequencies, then to a vector of elements in a matrix, and from there through the processes of normalization, dimensionality reduction, and analysis -- this essay argues against the commonly held notion that counting words reduces complexity, suggesting instead that semantic models embed textual objects in highly complex structures that are extremely sensitive to historical context and subtle nuances in meaning. Word frequencies aren't static, given things that simply exist in a text. They're produced through the act of modeling, and the mathematical structures they imply dissolve both words and texts into elaborate systems of mutual interrelation.

Several months have now passed since *Critical Inquiry* published Nan Z. Da's "A Computational Case Against Computational Literary Studies" and hosted an online forum to debate the issues Da raised.[1] In the piece, Da argued that literary criticism and computational research are fundamentally incompatible. Among studies that use computers to interpret literature, "what is robust is obvious," she argued, "and what is not obvious is not robust."[2] Computational research in the field lacks statistical rigor, she asserted, and therefore, she advised, journal editors should consult statisticians during peer review and seek their approval before publication.[3] The evidence Da marshalled came primarily in two kinds. In some cases, she showed that tweaking the analytical procedures in some small, seemingly arbitrary way could change the resulting analyses. Adding stopwords back into the system skewed the results of Hoyt Long's and Richard Jean So's study of modern fiction; stemming the Latin altered the distribution of book chapters in Andrew Piper's graph of Augustine's *Confessions*.[4] Other times, she advanced alternative, deflationary

readings of other critics' results. For example, she argued that Matthew Jockers's network graph of nineteenth-century novels just reflects what we already know about language and doesn't tell us anything interesting about their history as novels, per se.[5] Scholars whose work she critiqued responded with predictable defensiveness, most pointing out places where, they believed, Da had misrepresented some aspect of their work or where she had made some other kind of mistake.[6] In the following days of public debate, Da admitted several errors and disclosed a surprising fact about her article's peer-review process: the "out-of-field" scholar who reviewed her essay said she failed to demonstrate her core argument. "After assessing the validity of my empirical claims," Da conceded, "my out-of-field peer reviewer did not finally agree with me that computational methods works [*sic*] poorly on literary objects."[7]

Whether those admissions are totally discrediting or totally beside the point depends, as such things always do, on one's perspective. Scholars whose work had been impugned tended to react as if the most important thing was whether or not the criticisms leveled at themselves were justified. Several tried to redirect the conversation toward methodological questions regarding the role of statistics in humanistic inquiry, but the sheer drama of mutual accusation swamped such concerns. Few readers, I think, shared their sense of outrage over Da's improper definition of *p*-values.[8] For other commenters, the real issues were political. Is the most important thing that neoliberal administrators are destroying the humanities? If so, digital humanists ought to be scorned as opportunistic confidence men, at best, and as collaborating saboteurs, at worst.[9] Is the real problem instead that the digital humanities replicate old-school academic politics, elevating the work of white guys at the expense of scholars who aren't male and pale? If so, the entire debate could only provoke frustration while motivating a push to acknowledge more socially conscious work by digital humanists who are women and people of color.[10] Wrapping up the entire debacle, Stanley Fish admitted that he couldn't follow Da's reasoning but was sure she was right, and, caring nothing for the political concerns that motivated other commenters, used the opportunity to repeat his decades-old gate-keeping admonition that data analysis has no place in literary criticism.[11] The whole affair was 1) a mess and 2) a reminder that people can have incompatible opinions about the grounds of valid knowledge despite working in a common academic field.

When I first imagined what I might say in response to the controversy, I imagined writing a piece that would castigate Da's partisans for their lack of curiosity. We are, after all, not just professors of literature but also professors of language, and as such we have a professional responsibility to be curious about how language works and a shared obligation to learn how scholars from other disciplines study the topic. In the last forty years, major advancements have been made in the fields of library and information science, natural language processing, and corpus linguistics. On this score, I'd have faulted Da for dismissing research on the topic of "information retrieval" — that's a line of inquiry that studies how search engines work.[12] Think about this for a moment: How could the people at Google invent software that searches through billions of documents and returns just the right ones, unless they had an extremely sophisticated and robust *theory* to explain how meaning is distributed through language? I believe we should be curious about that theory and learn something about it before dismissing it as worthless. Even if Fish were right, and criticism written under this theory does nothing for us as critics, isn't the theory, in and of itself, profoundly interesting? Any English professor who replies "No!" or "Sure, but . . ." is simply being obtuse, whatever their political or disciplinary rationalizations.

Or so my imaginary table-pounding essay would have concluded. However, I don't want to focus on this critique because, though true enough on its own terms, it's quite unfair. We partisans of cultural analytics can't really fault our colleagues. As I've written elsewhere, we have an unfortunate habit of skipping to the end — of showing off our statistical analyses of corpora without pausing to reflect on the extraordinary gap that separates the theory of meaning that informs computational semantics from the subjectively felt experience of reading and knowing about a text. If I read a paragraph and come to a conclusion about that paragraph's meaning, no statistical analysis is going to change my mind. Full stop. Ain't gonna happen. When literary historians generalize across lots of books, their generalizations are grounded in this fundamental activity of reading comprehension, an activity that has been rigorously theorized going back to Aristotle. When we flip the script and start tinkering with statistical models, we're effectively throwing that entire activity and all the critical history that supports it out the window. And for what? For counting words. And what does counting words get us? A few charts we can comprehend only if we've

already read enough of the books to infer where the statistical signals are coming from.

We can cut our colleagues some slack for suspecting we're all a bunch of crackpots.

And so it is with some trepidation that I cast a few ideas into the teapot where this tempest has so vigorously blown. I should acknowledge at the outset that I don't identify as a literary scholar. I'm an intellectual historian and a book historian by training, so I have little personally at stake in debates about interpretive validity in criticism.[13] For me, The Most Important Thing is a rather small, idiosyncratic, perhaps even pedantic thing, but it's a belief that motivates all my work in the field. My belief is this: Counting words is *interesting*. My goal is to share with you why I find it so. Rather than ask whether computational methods can be good for the study of literature, I want to back up, take a few deep breaths, and just think a bit about the very nature of computational textual analysis.

This essay is the first in a two-part series. In this first installment, "On Counting Words," I'll respond to a single comment made almost in passing in Nan Z. Da's original piece. She says that "all the things that appear in [computational literary studies]—network analysis, digital mapping, linear and nonlinear regressions, topic modeling, topology, entropy—are just fancier ways of talking about word frequency changes."[14] This comment is wrong in little ways that won't matter to most literature scholars. Much of what happens in network analysis and digital mapping, in particular, has nothing to do with word counts. But Da is not completely off the mark. To be honest, I spend an embarrassingly large amount of my time trying to think up fancy ways to talk about word counts, and that's exactly what I'll do in this essay. I invite you to think along with me about a few very basic questions: What does it mean to count words in a text? What happens to the text, and to our understanding of it, when we decompose it into a series of word counts? What relation exists between the textual domain and its numerical image? Or, to restate this question with a nod to Professor Fish, "Is there a text in my data?"

My answer will be, "No, not really." To show you how I get to that answer, I'll follow one document through a series of typical transformations: first into a simple list of words and their frequencies, then to a vector of elements in a matrix, and from there through the processes of normalization, dimensionality reduction, and analysis. At

each step along the way, I'll try to describe what has happened to this lexical thing — this textual object that, we'll see, no longer resembles a "text" in any accepted use of that term. However, I'll argue against the commonly held notion that counting words reduces complexity, and I'll suggest instead that semantic models embed textual objects in highly complex structures that, when constructed using relevant corpora, are extremely sensitive to historical context and subtle nuances in meaning. Word frequencies aren't static, given things that simply exist in a text. They're produced through the act of modeling, and the mathematical structures they imply dissolve both words and texts into elaborate systems of mutual interrelation.

The second installment, "Notes Toward a Mathematical Theory of Authorial Intention," will tackle a somewhat thornier problem. In his response to the controversy, Fish made a strange comment that has, as far as I can tell, escaped notice. He wrote, "The excavation of verbal patterns must remain an inert activity until added to it is the purpose of some intentional agent whose project gives those patterns significance."[15] This comment is strange because I doubt very many literature professors would agree with it — or, at least, few would consider it obvious enough to be dropped in passing as the justification and proof of their position. The whole question of intention is famously vexed. It's hard to interpret a poem and to say with confidence that your interpretation represents the intention of the author. On the other hand, it's hard to interpret a poem without having in mind the fact that somebody wrote it for a reason. The tradition of literary theory is riddled with essays puzzling over this conundrum. Given their lack of consensus in even the simplest of cases, who's to say we can't find intentions in literary data? I'll argue that we can. However, to make that argument, I'll first have to ask, "Is there an author in my metadata?" And I'll answer, "Sort of."

But more of that in the second installment. For now, I want to keep focus squarely on more basic questions. I want to take my time to think slowly about what happens when we transform a text into data. In particular, I'll create a *latent semantic map* of a document and a corpus. The procedures of latent semantic analysis have been well established since the 1990s.[16] When compared to more recent machine-learning applications, the math is a lot simpler and the process is much more straightforward, but everything I'll say about this technique applies, *mutatis mutandis*, to popular applications like topic modeling.[17] The document I'll focus on is drawn from

the *Early English Books Online* (EEBO) collection, but just for fun, I'll withhold its original title and refer to it only by its EEBO identifier: A43998. Most people reading this essay will have read the book or will at least know it by reputation, but I ask you not to cheat by looking up the title online. Instead, let's play a little game. As you read, at what point are you able to guess the author and title? What meanings has the analysis drawn forth, such that you're able to connect the dots between the numbers and the text? To what extent and in what ways does the quantitative analysis overlap with your qualitative sense of what the book is all about? The answers to these questions will differ for everyone. However, my hope is that all but the most entrenched partisans will find something of interest in the analytical perspective afforded by counting words. There might not be a text in my data, but there's something else, and that something else is fascinating, complicated, and worthy of further study . . . even by English professors.

From Texts to Tokens

The EEBO-TCP files are available for download from Github (an online file management service popular among programmers) in eXtensible Markup Language (XML) format. They were transcribed into XML by workers who visually inspected the image files scanned from the *Early English Books* microfilm collection. As documentary transcriptions of the underlying books, EEBO-TCP files record the original sequences of letters and punctuation as faithfully as possible while marking structural features like page breaks, chapter divisions, paragraph breaks, and the like. There are many ways a researcher might begin working with such a file — a well-written XML file can enable a wide range of analytical transformations — but for our purposes here I'll follow only the simplest and most common procedures.

The first step uses the XML structure to differentiate the textual data from the metadata (that is, from the bibliographical information stored at the top of the XML file); then captures anything that was transcribed from the original source; then dumps that information into a single, unbroken string of characters. At this point, A43998 is represented as a single sequence of letters, numbers, spaces, and punctuation marks. The other structural features of the book have all been erased. Yet, at this point, it still retains many features of the original. The sequence of letters and punctuation are all preserved, and with patience a person could conceivably read

the document in this format and come to a fairly trustworthy conclusion about what the original book was saying. At this stage, the text is not particularly amenable to counting, but there are two ways you might describe its length. On the one hand, its length is 1, because it exists in the computer as a single, unitary object. On the other hand, its length is 1,232,337, because that's how many characters A43998 includes, counting all spaces and punctuation marks.

The second step is to break that string into words. This step already involves making dangerously arbitrary decisions. The researcher must decide whether to separate contractions into distinct words, whether to "stem" the words (collapsing all conjugations of verbs or versions of nouns into canonical forms), and whether to retain capitalization. Generally speaking, these decisions remain undertheorized in the field of digital humanities, and I know of no study that tries to account systematically for their effects. For now, we'll adopt a willfully naive perspective and provisionally define a word as *any continuous sequence of alphabetical characters* contained within a larger string of characters. This definition will have to be further complicated in a few moments, but let's stick with it for now. After the large character string has been broken up in this way, a reader could still probably follow the text pretty well, but they'd have to imaginatively fill in the missing punctuation, as well as correct for transcription errors in EEBO's XML, so any number of interpretive mistakes would likely result. This would make a very bad reading text, but it would still be plausible as such.

The third step is to count the words. Easy enough. There are 211,997 of them, making our text about twice the length of a typical novel. That's very big for the EEBO collection, which includes lots of single-sheet broadsides and pamphlets, but it's a familiar length to us now, so we have good reason to believe that, whatever else might be true of A43998, it's probably a book in the enduring common sense meaning of the word, "book." However, to say that A43998 is 211,997 words long does not say much more than that. We've reduced the text down to a single *scalar* value. It gives you a vague sense of scale — "This thing is made of about two hundred thousand pieces." — but nothing more.

The question immediately arises: How many unique words can be found in the text, and how many of each can we find? And here is where things start to become a little

more interesting. Counting words requires differentiating each signifier from its various enunciations, a distinction corpus linguists make by differentiating word "types" from word "tokens." A type is a particular form of a word, and a token is every particular instance of that type. So any time you point to a word in a corpus, you're pointing at a token, and that token has a type. In A43998, there are 211,997 tokens of 10,328 types, giving it a type-to-token ratio of 4.9%. That's roughly twenty tokens for each type, a number that's really high because the book is so big. The ten most frequently appearing types in A43998, after all the words have been converted to lowercase, are *the* (14,849), *of* (10,850), *and* (7,305), *to* (7,236), *is* (4,864), *that* (4,786), *in* (4,194), *a* (3,122), *by* (2,636), and *for* (2,539). Those are just the top words, of course, and the list goes on from there. At this point, A43998 exists as a list of numbers exactly 10,328 items long. Each item in that list has both a value (the word frequency) and a label (the word type).

We should stop for a moment to consider what this thing is that we now have on our hands. We might still call this list of word counts a textual object, because it certainly depends for its existence and form on the original text from which it's drawn, but it isn't a text. A list of word frequencies like the one I've just created is not even a thing unto itself, really. Rather, it is a combination of three entirely separate mathematical objects: the set of word types, the set of documents, and the set of natural numbers. The set of word types contains 10,328 unique elements, all of which are related to each other because each is contained in A43998. The status of A43998 as a text survives only as principle of relatedness — of belonging — that binds these types and their tokens together. In his classic textbook on set theory, Paul R. Halmos writes:

The principal concept of set theory, the one that in completely axiomatic studies is the principal primitive (undefined) concept, is that of *belonging*. If x belongs to A (x is an element of A, x is *contained* in A), we shall write: $x \in A$.[18]

Any time we start counting words in a document, we start building sets, and therefore we find ourselves making propositions about what sorts of things belong together with what other sorts of things in what kinds of ways. In this case, we've already identified two conceptually distinct modes of belonging. All the types belong together because they all appear in A43998. All the tokens belong together for the

same reason. However, some of the tokens belong together in a different way by virtue of being the same type.

Proposition 1. *Word frequencies describe the observed intersection between two sets of elements: the set of all allowed word types and the set of all allowed textual objects.*

As we'll see, a great deal rests on the word "allowed." For now, we're still taking small steps. In the simplest case we've been following so far, this transformation happens by mapping the observed tokens onto the set of natural numbers. (The "natural numbers," you'll recall, are the integers greater than zero.) A list of word frequencies can have no values lower than 1 and no values that aren't integers, because we're just naively counting words that happen to appear in A43998 one at a time. Further, that list unfolds over a single axis because we've constrained our set of documents to a single element. The only allowed document is A43998.

This procedure represents, as far as I can tell, what most literature professors imagine when they complain that counting words is hopelessly reductive. And if that were really the end of it, they'd be right to complain. The dataset we've compiled contains 10,328 nearly meaningless facts. For each word type, we're told that $x \in$ A43998, but that's true of every word in the list and so provides no means to differentiate among them. The frequency values themselves tell us little else. We know that *the* appears more often than *of*, but we have no framework for deriving any meaning from that fact. Even Jerome McGann or Franco Moretti would struggle, I think, to pretend to interpret this collection of factoids. From the perspective of cultural analytics, the problem with word-frequency counts is not that the rich plenitude of the text has been evacuated (the typical outsider's complaint) but that the document has been dissolved into a messy lexical goo that lacks any meaningful structure. You simply can't do anything with lists of word counts, which means you can't know anything about them.

To escape from this mess, we'll need to enlist the aid of something very special. Not a machine-learning algorithm; not some fancy software. We'll need help from the number zero.

From Tokens to Matrices

By themselves, word frequencies can't be analyzed in any but the most cursory way. To be compared mathematically, they must be converted into a regularized form, sometimes called a "vector" or a "distribution over a fixed vocabulary." A fixed vocabulary is a set of word types used for analysis over an entire corpus, regardless of whether the words are included in any individual document. For words that are both in the vocabulary and in the document, the process is pretty much the same. But words that aren't in the document get counted as zero, and words that aren't in the vocabulary get discarded completely. Rather than a list of word counts, this process gives you a structured vector — a lexical frame, derived from the corpus, over which the document's word frequencies are now stretched. Vectorization has two important consequences: 1) In practical terms, it renders word counts from one document numerically commensurable to others in the corpus, so you can perform statistical comparisons across them. 2) In theoretical terms, using zeroes profoundly alters the ontological condition of the textual object.

Let me explain that second point before coming back to the first.

When creating a fixed vocabulary, you're demarcating the field of possibility within which your documents are described. Nobody asks how often William Shakespeare used the word *ipad* because the word didn't exist until hundreds of years after he died. It's neither in any of his books nor in any of his contemporaries'. To say that Shakespeare wrote *ipad* zero times would be true, but a very weird thing to say. However, it's not at all weird to say that he never used the word *theology* — that little factoid would differentiate his works from at least some of EEBO's books — and it might be even more interesting to know that Shakespeare's comedies never use the word *christ*, even though a handful of his history plays do. Consequently, *ipad* doesn't belong in a fixed vocabulary of Shakespeare's works, but *christ* definitely does, and *theology* might, depending on the kinds of claims you hope to make.

The most important step in data analysis is deciding where to put your zeroes. I want to interpret A43998 in historical context, so to build a fixed vocabulary for it, I begin by randomly selecting 4,999 other documents from the corpus, counting the words

in each, and selecting only those words that appeared in at least one hundred of them. I then further pared the vocabulary down to include only the 10,000 most frequent words among those that are left. This had two immediate consequences for our list of frequencies. First, the initial set of 10,328 unique words is scrapped. Only 5,890 types made it through to the final vocabulary. Second, an additional 4,110 bits of data are inserted into the vector, all of which are zero values. These 4,110 zeroes represent words that occur prevalently throughout our sample of EEBO but nowhere in A43998. Thus, only 59% of the data in our A43998 vector actually comes from A43998. The other 41% comes from the corpus and is made up of words that A43998 might have used, but didn't. The resulting vector is therefore a strange kind of hybrid entity, both an empirical statement about what was found in the document and a hypothetical statement about what's missing. Keep in mind, too, that A43998 is an unusually large document. Most books in EEBO are much shorter. On average, documents in our sample use only 1,473 words from the vocabulary, which means that about 85% of all values in our data are zeroes. This situation is quite typical. In data analysis, objects exist primarily in terms of what you think they might have been.

At this point, our vector of word counts bears little resemblance to the original text. Yet, this form gives us lots of new information about the original. Each zero in a word-frequency vector represents a counterfactual proposition about the text based on what we know about the corpus overall. (See Table 1.) You'll notice that many of the most frequent vocabulary terms missing from A43998 are short nonsense words, like *per* and *ter* that reflect anomalies in the underlying XML, typical of EEBO files, which were transcribed in a way that sometimes cuts words in half. In most studies, they'd be excluded as stop words or otherwise corrected during text processing. For our purposes, they don't hurt anything. Looking past those anomalies, we can see that many documents in our selection of EEBO use relational terms like *near* and *met*, or terms of affect like *cry* and *tender*, but those words are completely absent from A43998. This tells us something — not much, perhaps, but something. The absent terms seem to connote personal and intersubjective experience, so whatever our text is, it's probably neither a novel nor a collection of poems or plays. We can get a more positive sense by looking at the words that are stripped away. Terms like *soveraigns* and *representative* appear dozens of times in A43998 but not enough over the corpus as a whole to be included in the vocabulary.

This means our book is probably a treatise on political theory or natural law, perhaps with an unusual concern for *incorporeall* things like *ghosts*. Weird or archaic spellings of common words suggest that it was probably published no later than 1680 or so.

Table 1. A comparison between the word frequencies in A43998 and the fixed vocabulary drawn from the sample corpus. The left column shows the most prevalent types not present in A43998. The right column shows types that appear most frequently in A43998 but are excluded from the vocabulary.

Document frequency (of 5,000 possible)	Number of tokens (in A43998)
per (2394)	soveraigns (87)
ter (2394)	representative (76)
al (2321)	incorporeall (36)
m (2167)	dependeth (32)
ted (2076)	politique (30)
es (2075)	ghosts (26)
pro (2017)	judicature (24)
near (1919)	legislator (24)
hall (1878)	forraign (23)
com (1868)	fundamentall (23)
tender (1860)	dammage (22)
er (1820)	disposeth (22)
view (1807)	expressely (22)
met (1784)	politiques (22)
publick (1749)	supernaturally (22)
cry (1729)	democracy (21)
fore (1711)	aristocracy (20)
sed (1710)	artificiall (20)
ment (1702)	latines (20)
ting (1680)	phantasmes (20)

The next step in data curation is to read back over the corpus again, using the fixed vocabulary to record frequency values for each document and to place those values in a large matrix, with a row for each of the 10,000 word types and a column for each of the 5,000 documents. In the jargon of linear algebra, each of these sets represents its own vector space — the vocabulary defines what might be called "lexical space," and the document titles define what could be called "bibliographical space."[19] Inserting A43998 into a structure like this again changes its ontological status. Our data is no longer a vector by itself but now exists as part of a larger

semantic system — a "linear map" that transforms one vector space into another and so describes the relations among words and books that in fact now constitute words and books as such.

This last point is very abstract. At this stage of analysis, every document in the corpus is now represented over the same fixed vocabulary (lexical space), and every word is represented over the same series of titles (bibliographical space). While it's true, from one perspective, that documents are represented as frequencies of words, it's simultaneously true that words are represented as frequencies of documents. That is to say, documents are vectors of words and words are vectors of documents.

Proposition 2. *In a vector-space semantic model, words and documents are mutually constituted by the linear transformation of lexical space into bibliographical space.*

For this reason, once placed in a matrix, the numbers of any given row or column are never quite identical to themselves, because the matrix is an elaborate proposition about their mutual interrelation. It's no longer quite right to say that *for* appears in A43998 2,539 times, because *for* is no longer a word in any conventional sense — it is neither type nor token. Instead, *for* is a sequence of 5,000 values, labeled by EEBO number, of which "A43998" is just the first. Indeed, it makes just as much sense to say that "A43998" appears in *for* 2,539 times. Neither perspective captures the full truth because both are true simultaneously.

Here's the main point: Word counts aren't word counts at all, as literature professors understand the phrase. Why? Because every document is a system of words made of documents; every word, a system of documents made of words. Rather than say we're representing a document as a series of word frequencies, it's more accurate to say that each text is represented as a structured set of historically relevant relations to historically relevant contexts. As we'll see, this dialectical structure makes semantic models extraordinarily useful for describing the distribution of difference in any corpus.

Normalization & Analysis

At this point, you might be hoping for an example or two, but the term-document matrix would not be very helpful in that regard. We could look at the ten books that

appear in *for* most frequently, but that would just give us the ten biggest books. They wouldn't have any relationship other than that. Instead, we need to know which values are unusually high or unusually low, and so we need to normalize the matrix to better reflect those variations. Many techniques exist for this task: TF-IDF, entropy weighting, z-scores, etc. They all have subtle differences, of course, and information scientists debate which weighting schemes are best for which tasks, but they all work pretty much the same way and have more or less the same effect. For our dataset, I'll transform our matrix using a formula called *positive pointwise mutual information* (PPMI), a statistical process commonly used when preparing data for semantic models. The basic idea behind PPMI is to weight each value in the matrix by type frequency and document size, then to see how far the actual values deviate from these baseline expectations.

Every value in our textual object now represents an element in a statistical model of A43998's significant relationships to all other documents in the corpus. After processing, our data has become fairly sophisticated — it's gotten fancier, you might say — and so we are now able to describe major themes in the text, as well as to identify exemplary historical frames of reference for each of its keywords. In A43998, the words with the highest PPMI scores are *sove* (4.08), *civill* (4.07), *representative* (3.87), *soveraign* (3.82), *soveraignty* (3.79), and *nevertheless* (3.58). Each of these terms represents a distribution over the model's bibliographic space and so embeds A43998 within a network of similar documents. (See Table 2 and Table 3.) Documents where the term *soveraign* is over-represented tend to be relatively short political works, often addressing the crown directly, making proclamations, or referring to specific events. The term *commonwealth* appears most distinctively in political discourse published during the Interregnum period. Those are just two examples. The normalized matrix represents within itself all of these intertextual connections.

Table 2. EEBO documents with the highest PPMI score for the term *soveraign*. These values represent one small snapshot of the term's complex representation within the model. The documents all become connected to A43998 by virtue of their shared emphasis on the term *soveraign*.

TCP	Date	PPMI	Title
B03109	1666	6.19	*Englands tryumph, and Hollands downfall.*
A58549	1685	5.80	*Act anent the covenant Edinburgh, May 8, 1685.*
B09606	1695	5.60	*The Earl Marshal's order for going into second mourning . . .*

B04010	1660	5.50	*Lætitiæ Caledonicæ, or, Scotlands raptures . . .*
N00665	1697	5.45	*By His Excellency Collonel Benjamin Fletcher captain general . . .*
B43921	1681	5.23	*Act ratifying all former lavvs for the secuity of the Protestant religion. . .*
N29539	1699	4.87	*Province of the Massachusetts-Bay . . . A proclamation . . .*
B05164	1696	4.83	*Act anent the old fourteen shilling pieces and their halfs . . .*
A91202	1657	4.74	*King Richard the Third revived. Containing a memorable petition. . .*
A92469	1693	4.68	*Act against corresponding with France. Edinburgh. . .*

Table 3. EEBO documents with the highest PPMI score for the term *commonwealth*.

TCP	Date	PPMI	Title
B09282	1652	5.57	*Ireland. By the Commissioners of the Parliament of the Common-wealth. . .*
A82953	1660	5.38	*Die Mercurii 9. Maii, 1660. Ordered by the Lords and Commons . . .*
A74528	1653	4.98	*An ordinance declaring that the offences . . . shall be adjudged high treason. . .*
B02497	1653	4.98	*A proclamation of His Highnes, with the consent of his Council. . .*
A87132	1659	4.97	*The spirit of the nation is not yet to be trusted with liberty. . .*
A91095	1659	4.90	*A proposition in order to the proposing of a commonvvealth . . .*
A83734	1643	4.90	*It is this day ordered by the House of Commons. . .*
A74137	1654	4.79	*By the Lord Protector. Whereas the enemies of the peace of this nation ...*
A82752	1653	4.78	*A declaration of the Parliament of the Commonwealth of England. . .*
A38099	1652	4.75	*Resolved by the Parliament . . . that any cattle, sheep, horses, corn, or grain. . .*

Understanding this point is absolutely essential if you want to understand how these models work, but nowhere in the critical discourse have I seen it raised, and in fact the whole question of "distant reading" directs attention away from it. The explanatory power of data analysis doesn't come from its ability to show "the big picture." Nor does it come from reducing the complexity of its subject. Quite the opposite, in fact: A43998 is now several orders of magnitude more complex than the original book. After this kind of processing, you could never read A43998 and comprehend it because at every single point it's more complicated than any thought you could ever hold in your head. The explanatory power of data analysis instead comes from systematically gathering knowledge about a total population — in the humanities, that's usually a population of words — then using that general picture as a framework to characterize every individual member of the population. Through this dialectical process, every instance is stamped with an image of the whole against

which its unique properties become more clearly visible. At this point, every single datum in A43998 is derived from corpus-level measurements.

Proposition 3. *Every value in a vector-space model reflects information gathered over the entire corpus*.

For this reason, I believe that literary scholars have been misled by distant reading's leading polemicists. (I'm speaking here most directly of Moretti, but we all share some blame.) Quantitative literary analysis doesn't work by reducing the complexity of an object in order to expose big patterns or macro trends. That idea took root in the discipline long before any of us knew enough about the subject to know whether it was true. Finding simple patterns might be what scholars want from their analyses, but it's not what the methods most directly entail. Instead, *analysis introduces complexity by representing explicit relationships among all elements in the data*. In a printed book, words are connected by mere punctuation and pagination. In a vector-space model, every word is connected to every other in a complex network. The only reason it feels like words on a page have greater complexity is because only there are they simple enough for us to read them.

So what do we do, trapped as we are in our puny human brains? We draw charts. We tabulate keywords. We fumble summary. Anything to reduce the enormous, monstrous complexity of our creation.

Much research in natural language processing and information retrieval is devoted to precisely this task. The final step of data processing is dimensionality reduction. This is a term for the mathematical operation called singular-value decomposition, through which a matrix is decomposed into three component parts. In a latent semantic model, these components include a matrix showing relations among word vectors, another showing relations among the documents, and a third that identifies the scale of each dimension. Those scales are called "eigenvalues." Because of how matrix multiplication works, and because of how eigenvalues are computed, the largest dimensions identify the most common points of collocation among the words and documents. Identifying the eigenvalues allows you to disregard the smallest dimensions, thus reducing the sensitivity of the model to subtle variations in the data and generating results that more closely approximate human ways of thinking. Whereas the original term-document matrix was sparse — meaning most of the

values were zeroes — the resulting eigenvectors are dense. All the zeroes are removed and replaced with scalar values that locate each word and each document along the latent axes of a common semantic space.

Proposition 4: *The distribution of tokens in lexical space will tend to correlate with the distribution of tokens in bibliographical space.*

Once the dimensions of this space have been mapped, it becomes possible to identify which words and documents share the most points of overlap. Because such similarities often correspond with intuitive notions of what documents are about and what words mean, information scientists refer to these patterns as "semantic." However, these statistical patterns relate only indirectly to the kinds of meaning readers experience while reading. The meanings of the words don't exist as paraphrasable definitions but as structured distributions over thousands of documents which are, themselves, structured distributions over thousands of words.

Let's see what happens when we take another look at A43998 through this new lens. Identifying terms that cluster together within semantic space reveals the underlying geometry of associations that subsists among words. Using data gathered at the corpus level, Table 4 shows keywords from A43998 again, which I've now grouped using a technique called *k*-means clustering. I found ten anchor points at various locations in the document's semantic space, then found the words that sit closest to each point.[20] The most conceptually dense region of this space — where the terms are most overrepresented in A43998 and where they overlap most visibly — contains terms relating to theories of natural law. Our document is most predominantly concerned with questions about *government*, *authority*, and the *power* of *law*. Underneath this specific concern is also a sustained engagement with scriptural tradition, including words related to the Old Testament (*moses*, *prophet*, *israel*, *jews*) and the New Testament (*scripture, doctrine*, *teach*, *apostles*, *christian, taught*). Alongside these political and historical concerns we find also a more general intellectual engagement with a discourse of philosophical disputation — words like *reason* and *nature* suggest that the author of A43998 hoped to ground the political theory within a more fundamental account of nature and reality. (Notice how I said that the measurements suggest what the author of A43998 hoped to communicate. I'll return to this point in the next installment.)

Table 4. Semantic clusters in A43998. Similarity is measured by taking the cosine among word vectors in an LSA model with 100 dimensions. Terms were selected if they met two criteria: if their overall word frequency was above average over the 5,000 document sample from EEBO, and if their PPMI score was above average for A43998. Within this semantic space, ten anchor points were identified using *k*-means clustering, and each group shows the words closest to each point. Groups are ranked by the total PPMI score for the twelve words that sit closest to each anchor.

Total PPMI	Word Groups
17.54	laws, liberty, government, power, private, subject, authority, bound, right, law
16.82	subjects, command, whereas, contrary, laws, private, subject, authority, obedience
15.90	moses, prophet, prophets, israel, jews, spoken, worship, scripture, teach, ghost, saviour
13.89	own, again, living, judge, every, another, private, honour, words, subject, already
13.61	reason, proceed, else, subject, sometimes, because, discourse, use, question, nature
13.55	salvation, acts, obedience, covenant, saviour, judge, therefore, own, jews, ghost
13.30	scripture, doctrine, teach, ghost, apostles, spoken, christian, worship, taught, say, false

While Table 4 shows how the semantic model organizes words inside A43998, Table 5 shows how our document is now situated among a field of contemporaries. I searched for the thirty documents most similar to A43998 and grouped them, again, using *k*-means clustering. These documents were mostly published in the 1640s and 1650s, during the English Civil War and Interregnum. (The median publication date among these thirty books is 1647.) Understandably given this history, they are almost all devoted to questions about religion and politics, often focused very specifically on debates over ecclesiastical authority and its relation to state power. This list returns political treatises by Spinoza, John Milton, John Bramhall, Robert Filmer, and Henry Parker. The document with which A43998 is most similar — indeed identical, because it's the book A43998 was derived from — is Thomas Hobbes's *Leviathan* (1651).

Table 5. Documents most similar to A43998. Results are grouped by *k*-means clustering and, within each cluster, sorted by cosine similarity to A43998.

Author	Date	Title	Similarity
Hobbes, Thomas.	1651	*Leviathan. . .*	1.00
Tenison, Thomas.	1670	*The creed of Mr. Hobbes examined . . .*	0.78
Spinoza, Benedictus de.	1689	*A treatise partly theological. . .*	0.77
Bucer, Martin.	1644	*The Ivdgement of Martin Bucer concerning divorce. . .*	0.71
Erastus, Thomas.	1659	*The nullity of church-censures. . .*	0.71
Ellis, John.	1700	*A defence . . . of the Church of England . . .*	0.68

Hobbes, Thomas.	1652	*De corpore politico. . .*	0.89
Filmer, Robert.	1680	*Patriarcha, or, The natural power of Kings. . .*	0.72
Bramhall, John.	1655	*A defence of true liberty . . .*	0.72
Coke, Roger.	1660	*Justice vindicated . . .*	0.69
Ball, William.	1656	*State-maxims. . .*	0.69
Wren, M. (Matthew).	1659	*Monarchy asserted. . .*	0.68
Milton, John.	1649	*The tenure of kings and magistrates . . .*	0.77
Well wisher to the Church.	1642	*The unlimited prerogative of kings subverted . . .*	0.77
Mayne, Jasper.	1647	*Ochlo-machia. Or The peoples war. . .*	0.76
Milton, John.	1642	*A reply to the Answer . . .*	0.72
J. L.	1649	*Illumination to Sion Colledge . . .*	0.71
Gee, Edward.	1650	*A vindication of the Oath of allegiance. . .*	0.70
Ball, William.	1645	*Tractatus de jure regnandi, & regni . . .*	0.69
Anon.	1652	*The key of true policy . . .*	0.68
Nicanor, Lysimachus.	1639	*The ungirding of the Scottish armour . . .*	0.67
Daniel, Samuel, 17th cent.	1642	*Archiepiscopal priority instituted by Christ. . .*	0.76
Parker, Henry.	1641	*A discovrse concerning Puritans . . .*	0.73
Baker, Richard.	1641	*An apologie for lay-mens writing in divinity. . .*	0.70
Maxwell, John.	1641	*Episcopacie not abivred in His Maiesties realme. . .*	0.70
Morley, George.	1641	*A modest advertisement . . .*	0.70
Strode, William.	1644	*A sermon concerning svvearing . . .*	0.69
Jackson, John.	1640	*The key of knowledge . . .*	0.68
Noyes, James	1646	*The temple measured . . .*	0.68
Ainsworth, Henry.	1609	*A defence of the Holy Scriptures. . .*	0.67

Conclusion

I began this essay by asking the question, "Is there a text in my data?" My answer is "No." The reason, I think, is fairly obvious. When you decompose a document into its constituent elements, you not only lose the pagination that makes reading possible, you also fill it with information derived from elsewhere, and, in doing so, you assemble that information into a new structure. Rather than reduce the sophistication of the text, a semantic model embeds that text in a web of others. To say that such methods are "just fancier ways of talking about word frequency" is to misapprehend them rather profoundly. Computational literary analyses are just made

of word counts like computer graphics are just made of pixels, like human bodies are just made of cells, and like societies are just made of people. Complex systems are built of simple blocks. A semantic network is a great leviathan of words.

In the above discussion, I advanced four propositions that I believe are fundamental to the practice of quantitative text analysis. First, that word counts represent an observed relationship between two very different kinds of elements: documents, which demarcate syntagmatic stretches of written discourse, and word types, which identify paradigmatic points of connection across those stretches. Second, that, in such a model, words and documents are dialectically constituted systems of collocation. Third, that, after statistical processing, every individual value is represented in relation to measurements taken over the whole. And, fourth, that words and documents correlate in meaningful ways. Together, these propositions explain why semantic models have such incredible expressive power. They allow for confident generalizations across large collections of texts as well as highly detailed examinations of individual cases.

I did not attempt to put forward a reading of *Leviathan*, whether close, distant, or otherwise. Nonetheless, I think it should be clear how a reading could be supported using these techniques. Scholars with a purely instrumental view of computation — who think analyses like these are "tools" that help answer questions they care about for other reasons — are perfectly free to think of semantic models simply as concordancing and indexing methods. In the time it took to write this short essay, I created a detailed concordance of *Leviathan*, identifying its keywords and their most historically relevant correlations, as well as a structured bibliography of early modern books most relevant to each term. The tables presented above reveal only the tiniest of glimpses at the data I've collected. To say that there's nothing we could learn using these methods is like saying there's nothing we can learn from consulting dictionaries or bibliographies.

It will also be clear, I hope, why advocates for cultural analytics believe these methods can be scaled up to describe, for example, differences among genres or change over time. If we can correctly identify information about individual documents like A43998, there's no reason to assume that systematic comparisons based on word-frequency data are ill-founded or based only on metaphors or false

analogies. Scaled up to the entire EEBO dataset, I'd have the same detailed information about more than 60,000 early modern books. To generalize from such evidence is perfectly reasonable.

As I said in the introduction, I'm not really a literary scholar, but, like Ted Underwood and others, I sometimes use numbers to show change over time or variation across geography. I sometimes use computational methods as tools for answering literary historical questions. But those are rarely the true stakes for me. My interests are mostly theoretical. Looking again over the data objects that now litter my computer's desktop: What are these things I've created? What are their most important features? How do they work? To what in the world do they relate? How can they best be described?

These aren't texts. They're an entirely new form of textuality.

Seriously, what the fuck are these things?

Because researchers in the information sciences have adopted a purely instrumentalist view of their own inventions, they've left these questions almost completely unexplored. It is perfectly legitimate, and indeed I believe quite important, for scholars of language and literature to pick up where they left off.

Notes

1. I would like to thank the following people for reading earlier versions of this essay: Eric Gidal, Jeanne M. Britton, Ed Madden, Seulghee Lee, Jonathan Edwards, and Brian Glavey.
2. Nan Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45 (Spring 2019): 601-39, 601.
3. In Section 9, item 4 of her Appendix, Da argues that journal editors should, during peer review, "Enlist a statistician to a) check for presence of naturally occurring, 'data mining' results, implementation errors, forward looking bias, scaling errors, Type I/II/III errors, faulty modeling, straw man null hypothesis, and others; b) see if data work is actually robust or over-sensitive to authors' own filters/parameters/culling methods; c) see if insights/patterns are actually produced by something mechanical/definitional, d) apply Occam's razor Test — would a simpler method work just as well?"
4. Da, "Computational Case," 623-24, 612-13.
5. Da, "Computational Case," 610-11.

6. On *Critical Inquiry*'s blog, Ted Underwood, Mark Algee-Hewitt, Richard Jean So, and Hoyt Long all defend their research against specific critiques. Andrew Piper does not directly address Da's critique of his study of *Confessions*, but he cites Benjamin Schmidt's "A computational critique of a computational critique of computational critique," http://benschmidt.org/post/critical_inquiry/2019-03-18-nan-da-critical-inquiry/ which offers a detailed defense.
7. "Final Comments." https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-4/ Da elaborates on her decision to disregard the advice of her peer reviewer this way: "Statisticians or computer scientists can check for empirical mistakes and errors in implementation; they do not understand what would constitute a weak or conceptually-confused argument in literary scholarship. This is why the guidelines I lay out in my appendix, in which many people are brought into peer review, should be considered." Nowhere does she explain under what conditions a statistician's review, once consulted, ought to be disregarded.
8. Da offers a corrected definition of *p*-values in her response,"Errors." https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2/
9. Sarah Brouillette's response represents and reflects this more general line of critique. https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-3/. Brouillette writes, "Universities for their part often like DH labs because they attract these outside funders, and because grants don't last forever, a campus doesn't have to promise anything beyond short-term training and employment." See also David Allington, Sarah Brouillette, and David Golumbia, "Neoliberal Tools (and Archives): A Political History of Digital Humanities," *Los Angeles Review of Books* May 1, 2016. https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/#!
10. In her second contribution to the forum, Lauren F. Klein points to "structural deficiencies" that disregard work "performed disproportionally by women and people of color," concluding: "In the end, the absence of the voices of the scholars who lead these projects, both from this forum and from the scholarship it explores, offers the most convincing evidence of what—and who—is valued most by existing university structures; and what work—and what people—should be at the center of conversations to come." https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3/ See also Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern University Press, 2018).
11. Offering a novel perspective on the nature of professional expertise, Fish argues that "ignorance is no bar to my pronouncing on the Digital Humanities because my objections to it are lodged on a theoretical level in relation to which actual statistical work in the field is beside the point. I don't care what form these analyses take. I know in advance that they will fail." Fish continues to say, "I was pleased therefore to find that Professor Da, possessed of a detailed knowledge infinitely greater than mine, supports my relatively untutored critique." Stanley Fish, "Response." https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-5/ Something like Fish's position was anticipated by Benjamin Schmidt, who, writing two weeks earlier, argued that Da's statistical claims were written to appeal to readers ill-equipped to evaluate them: "This blizzard of terminology establishes for the innumerate reader that they finally have an expert who will debunk statistics for them, while freeing them of the burdensome requirement to think for themselves." http://benschmidt.org/post/critical_inquiry/2019-03-18-nan-da-critical-inquiry/
12. Da mentions research in the field only obliquely by describing their "applications": "Typical applications of textual data mining involve a trade-off: speed for accuracy, coverage for nuance. Such methods are efficient for industries, sectors, and disciplines that are dealing with so much textual data at such fast speeds that they cannot possibly (nor would want to) read it all or where

one wants to extract from a large data set a relatively simple piece of information that is either actionable or that can be quickly labelled and classified along simple features . . . The information that is extracted is not supposed to be semantically complicated." "Computational Case," 620.

13. In this respect to this issue, Da remarks: "People who *can* do this work on a high level tend not to care to critique it, or else they tend not to question how quantitative methods intersect with the distinctiveness of literary criticism, in all its forms and modes of argumentation." "Final Comments" https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-4/ At the risk of citing myself, allow me to note here that I once wrote a book about the history of literary criticism, and in that book I argue at length against any notion of the field's "distinctiveness." I also compare the debates that surrounded early English criticism to debates that have surrounded digital humanities: "In their respective moments of media flux, both early criticism and the digital humanities confront the worry (or the exhilarating promise?) that old forms of literary knowledge will become obsolete and forgotten while new genres dissolve into mere chatter. Then and now, criticism's fault-lines and boundaries are marked by sharp, biting polemical debate. . . . The field's peculiar rallying cry has long been the same: What we do is vitally important. Everything we're doing is horribly wrong." *The Invention of English Criticism, 1650-1760* (Cambridge University Press, 2015), 8, 23.
14. Da, "Computational Case," 607.
15. Fish, "Response." He continues, "Once you detach the numbers from the intention that generated them, there is absolutely nothing you can do with them, or, rather (it is the same thing) you can do with them anything you like." Fish does not explain how computational analyses differ in this regard from any interpretive paradigm that de-prioritizes intention as a guiding framework.
16. The history and theory of latent semantic analysis is covered in detail in Thomas K. Landauer, et al., eds., *Handbook of Latent Semantic Analysis* (Lawrence Erlbaum, 2007). In that volume, see in particular the essays, "LSA as a Theory of Meaning," by Landauer, and "Mathematical Foundations Behind Latent Semantic Analysis," by Dian I. Martin and Michael Berry. A classic description of the method can be found in George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum, "Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure," *Proceedings of the ACM-SIGIR* (1988): 465-80. http://susandumais.com/SIGIR1988-LSI-FurnasDeewesterDumaisEtAl.pdf For a concise and highly readable summary of the method, see Jerome Bellegarda, *Latent Semantic Mapping: Principles & Applications* (Morgan Claypool, 2007). I follow Bellegarda's procedures closely, with one exception. When normalizing the model, I do not follow his suggested entropy weighting technique, but instead use positive pointwise mutual information (PPMI), as described in Peter D. Turney and Patrick Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (2010): 141-188.
17. For the relationship between latent semantic analysis and probabilistic topic modeling, see Mark Steyvers and Tom Griffiths, "Probabilistic Topic Models," in *Handbook of Latent Semantic Analysis*, 427-48.
18. Halmos, *Naive Set Theory* (1960), p. 2.
19. In the simplest and most typical case explored in this essay, the bibliography corresponds to its common definition as a list of document titles, but I use the term "bibliographical" in an idiosyncratic way to refer to syntagmatic sets of tokens, whether organized by document titles or by supratextual categories (e.g. authors' names, genre labels, dates), intratextual structures (e.g. paragraphs, sentences, context windows), or any other configuration of belonging that corresponds to a colloquial sense of "appearing together" somewhere in the textual record. This definition does not depend on book title for its organizing scheme, but refers more generally to any index of syntagmatic relations. Within the topology of a corpus, bibliographical sets differ from lexical sets, because lexical sets posit paradigmatic relations based on word type.

20. Terms were selected for inclusion by two criteria: if they are both over-represented in the vocabulary (to ensure I'm getting the most important words) and have above-average PPMI scores (to ensure I'm getting words most relevant to A43998). Keep in mind, of course, that if any of these settings were changed, Table 4 would include different words in different groups. I had to play with the model a bit before I discovered the hyperparameters that would return lists that most clearly resemble what human readers might call "topics." Critics are sometimes scandalized by this aspect of quantitative literary research, and no doubt if any such critic were to tinker with my data, they'd be able to make the semantic clusters come out in ways that seemed to scramble the results. Nothing in my argument hinges on the particular contents of these groups, nor does this mean the model fails any test of robustness. Because of how PPMI and similarity are measured, no two words could appear together in a table like this without a better than fair-odds chance of appearing together in the same documents. The odds that 100 words (out of 10,000) with randomly generated vectors would sort so coherently, no matter how you might fiddle with the hyperparameters, are less than miniscule. There are so many null hypotheses I could reject, I wouldn't know where to start.