A fragmentising interface to a large corpus of digitized text: (Post)humanism and non-consumptive reading via features

Sayan Bhattacharyya, Peter Organisciak and J. Stephen Downie
sayan@illinois.edu, organis2@illinois.edu, jdownie@illinois.edu
Graduate School of Library and Information Science,
University of Illinois, Urbana-Champaign

Abstract:

While the idea of distant reading does not rule out the possibility of close reading of the individual components of the corpus of digitized text that is being distant-read, this ceases to be the case when parts of the corpus are, for reasons relating to intellectual property, not accessible for consumption through downloading followed by close reading. Copyright restrictions on material in collections of digitized text such as the HathiTrust Digital Library (HTDL) necessitates providing facilities for *non-consumptive* reading, one of the approaches to which consists of providing users with *features* from the text in the form of small fragments of text, instead of the text itself. We argue that, contrary to expectation, the fragmentary quality of the features generated by the reading interface does not necessarily imply that the mode of reading enabled and mediated by these features points in an anti-humanist direction. We pose the fragmentariness of the features as paradigmatic of the fragmentation with which digital techniques tend, more generally, to trouble the humanities. We then generalize our argument to put our work on feature-based non-consumptive reading in dialogue with contemporary debates that are currently taking place in philosophy and in cultural theory and criticism about posthumanism and agency. While the locus of agency in such a non-consumptive practice of reading does not coincide with the customary figure of the singular human subject as reader, it is possible to accommodate this fragmentising practice within the terms of an ampler notion of agency imagined as dispersed across an entire technosocial ensemble. When grasped in this way, such a practice of reading may be considered posthumanist but not necessarily antihumanist.

A fragmentising interface to a large corpus of digitized text: (Post)humanism and non-consumptive reading via features

For large corpora of text to be useful at scale, "distant reading" — a practice of reading so called on account of its abstracted, bird's-eye view of many texts in aggregated form — is a condition of knowledge (Moretti 2013, 48). Ted Underwood has recently pointed out that recent advances in algorithmic processing of big data at scale and the uptake of these methods by practitioners of the humanities have converged to make the inauguration of an unprecedented interdisciplinary conversation between computer science and the humanities overdue. Underwood points out that while the "hermeneutic cycle is intuitive enough when we're talking about a single text," it is much less so when what is being interpreted is a large *collection* of texts. When "a collection [is] too large to be surveyed by a single reader," then the task of data mining is to explain how it can work at that scale (Underwood 2014a, 67) — an explanation which may turn out to be not at all obvious. Underwood suggests that "humanists are gearing up to have a conversation about digital research methods" (64), and that "a new kind of interdisciplinary conversation" between humanists and computer scientists is about to begin, one in which "a rare opportunity is emerging for a genuinely productive exchange between scientific methodology and humanistic theory" (70). In this paper, we attempt to take up Underwood's suggestion in the context of the feature-extraction service that we are developing at the HathiTrust Research Center (HTRC). We attempt to situate the technological design of the functionality of the service (which is driven by the practical necessities imposed by laws pertaining to intellectual property) within the frame of larger conversations in cultural studies and science studies by making the fragments of text in the form of features, generated by our reading interface, serve as

a proxy for the trope of fragmentariness in general.[1] The question of fragmentation troubles the relationship between digital techniques and the humanities in general; this is why, although the HTRC is pursuing work with features out of a practical necessity, this work is worth our attention and engagement at a theoretical and philosophical level as well.

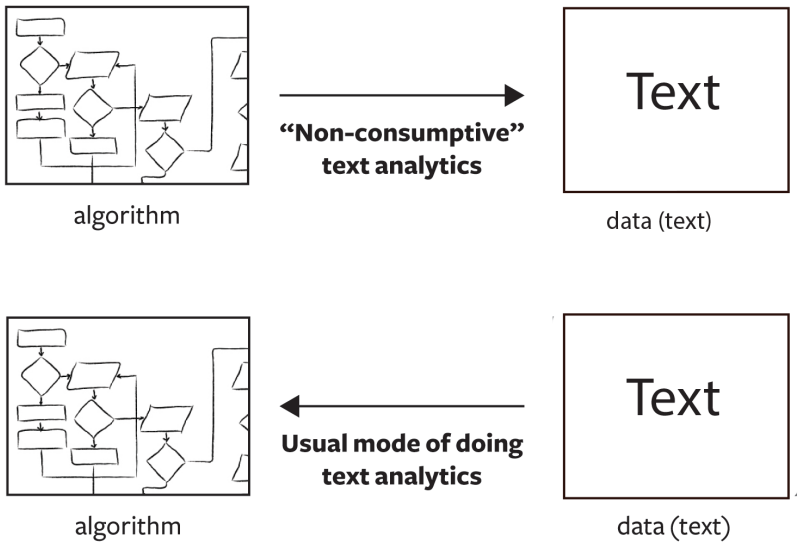

Fig. 1: Non-consumptive text analytics using a "data capsule" strategy. Instead of the user downloading the text and running algorithms against the text data (that is, "bringing the text to the algorithm"), in non-consumptive reading using a "data capsule" strategy, the algorithm is made to run securely against the text data, which is kept encapsulated from the user ("bringing the algorithm to the text").

The HathiTrust Digital Library (HTDL) comprises digitized text from the holdings of the world's great research libraries that are members of the HathiTrust consortium, and constitutes a valuable part of mankind's cultural legacy. However, more than seven million volumes out of a total of the more than eleven million volumes in the

[1] For reasons that will become clear later in the paper, we use the concepts of feature *extraction* and feature *generation* interchangeably. From the point of view in this paper, at the abstraction that we denote as the reading interface, features are generated by the interface itself. That they are actually extracted from the underlying text is hidden by the abstraction and is not relevant. As we will see later in the paper, this allows us, following the ideas of the agential "cut" and posthumanist performativity suggested by the philosopher of science Karen Barad, to think of the agency implicated by the overall reading process as partially distributed over the interface itself.

HTDL corpus are protected by some kind of copyright restriction. A bulk offering of texts that can be read individually would violate intellectual property laws in many countries. As the HTRC evolves, we have started to explore methods for "non-consumptive reading," which has become necessary to allow for the insights of large-scale distant reading to be available to researchers while operating within the above constraints. In the paradigm of non-consumptive reading, text data cannot be downloaded ("consumed") by a user and brought to the algorithms on the user's side for algorithmic distant reading for statistical analysis. Instead, non-consumptive reading can follow one of two strategies: the strategy of bringing the algorithms to the text, with the algorithms allowed to execute against the text in a secure, encapsulated environment in which the text is strictly isolated from the user [Fig. 1] (Zeng *et al.* 2014); and the "feature-extraction" strategy, in which only certain notable or informative characteristics extracted from the text, but not the text itself, are brought to the algorithm [Fig. 2]. While both these strategies are currently being explored at the HTRC, this paper will focus exclusively on the second strategy, namely feature-extraction.

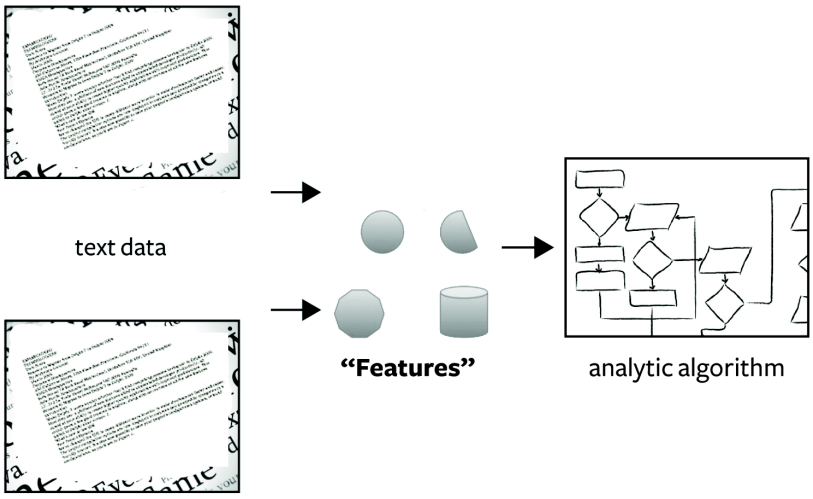

Fig. 2: Non-consumptive text analytics based on feature-extraction. Features (in the form of annotated fragments of potentially different kinds, such as: words (tagged by part-of-speech), with frequencies (per-page); particular types of characters and their frequencies (per-page), etc. (The various shapes symbolically representing the features are meant to indicate that many different kinds of features can potentially be provided.

The HTRC was created as a research arm of the HathiTrust consortium to consider tools for research for scholars who hope to analyze and interpret text at large scale (Unsworth 2011, Kowalczyk 2012, Kowalczyk *et al.* 2013). The services that the HTRC provides include, to date, support for scholar-created custom research collections ("worksets"), statistical text analysis tools and online interfaces to them, an application programming interface (API) for metadata, and a data API for public domain materials. The new feature-extraction service of the HTRC, now in alpha-release and presently operating in a pilot version on 250,000 volumes of text, provides per-page features packaged (in the JSON (JavaScript Object Notation) lightweight data interchange format) in one file per volume, with the relevant page-section (that is, header, footer and body) identified for each page [Fig. 3].[2] The per-page features currently provided are per-page "bags of words" with frequencies (that is, counts of part-of-speech tagged words per page-section) and some line-level information such as counts for the initial character and the final character of each line in a page section. We can think of these "features" as a translation of text from a language that humans understand to machine-readable *fragments* [Fig. 4].[3] For the purposes of this paper, we will focus on the particular feature consisting of per-page words and their counts.

[2] This pilot version is available via the following URL at the HathiTrust Research Center's sandbox portal, which is where new functionality is introduced on a smaller public domain subset of the HathiTrust corpus: https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/Features.

[3] Features can be defined on any domain of culture, including non-textual ones. The primitives of human motion ("movemes") that researchers seek to identify from video sequences are arguably features in this sense (Bregler 1997), and such motion features have been applied to dance (Shiratori *et al.* 2006); likewise, there are ongoing initiatives to extract musical features such as chords from musical scores (Raphael and Wang 2011). For text, of course, term occurrences (words) are a particularly obvious feature.

```
features: {
   "pagecount": 230,
   "pages": [
      {
        "tokenCount": 212,
        "lineCount": 38,
        "sentenceCount": 7,
        "header: {
             "tokenCount": 1,
             "lineCount": 1,
             "sentenceCount": 1,
             "tokens": {
                 "INTRODUCTION: { "NN": 1}
             }
             "beginLineChars": { "I": 1 },
             "endLineChars": { "N": 1 }
      },
```

```
"body": {
    "tokenCount": 211
    "lineCount": 37,
     "emptyLineCount": 10,
     "sentenceCount": 6,
     "tokens": {
         "priests": { "NNS":1 },
         "development": { "NN": 1 },
         "extraordinary": { "JJ": 1},
         "striking": { "JJ": 1 },
         "which": { "WDT": 3 },
         "sprang": { "VBD", 1 },
         ".": { ".":7 },
         ",": { ",": 10 }
     },
     "beginLineChars": {
         "f":2, "d":2, "b":2,...
     }
     "endLineChars": {
         "f":1, "g":2, "d":2, "r":1,...
     }
}
```


Fig. 3: Selected sections from the JSON package of features for a page of text

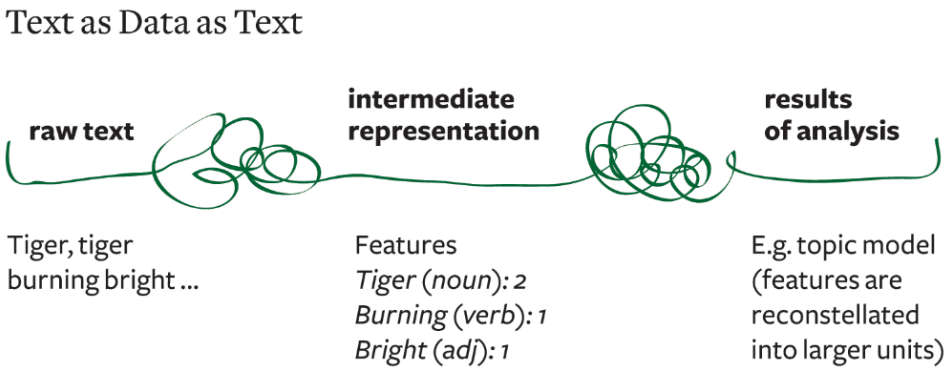

Fig. 4: Feature extraction as translation of text to fragments. We illustrate this translation process for the representative feature we use in this paper, namely the word (annotated with frequency on a per-page, part-of-speech-tagged, basis). Features are the intermediate representation of the text in the form of fragments, which can be reconstituted into human-readable, meaningful units by means of algorithms (such as topic modeling algorithms) that can analyse and reconstellate the fragments.

If we think of the act of reading a collection of texts in terms of the notion of architecture, we can conceive of the act of reading both in terms of a process view and in terms of a components-and-connectors view. These are two equivalent, but conceptually distinct, views in terms of which architecture can be conceptualized. A

process view consists of elements that are processes, and of relations between the elements that define process-related interaction (Bass *et al.* 2003, 208). While the process view of an architecture is useful for uncovering the interaction relationships among the architecture's components by revealing the path of data as it flows through the architecture, the interactions in any real system, such as the feature-extraction service which is our object of interest in this paper, will usually involve such a large number of components that expressing the architecture through a process view may lead to obscuring of the view by details (Fielding 2000, Ch. 5). Instead, looking at non-consumptive reading via feature-extraction in terms of the components-and-connectors view, that is, in terms of components and of interfaces, is more useful for our purposes. It provides an abstraction, through encapsulation, of the processes that are involved in feature-extraction. From the components-and-connectors perspective, it is possible to think of the HTRC's feature-extraction service as an interface producing fragments.

An interface is a metaphorical space. It is important for humanists to examine, when they use a space, what the nature of the space is and what it *means*. This holds for real spaces as well as for metaphorical spaces such as the interface between two conceptual entities. In the remainder of this paper, we will be concerned with how we can situate the fragment-producing interface for reading that is the HTRC's feature-extraction service within historical and philosophical contexts. Such fragmentizing interfaces are paradigmatic of all too many interfaces that mediate and fragment our experience in contemporary life and threaten to do so even more in the future. This in turn makes the question of the agency of such a fragment-producing interface an important one, and aligns that question with broad debates that are currently taking place about the hopes and fears that are being raised by the spectre of the digital — a spectre that, some fear, may herald a posthumanist future for the humanities.

The anxiety fragments and fragmentation tend to produce is hardly new. The trope of fragmentation has long been associated, in art and literature, with mourning for the inevitable disarray and decline to which all order is ultimately susceptible. John Donne, in his poem 'The Anatomy of the World,' published in 1611, speaks of how "new philosophy calls all in doubt," expressing the quintessentially modern anxiety

of the human having been displaced from the discursive center of the world after the scientific revolutions of the early modern era. The entropy of scattering and dispersion, as a metaphor for decay and thus for the shortness and transitory quality of life, takes centre stage in Donne's poem. Punning on the word "volume," Donne describes new volumes failing to retain all the old text that has been inherited from the past. He writes:

> ... mankind decays so soon,
> We'are scarce our fathers' shadows cast at noon,
> Only death adds t'our length: nor are we grown
> In stature to be men, till we are none.
> But this were light, did our less volume hold
> All the old text; or had we chang'd to gold
> Their silver; or dispos'd into less glass
> Spirits of virtue, which then scatter'd was.
> But 'tis not so; w'are not retir'd, but damp'd;
> And as our bodies, so our minds are cramp'd...
> We seem ambitious, God's whole work t'undo;
> Of nothing he made us, and we strive too
> To bring our selves to nothing back; and we
> Do what we can, to do't so soon as he.

There is a sense of foreboding in the poem suggesting that human experience becomes more "scatter'd", fragmentary and less satisfying over time, the containers for texts and spirits proving inadequate to their contents as the new age of uncertainty and disorder descends upon the world, paralleling the unraveling and decay of all complex order including that of the human body and mind.

Moving from the seventeenth century to the eighteenth, we discover that the imagery of architecture pervaded eighteenth-century thought and, not unexpectedly, eighteenth-century writers often referred to architectural ruins when they thought about the fragmentary (Harries 1994, 57). Reflections on ruins was often associated with a kind of elegiac humanism mourning the decline of a more desirable past order. Paul Fussell writes that, in the eighteenth century, "the ruins of Rome provide[d] the humanists with a powerful image of the kind of desolation inevitably wrought by innovation, novelty and wilful change" (Fussell 1965, 293). As for the nineteenth and twentieth centuries, the pervasive anxiety, from the industrial

revolution onwards, about the fragmentation of life is well-known enough not to require repetition: both European romanticism in the nineteenth century, and modernism and postmodernism in the long twentieth century extending into the twenty-first, can arguably be thought of as a sustained response to the experience of the fragmentation and destabilisation of a hitherto stable régime of relationality that had earlier been taken for granted: "All that is solid," in Marx's famous description, "melts into air" after capitalism shakes things up and dissolves "all fixed, fast-frozen relations" (Marx and Engels 1998, 38). When T.S. Eliot writes in *The Waste Land,* in 1916,

> Shall I at least set my lands in order?...
> These fragments I have shored against my ruins
> Why then Ile fit you. Hieronymo's mad againe

we seem to have returned to a crumbling, fragmenting and disoriented world not unlike that which Donne had described in the seventeenth century in 'The Anatomy of the World':

> When in the planets and the firmament
> They seek so many new; they see that this
> Is crumbled out again to his atomies.
> 'Tis all in pieces, all coherence gone,
> All just supply, and all relation...

This anxiety about the fragmentation of human experience produced by technology — and about its supposedly deleterious effect on our well-being — reaches a crescendo in our contemporary, digital, era, especially within the past decade. Journalists and commentators increasingly often associate this fragmentation in the context of practices of reading with reduced attention spans and with a lack of sustained, deep, or immersive reading, as it becomes increasingly easy to simply surf fragments of text and move on. In a recent article in the *New York Times*, for instance, Ravi Somaiya reports that we have entered "a world of fragments, filtered by code and delivered on demand," in which news is consumed not via the integral entity of the print edition of a newspaper, purveying many stories together on the physical printed page, but through "social media and search engines driven by an algorithm" (Somaiya 2014). Furthermore, the practices of reading that

are perceived as relational and immersive (and thus as a bulwark against fragmentation) tend to be associated by newspaper columnists and cultural commentators with the traditional humanistic disciplines — with literature, and with philosophy. For example, David Brooks pays nostalgic tribute to the practices of reading magazines in the twentieth century and, lamenting the new practices that have taken their place, which he finds to be fragmentary and inadequate, observes:

> During the 20th century… [e]ach magazine had its own personality, its own community of writers and readers and defined its own spot on the intellectual landscape.
>
> Today, the Internet has made magazine communities less cohesive. Most of those magazines still exist, but people surf through them fluidly and click on individual articles. Writers are identified more as individuals and less as members of a circle…Something important has been lost in this transition… [This] pragmatist mind-set itself [is] the mind-set of people who try to govern without philosophic or literary depth. (Brooks 2014)

Thus, technological changes (and the discursive practices associated with them) that seem to favor a fragmentation of the reading experience at a sociotechnical level are interpreted as a threat to "philosophic or literary" inquiry. Scholars in the humanities have also expressed concern with how digital scanning of texts, in general, dissociates the product from the process that underpins it, so that digital versions of texts have a propensity to be disseminated, as Bonnie Mak mentions, with scant regard for their "history of construction," so that they all too easily end up "deployed as data in the crafting of other narratives" (Mak 2013). It is to be quite expected, then, that some scholars in the humanities are likely to react with consternation at the idea of non-consumptive "reading" that we have described, which takes place across an interface that generates bags of words while obliterating all traces of the positional relations between words: "'Tis all in pieces, all coherence gone," indeed, to borrow from Donne's poem!

Non-consumptive reading is also likely to be especially anxiogenic to the traditional humanities for another reason: distant reading by itself does not forestall the possibility of supplementing distant reading with traditional close reading, but non-consumptive reading does not allow a space for close reading at all. The most

trenchant criticism of Franco Moretti's advocacy of a place for distant reading in literary theory has tended to come from critics like Christopher Prendergast, who claims that Moretti is "placing a very large bet on bringing the laws of nature and the laws of culture far closer than they are normally thought to be" (Prendergast 2005, 56). However, while Moretti has indeed provocatively called for distant reading (and the discovery, through such reading, of large-scale patterns in corpora similar to how scientists discover natural laws) to supplant closed reading, most literary scholars who discern distant reading to be valuable see it as a way to supplement, rather than supplant, close reading.[4] However, non-consumptive reading by means of the feature-extraction interface does away with even the very possibility of such supplementation of distant reading with close reading, and *only* distant reading is left on the table as a possibility. Here, one could of course object that non-consumptive reading does not rule out all forms of close reading in any *absolute* sense; although downloading and physically reading a specific book in question is prohibited by non-consumptive reading on account of copyright issues, there is after all no injunction against the reader going to a library and borrowing the book to read it closely, or against the reader simply buying the book and reading it. Nevertheless, in the general case, we must assume that the non-consumptive reader is not practically able, for any one of a number of logistical reasons, to simply read the book in this way. For example, the reader performing the non-consumptive reading over the Internet may be located in a country where the book may be impossible to obtain for close reading, for political or economic reasons. Thus, in principle, non-consumptive reading across a fragmentising interface does not allow for the possibility of supplementing or extending distant reading with close reading. Non-

[4] See, for example, Matthew Kirschenbaum's suggestion that "the existence of collections of millions of books in machine-readable form" will be "supplementing" rather than necessarily "replacing the libraries of individuals and institutions" (Kirschenbaum 2007); Mike Frangos's argument that explanation, having to do with the traces of social and material forces, as well as the "literary itself," which can be grasped, as abstract form, only through distant reading, has to be in a dialectical relationship with close reading of individual works in order to provide an adequately rich interpretation (Frangos 2013); and Moretti's own reply to Prendergast, in which he himself seems to acknowledge that interpretation (arguably, the task best achieved by close reading) and explanation (which distant reading is arguably capable of on its own) are intertwined (Moretti 2006, 81-83).

consumptive reading is, then, likely to incur even greater hostility from traditionalist scholars than does Moretti's proposal for distant reading, which after all, as we saw, does not, in and of itself, rule out the possibility of close reading.

It is certainly true that technological change does in fact create much fragmentation in our lives and our experiences of the world. Today's nostalgic cultural commentators, much like the Augustan intellectuals described by Fussell, can hardly be blamed if such fragmentation seems to them like "desolation… wrought by innovation", or if it makes them fear that the humanities are about to be desolated by the imminent advent of a posthumanist future. Of course, these fears are not totally groundless. However, posthumanism (if that is the name by which we describe the era that we are entering) offers the possibility of a neo-humanism at least as much as it does that of an anti-humanism. First of all, our relation to books may always have been fragmentary and discrete to begin with, rather than integrated and continuous. As Pierre Bayard writes,

> Our relation to books is not the continuous and homogeneous process that certain critics would have us imagine, nor the site of some transparent self-knowledge. Our relation to books is a shadowy space haunted by the ghosts of memory, and the real value of books lies in their ability to conjure these specters. (Bayard 2009, xix)

Whether or not we find this argument convincing, the so-called "ontological turn" in recent decades in continental philosophy provides us with another set of useful ideas to think about this issue in a new light. The notion of treating text as a collection of fragments out of which meaning will be made through a subsequent, algorithmic re-constellation of the fragments (as is the case with our fragment-generating interface), rather than conceiving of text as consisting of an *a priori* orderly sequence of syntactically well-formed and semantically meaningful sentences, can seem antihumanist in the traditional understanding of that term. However, the concept of a posthumanist performativity, introduced by the agential realist ontology of Karan Barad, can provide us with a way to approach this notion from a very different, and promising, direction. Barad's idea of the "agential cut" is a useful way to think of the agency of the interface (Barad 2003, 815). Once features are extracted, there is not much, other than the individual, separated words, that humans can actually "read" in

the traditional sense of the term until such time as the fragments are algorithmically re-constellated into meaningful wholes — such as into "topics" (discovered by a topic model[5]), or into clusters (generated by a clustering algorithm[6]), that are meaningful to a human reader. In this sense, the practice of reading inaugurated by the algorithmic means of partitioning a text into fragments followed by the reassembly of fragments into larger wholes (but wholes that are different from any previously existing chunks from the original text) merits being termed posthuman. However, such a posthumanism is not necessarily antihumanist. Barad's notion of the "agential cut" draws our attention to the fact that there is always an arbitrariness involved in how we make a "cut" in terms of deciding to what or to whom we ascribe agency. How do we draw a demarcating line (a "cut") around agency, and determine which algorithmic components, and which "human" subjective elements are encompassed within the cut? The answer may have the appearance of a consensus, but it is merely the result of an agreement reached, through performativity, by a community of practice and discourse — and is thus as arbitrary as any other. Barad writes:

> A crucial part of the performative account that I have proposed is a rethinking of the notions of discursive practices and material phenomena and the relationship between them. On an agential realist account, discursive practices are not human-based activities but rather specific material (re)configurings of the world through which local determinations of boundaries, properties, and meanings are differentially enacted… And performativity is not understood as iterative citationality ([as in] Butler) but rather iterative intra-activity. (Barad 2003, 828.)

Barad thus distinguishes her view of performativity from that of Judith Butler by characterizing the latter's view of performativity as citational: for Butler, performativity is the end product of linguistic or discursive acts, that is, it is ultimately a surface effect of human bodies. For Butler, performativity is "iterative *citability*": it is an ongoing (iterative) discursive practice that establishes signification for matter by establishing conventions through repetition (Barad 2003, 822). By contrast, Barad's notion of performativity is that of "iterative *intra-*

---

[5] Latent Dirichlet allocation (Blei *et al.* 2003) was the first widely used topic model, and remains the classic reference.

[6] A classic reference on the clustering of words in text is: Pereira *et al.* 1993.

activity": not a matter simply of ongoing discursive practice, but of ongoing *activity* of *all the actors* that are implicated within the entire technosocial assemblage that produces signification. In this conception, agency is distributed throughout the assemblage, and all the agencies involved in the sociotechnical system participate in the production of signification. Agency simply gets ascribed through the making of an "agential cut," that is, by carving out a subset of putative agents as the ascribed source of agency. Whom the cut encompasses (that is, to which part(s) of the sociotechnical assemblage agency gets ascribed) is thus the result of a relatively arbitrary decision, one based, typically, solely upon the rules of social convention as they have been fixed by repeated, iterative practice.

Turning to our fragment-generating (that is, feature-extracting) interface, we can now begin to see how, within the terms of Barad's ontology, the fragments produce meaning. The fragmentary features, in such a view, are not simply mutilated or deficient text, nor merely parts of a fallen or disassembled whole, of what the text once *used to be*. Rather, according to Barad's conception, the features — along with the technosocial processes that comprise the topic-modeling or clustering algorithms that run on them and re-constellate those fragmentary features into topics or clusters — could be thought of as participants in the production of meaning. Discursive subjectivities as well as intersubjectivities in the form of social determinations go into the selections that are made about the constellative algorithms and into the selection of their parameters' values, just as such subjectivities and intersubjectivities underlie the human act of interpretation. As Barad explains: "Meaning is not a property of individual words or groups of words. Meaning is neither intralinguistically conferred, nor extralinguistically referenced. Semantic contentfulness is not achieved through the thoughts and performances of individual agents but rather through particular discursive practices." (Barad 2003, 818). Conceived of in this way, the meaning-making achieved through the production of features by the feature-extraction service and the constellation of the features into larger units such as topics or clusters is a posthuman performativity.

The meaning-making apparatus here does, then, in fact allow for subjective difference and intersubjective variation (through such choices as the choice of

algorithm and through variability and parameter tuning in how the algorithms are executed). A space is thus opened for accommodating ambiguity and non-fixity — qualities that we tend to associate with humanism as traditionally understood. Viewed in this light, such a posthumanist practice of reading actually proves to be a species of neo-humanism rather than an anti-humanism — and this is why such a practice of reading ought to be welcomed, rather than avoided or feared, by humanists. This is not to say that the fear of an anti-humanist dystopia in which human agency is devalued and the quality of all human experience (including the phenomenological experience of reading) is tragically degraded is totally irrational. David Golumbia points out that, ever since Leibniz, one strand of computationalism within Western thought has habitually taken the form of a response to the distress caused by the troublesome unpredictability of ambiguity which is part of the human experience, and has sought a stabilization or fix to, precisely, excise this element of ambiguity by computational means (Golumbia 2009, 15). But as Rosi Braidotti has recently argued, posthumanism does not necessarily *have* to be an anti-humanism; rather, posthumanism can, instead, also contain within it elements of a return to humanism — that is, posthumanism can also be a neo-humanism, with "posthuman thinkers embrac[ing] creatively the challenge of our historicity without giving in to cognitive panic" (Braidotti 2013, 159). The "unitary vision of the humanist subject," Braidotti says, "cannot provide an effective antidote to the processes of fragmentation" that mark our era (184). In addition, when the locked-up digital text is freed up for interpretation (even if only by means of fragments), then the locked-up text, although fragmented by the interface, is rendered accessible, albeit in transformed fashion, to circuits of even traditional humanistic inquiry. It is also worth noting that this notion of reading as a form of reassembly/rewriting has a long history: Bethany Nowviskie, for example, has written about the algorithmically combinatorial tool, the *Ars Magna*, created by the thirteenth-century polymath scholar Ramon Llull, to serve as a mechanical aid to hermeneutics, that was generative, analytical and interpretive all at once (Nowviskie 2014, 140); on a similar note, Lisa Jardine and Anthony Grafton point out that Elizabethan great houses employed scholars who read and excerpted texts (yet another form of disassembly and reassembly) on their employers' behalf (Jardine & Grafton 1990, 35), and Ann Moss describes how, during the Renaissance, students were expected to create their

own “commonplace books” consisting of quotations gathered from their reading (Moss 1996). Services such as our fragment-generating interface for reading can perhaps become a possible source of the kind of metaphors that may underlie the alternative visions of humanistic inquiry, constituted by processes of fragmentation and reassembly across vast swathes of text assembled on the fly from the depths of the world’s great libraries’ collections on the basis of complex queries driven by careful search using both content and metadata. But such metaphors will not be completely new, as antecedents for them have, as we just saw, already existed in the past, created by the same human propensity to take knowledge apart and then to re-assemble based on complex criteria of play and inquiry: a combinatorial impulse simultaneously ludic and purposive.

In his play *The Trackers of Oxyrhynchus*, the British poet and playwright Tony Harrison depicts two British Egyptologists sifting for papyri containing fragments of lost classical manuscripts at Oxyrhynchus in Upper Egypt, one of the most important archaeological excavation sites in the world. Oxyrhynchus, indeed, is a historical, not fictional, place. As Egyptian society was governed bureaucratically under the Greeks and the Romans, and since Oxyrhynchus was an important provincial capital, the material at the Oxyrhynchus dumps included large quantities of papyri, which were eventually discovered by archaeologists in the twentieth century in the form of fragments. The reconstruction of pieces of text from these fragments continues to be an ongoing project today. In Harrison’s play, the discoveries made by Grenfell and Hunt (who were modeled by Harrison after two real-life nineteenth century Egyptologists with the same names) include mind-numbing papyri fragments and, eventually, pieces of a lost satyr play by Sophocles (Harrison 2004, 30):

> Grenfell gets so anxious to recover even scraps
> It’s brought the poor chap almost close to a collapse…
> He heard Apollo yammering for scraps and tatters
> Of some lost Sophoclean play called *The Tracking Satyrs*…

The trope of the quest to reconstruct a “lost original” recovered from numerous small, discrete fragments, which constitutes a defining idea for his play (but which is an idea that the play ultimately resists) represents, in fact, well-established methodological practice in classical scholarship. Jerome McGann describes the “lost

original" which such quests seek to recover as a *terminus ad quem* or limit point, which, strictly, cannot actually be reached, but can only be progressively approached or approximated with the help of heuristics (McGann 1983, 56). To idealize the attempt to reach the unitary *terminus ad quem* may be one way to be a neoclassicist in today's world, but there are other ways, too, of being one — and these latter may consist of embracing the fragmentary and *refusing* to assimilate fragments into an idealized unitary. Raphael Lyne argues that Harrison's play, with its fragmentary verses, in fact celebrates the fragmentary, and that Grenfell and Hunt, converted into rebellious satyrs, embody a rebellious voice that will not be assimilated (Lyne 2008, 136-137). The play itself resists any attempt at the smooth assimilation of the fragmentary classical heritage into the prevalent, traditional notion of the classical, and, in resisting the pressure to do so, it paradoxically points to a possible way to be a neo-classicist in our contemporary epoch. Likewise, embracing the fragmentising interface and the posthuman quality of the enterprise that results from it— even though they may appear to some anxious humanists as a sure invitation to be led down the primrose path of anti-humanism — may in fact paradoxically be one of the possible ways of being a neohumanist in our times.

In his essay 'What do you do with a million books?' Gregory Crane points out that it would be useful for a large digital library to make its constituent elements accessible at a very fine grain-size (Crane 2006). This suggestion may appear paradoxical as large scale of a collection and its fine granularity may seem, dimensionally speaking, to be the antitheses of each other. However, it is precisely the combination of largeness of scale with accessibility at a fine level of granularity that is going to be most useful to an "algorithmic" reader. The algorithmic reader, as Stephen Ramsay suggests, imagines the text as "radically transformed, reordered, disassembled and reassembled" (Ramsay 2011, 1). The larger is the scale of the collection, the deeper, usually, is the potential insight that can be obtained by reassembling fine-grained fragments through generative algorithms that seek to characterize certain properties of the corpus as a whole — topic models (which depend on the accessibility of individual words) being the paradigmatic example of such prediction-through-generation. In conclusion, the feature-extraction service of the HTRC can be thought of as a mechanism for producing the fine-grained fragmentary elements of text that

can be subsequently processed to generate insight about non-public-domain material in the collection that will otherwise not be available for analysis. Since the features are at the page-level, they can be used to train classifiers to distinguish between pages belonging to such broad genre categories as fiction, nonfiction and lyric poetry. This facilitates comparisons between genres, as in the case of Ted Underwood's work on the emergence of literary diction in different genres (Underwood 2012) and the tracing of the history of different genres, such as comparing the preponderance of first person narrative as opposed to third person narrative in novels over time. That there is considerable metadata from the bibliographic records accompanying the books in the HTDL has the advantage that one could carry out this kind of analysis for very specific genres, allowing for exploration of a range of arguments about the durability, stylistic coherence, differentiation and social stratification of genres that researchers have already started preparing to undertake (Underwood, 2014b). In addition, any statistical analysis that does not require relational information about the words on the page can be carried out with words-as-features, as the latter are equivalent to a bag-of-words per page with occurrence frequency information associated with each word in the bag. These proposed studies are all instances of humanistic inquiry, which the posthumanist modality of non-consumptive reading via features will make possible in the case of texts that would otherwise have remained inaccessible.

Acknowledgements:

This work was generously supported by the HathiTrust Research Center.
We gratefully acknowledge valuable suggestions for improvement that were made by Ted Underwood and Stan Ruecker in their comments on earlier drafts of this paper, as well as comments made by the two anonymous reviewers. Ted Underwood, Boris Capitanu and Loretta Auvil have been involved in all aspects of the design and implementation of the HTRC feature-extraction service.

References:

Barad, Karen. 2003. Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs* 28, no. 3: 801-31.

Bass, Len, Paul Clements, and Rick Kazman. 2003. *Software architecture in practice*. Boston: Addison-Wesley Professional.

Bayard, Pierre. 2009. *How to talk about books you haven't read*. New York: Bloomsbury Publishing.

Blei, David M., Andrew. Y. Ng, and Michael. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of machine learning research* 3, March: 993–1022.

Bregler, Christoph. 1997. Learning and recognizing human dynamics in video sequences. Paper presented at the IEEE conference on Computer Vision and Pattern Recognition (CVPR '97), 17-19 June, in San Juan, Puerto Rico.

Braidotti, Rosi. 2013. *The posthuman*. Cambridge: Polity Press.

Brooks, David. 2014. The problem with pragmatism. *The New York Times*, October 3, Opinions section, International edition.

Crane, Gregory. 2006. What do you do with a million books? *D-Lib Magazine* 12, no. 3, March. http://www.dlib.org/dlib/march06/crane/03crane.html (31/10/14).

Fielding, Roy T. 2000. *Architectural styles and the design of network-based software architectures*. PhD diss., University of California, Irvine. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm (10/9/14).

Frangos, Mike. 2013. The end of literature: Machine reading and Amitav Ghosh's *The Calcutta Chromosome*. *Digital humanities quarterly* 7, no. 1. http://www.digitalhumanities.org/dhq/vol/7/1/000152/000152.html (29/10/14).

Fussell, Paul. 1965. *The rhetorical world of Augustan humanism: Ethics and imagery from Swift to Burke*. Oxford, U.K.: Clarendon Press.

Golumbia, David. 2009. *The cultural logic of computation*. Cambridge, Mass.: Harvard University Press.

Harries, Elizabeth W. 1994. *The unfinished manner: Essays on the fragment in the later eighteenth century*. Charlottesville: The University of Virginia Press.

Harrison, Tony. 2004. The trackers of Oxyrhynchus: The Delphi text. In *Tony Harrison: Plays, Vol. 5,* London: Faber and Faber.

Jardine, Lisa and Anthony Grafton. 1990. "Studied for Action": How Gabriel Harvey read his Livy. *Past & Present*, no. 129, Nov 1990: 30-78.

Kirschenbaum, Matthew G. 2007. The remaking of reading: Data mining and the digital humanities. Paper presented at the National Science Foundation (NSF) Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM '07*),* 10-12 October, in Baltimore, Maryland.

http://www.csee.umbc.edu/~hillol/NGDM07/abstracts/talks/MKirschenbaum.pdf (4/10/14).

Kowalczyk, Stacy T. 2012. Digital humanities at scale: HathiTrust Research Center. Paper presented at the Digital Library Federation (DLF) Forum, 4–5 November, in Denver, Colorado. http://d2i.indiana.edu/sites/default/files/dlf_fall2012_htrc_stk.pdf (24/11/14).

Kowalczyk, Stacy T., Yiming Sun, Zong Peng, Beth Plale, Aaron Todd, Loretta Auvil, Craig Willis, Jiaan Zeng, Milinda Pathirage, Samitha Liyanage, Guangchen Ruan, and J. Stephen Downie. 2013. Big data at scale for digital humanities: An architecture for the HathiTrust Research Center. In *Big data management, technologies, and applications*, ed. Wen-Chen Hu and Naima Kaabouch, 270-294. Hershey, Penn.: IGI Global.

Marx, Karl and Friedrich Engels. 1998. *The communist manifesto: A modern edition,* ed. Eric Hobsbawm. London: Verso.

Lyne, Raphael. 2008. Neoclassicisms. In *Tragedy in transition*, ed. Sarah A. Brown and Catherine Silverstone: 123-140. Hoboken, N.J.: Wiley-Blackwell.

Mak, Bonnie. Archaeology of a digitization. *Journal of the American Society for Information Science and Technology*. 65, no. 8 (2014): 1515-1526.

McGann, Jerome J. 1983. *A critique of modern textual criticism*. Chicago: The University of Chicago Press.

Moretti, Franco. 2006. The end of the beginning: A reply to Christopher Prendergast. *New Left Review* 41, Sep–Oct: 71–86.

Moretti, Franco. 2013. Conjectures on world literature. In *Distant reading*, 43-62. London, U.K.: Verso.

Moss, Ann. 1996. *Printed commonplace books and the structuring of renaissance thought*. Oxford, U.K.: Clarendon Press.

Nowviskie, Bethany. 2014. Ludic algorithms. In *Pastplay: Teaching and learning history with technology*, ed. Kevin Kee, 139-171. Ann Arbor, Michigan: University of Michigan Press.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. Paper presented at the 31st annual meeting of the Association for Computational Linguistics (ACL), 22-26 June, in Columbus, Ohio. *Proceedings*: 183–190. http://www.cs.cornell.edu/home/llee/papers/ptl93.pdf (11/24/2014).

Prendergast, Christopher. 2005. Evolution and literary history: A response to Franco Moretti. *New Left Review* 34, July-August: 40-62.

Ramsay, Stephen. 2011. *Reading machines: Toward an algorithmic criticism*. Urbana: University of Illinois Press.

Raphael, Christopher and Jingya Wang. 2011. New approaches to optical music recognition. Paper presented at the 12th International Society for Music Information Retrieval conference (ISMIR 2011). http://ismir2011.ismir.net/papers/OS3-3.pdf (31/10/14).

Shiratori, T., Nakazawa, A. and Ikeuchi, K. 2006. Synthesizing dance performance using musical and motion features. Paper presented at the IEEE International Conference on Robotics and Automation (ICRA 2006), May.

Somaiya, Ravi. 2014. How Facebook is changing the way its users consume journalism. *The New York Times*, October 26, Media section, New York edition.

Underwood, Ted and Jordan Sellers. 2012. The emergence of literary diction. *Journal of digital humanities* 1, no. 2, Spring. http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/ (31/10/14).

Underwood, Ted. 2014a. Theorizing research practices we forgot to theorize twenty years ago. *Representations* 127, no. 1. Summer: 64-72.

Underwood, Ted. 2014b. Personal communication.

Unsworth, John. 2011. Computational work with very large text collections. *Journal of the text coding initiative* 1, no. 1, June: *Selected Papers from the 2008 and 2009 TEI conferences*. http://jtei.revues.org/215 (31/10/14).

Zeng, Jiaan, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptive use of texts. Paper presented at the 5th workshop on Scientific Cloud Computing (ScienceCloud), 23 June, in Vancouver, Canada.