Reflections on Distant Reading

Pia Bhatia

2:52 PM I found Underwood's A Genealogy of Distant Reading to be a fascinating essay that gave me a lot of insight into how terms are being used in the field today — the author discussed various distinctions, such as the term digital humanist versus distant reader, and buzzwords such as big data that have developed over recent years. What was most compelling about the piece to me was the idea that large-scale literary history has an established, pre-DH past whose ambitions are now being enabled by the field's new developments. In addition, Underwood proposed the idea that literary history DH projects need to be approached at various intersections — linguistics, the social sciences, etc. When discussing the Radway Reading the Romance project, the author points out how certain gendered binaries are constructed in these novels and uses this distinction in her statistical analysis. I wonder how versions of this line of thinking might be used today — would distant reading softwares be able to identify narrative arcs or overarching themes, for example? That said, it's written that these "binary oppositions" supposedly turned various archetypes into a "symmetrical structure," which sounds like a generalized and possibly problematic view of the books to begin with.

Anya Kalogerakos

3:00 PM I enjoyed the context that Underwood brought to the concept of distant reading, and it made me wonder about other practices that are considered under the "digital humanities" umbrella that have a longer (less digital) history than we may realize. Previously, in this class, I have been viewing the introduction of the computer to the field of the humanities as an incredibly disruptive innovation. However, after reading this article, I realized that the methodologies that the digital humanities use have been around for a much longer period, making the introduction of the computer less of a disrupter, and more of a smooth continuation. That being said, I think there are ways to see the introduction of computing as both a very disruptive innovation to the field of the humanities and a regular continuation of the field. Tahmasebi's presentation covered a lot of ground on the actual methods of data science research in a humanities field and reminded me of earlier conversations we have had about the impact of data cleaning and selection of data. While it is easy to assume that the fact that some data must be left out to make certain conclusions weakens the conclusions themselves, I think it is important to remember that all fields in some way use this practice, and it is not a critique of only the digital humanities. Instead, it is a struggle that all research faces.

Raphaela Gold

3:44 PM I was most interested in exploring the role of historical context in Underwood's "A Genealogy of Distant Reading". I've observed how in general, scholars and individuals who see digital humanities as a newly emerging trend tend to have more negative views of the field (this is, of course, a generalization). The words "devolve" and "manipulate" in the Earheart quote from Underwood's essay, for example, both carry negative connotations. On the other hand, when people see digitization, distant reading, and big data as part of a much longer and richer history of the intersection of humanities with technology, they tend to view the field as less harmful.

Underwood noted that while digital humanities and distant reading might be very compatible and complement one another, it is important to note that they have not always coexisted, and they should not be

conflated. His deep dive into distant reading was fascinating to me. I would have loved it if he had included more on Marxist literary theory and how literary works reflect the social institutions from which they originate, but understand that it may have distracted from the main point of the essay. I thought that Radway's value of simplicity was an important one to keep in mind as we delve into our own data cleaning and analysis. I also thought it was very interesting that Underwood approached literature from the perspective of the social sciences and promoted an almost scientific method to literary research, in which scholars first draw hypotheses, then methodically test various theories in order to reach a conclusion.

Additionally, hearing from Professor Tahmasebi in the video was really fascinating both because her research was interesting and because it was an exciting new medium through which to learn about digital humanity. Professor Tahmasebi's excitement about her own research and the ease with which she explained it made it a lot easier to digest. I was struck by the intersections she highlighted between data and language, as well as the distinctions she drew between them and illustrated with her circle graphic. I also really appreciated her reminder that humans drive research, not data. I think it is very important to keep that in mind as we delve more into our own datasets. By not losing sight of the fact that human beings drive datasets, we can keep in mind that data cleaning always involves some choice made by a biased human, which can inform and nuance the way we interact with a dataset. (edited)

Talia Goldman

9:53 PM Both Underwood's A Genealogy of Distant Reading and Tahmasebi's talk "The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies" shifted my thinking about experimentation in DH. In her conclusion, Understood writes that "we need…to rediscover the guiding principle of experiment" (Underwood, Paragraph 43). In looking to historical precedent, Underwood demonstrates the experiment, bringing in scientific and sociological methods, is not foreign to the humanities, an argument that fits in nicely with Tahmasabi's more detailed explanation of what quantitative experimental methods can look like in DH. These materials had me thinking more deeply about how experimental or scientific research methods are sort of ingrained in humanities thinking, even without technology (as Underwood argues). For example, analyzing art can be similarly "experimental." Perhaps the process of text mining is similar to the selection of what specific elements of a work of art merit further exploration. Or, choosing a method or theoretical framework for analyzing a work can be seen as choosing the research methodology for a text mining project, which Tahmasebi describes with her transportation metaphor. Out of these choices come a research question. While it was especially interesting to frame these ideas in the specificities of text mining (and the ideas of hypothesis and results), I found myself breaking down boundaries in my mind of "humanistic" and "scientific" thinking.

Pippa LaMacchia

11:56 AM The most impactful part of both Underwood's "A Genealogy of Distant Reading" and Tahmasebi's lecture "The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies" are the ways in which my understanding of Digital Humanities and my judgment of the field is changing. As we dive further into the academic and intellectual implications of digitization and the role that technology is playing in traditionally humanitarian fields, I am becoming more accepting of its universality. Underwood explains the overlap and differences between digital humanities and distant reading, pointing out that "literary scholarship turned out to have a blind spot" and enables us to understand that the value of computation lies in the new questions it allows us to ask. Throughout the course thus far, I have been viewing digitization as a threat in some regards but this article is helping me to realize the ways in which technology in fact adds to scholarship. It is

still important to ask questions about whether or not digitization is disruptive and to address scholarly fear about the “fusion” of the fields, and Underwood addresses these concerns with a fascinating lens. In Tehmesebi’s lecture, I am interested in this continued investigation into the overlap of scientific methods and humanistic research. Digitization allows us to use these two facets simultaneously and it is important to better understand new forms of experimental methods as these fields come together. Finally, I was particularly struck by the clarified idea that “humans drive research, data does not drive research” because it reinforces scholarly autonomy in a humanities field that transitions to a more digital realm.

Andrew Huo

1:40 PM Nina Tahmasabi’s presentation, “The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies,” was very informative and eye-opening for me. She explained each concept slowly and clearly, and I was hooked from the beginning. Firstly, the difference between how data scientists and digital humanists see text (data vs. representation of language) was very interesting. She then clearly laid out the steps for data mining as going from concrete to abstract. But what I found the most fascinating was filtering out data and deciding what to keep and throw away. Tahmasabi gave examples of ignoring formatting such as white spaces, white pages, fonts, and capitalization of letters. Still, my thought was, would it be more efficient to collect every single piece of text information and then begin filtering out rather than starting out by ignoring information? My next question was, if you happen to have a different research question with the same texts, do you need to run the whole data mining process again? Another analogy that made sense to me was the castle that, through data mining, represents a kind of 2D picture but ignores the ‘3D’ provided by the filler words, context, and data that was filtered out. Since humanities as a subject and its research questions (brought up by one of our previous readings) are built around the context of language, history, what came before, etc., does data mining limit the possibilities for humanities studies?

Alison Fortenberry

6:13 PM The Underwood reading was really interesting, and I really appreciated the argument that distant reading is not a novel concept, even though it tends to be thought of as a technological reading beginning in the 21st century. It reminded me of the tensions we’ve been talking about all semester of trying to incorporate the digital into the humanities without losing what the humanities have meant from their core for a long time. I think the reframing that Underwood did of distant reading as a continued tool with a modern, technological equivalent is a helpful response to that problem. While he is clear that distant reading is not DH, I wonder if this sort of argument could be used to justify DH to humanists who may be doubtful about its applications. Reframing DH as a continuation of methods we’ve always used that is just broadened now with the advent of new technologies. I also really enjoyed the Tahmasebi lecture, especially the discussion of how to decide what is cut and what is included in word mining. While setting specific parameters to match one’s research question seems like a clear and logical thing to do, I found the idea that this could introduce a layer of selection bias into the process really interesting. How do we narrow the scope of massive projects like these while making sure we’re not cutting things that should be there to fairly contextualize our projects?

Clay Glovier

10:27 PM I enjoyed watching Tahmasebi’s lecture about “The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies.” It was really interesting to learn about data mining and the process of filtering out certain data points and keeping others. I am curious how accountability can be maintained so that data points that are important to some and not others are not lost from the conversation if a certain researcher

decides they are not worthwhile? I also enjoyed reading Underwood's A Genealogy of Distant Reading. I particularly liked the segment about the power of readers of novels vs reviewers. While critics may assume they have immense power over the trajectory of texts, this is not always the case, as readers may interpret plots very differently and take away different things from novels. When seeking to truly understand the impacts of texts, it does seem important to separate the questions one poses from the evidence they gather to address it and the conclusion they finally draw. Surveys can be a useful tool to understand how viewers interpret a text, but these are often time-consuming and can only be conducted with living studies. After weighing these benefits and drawbacks, it does make sense that modern distant reading as outlined in the text seems to be more concerned with textual analysis. (edited)

Ethan Haque

5:06 PM I thought the Underwoord reading brought up some interesting case studies like the "What are the Three Most Important Ingredients in a Romance" chart from Radway and Moretti's clue trees. I was expecting them to mention the early approaches of text mining that people like Claude Shannon used to model languages. Maybe I only have this expectation given how much LLM's have been on our minds recently, but I think it was a missed opportunity. How did people know about the frequency of words and letters in the English language before computers were able to process text like they are able to today? By manual inspection and some very cool sampling methodologies. Knowing this information helped researchers answer questions like how have books and language changed over time and how do they relate to the state of the world at the time. The Tahmasebi talk was interesting as well. There's a particular visualization that she uses which I liked a lot. Around 35 minutes in she uses a picture of the Eiffel Tower with only small patches visible to to about how when we summarize information, our goal is to distill it into something that captures as much of the underlying text as possible while presenting it in the most compact form. She uses two different sets of patches where one of them helps us glean that what we're looking at really is the Eiffel Tower.

Layla Williams 5:59 PM I will begin with a short reflection about my experience transcribing my assigned postcards. I thought the process of transcribing was interesting with a piece of work like the postcards; I actually thought it was easier to transcribe the pages from Shakespeare and Company because of the structure of the document. It had a typical header, body, and sign-off. In contrast, the postcards had a variety of different structures. Some of the postcards only wrote in the designated body area, but there was one that wrote all in the margins of the postcard, making it difficult to determine the baselines and create the transcription. This one experience took me more time than expected, so I can imagine what it would be like to transcribe a whole collection. This leads me into the readings for today's class, specifically about a "distant reader." I would like to know about what distant reading looks like in practice. I know something that came from the video presentation was the relationship between the amount of information and the usefulness of information, and I wonder how that plays into the practice of distant reading. Is it helpful to see long trends and large amounts of useless information, or is a singular piece of really helpful information more helpful to a larger research project?

Colin Brown

8:35 PM This week's readings felt very familiar in that they revolved around a very scientific approach to digital humanists studies. Tahmasebi structures her research around a driving research question, and Underwood views distant reading as highly empirical and scientific. Moreover, both readings sent messages about computational work that I have also heard in more engineering spaces. For instance, Tahmasebi

extensively discusses the impact that one’s data preprocessing will have on the results they find, but often times the details of this preprocessing are not completely shared in the literature; indeed, I have heard from a professor who studies computational molecular models in our Chemical and Biological Engineering department that he frequently finds little information on how other researchers handle their data, which makes reproducing results a tricky task. These kinds of overlaps cast distant reading as perhaps the most scientific field of DH that we have studied so far. I noticed an interesting line in Underwood that said literary scholars are quick to adapt social scientist’s findings but not their methods. From this I wondered if humanists ever consider themselves as following a somewhat scientific approach? Is deep literary analysis also motivated by a research question and then backed up or disapproved by collected data, or do scholars view their process as entirely unique? I’d be curious to hear and learn about how this shapes their approach to research.

Melissa Woo

9:52 PM In engaging with Tahmasebi and Underwood's works, the intricate frontier between traditional literary analysis and computational methods in digital humanities becomes increasingly apparent. Tahmasebi's examination of large-scale text mining not only sheds light on the opportunities and challenges within this domain but also underscores the necessity for interdisciplinary cooperation and a nuanced comprehension of computational literary studies. On the other hand, Underwood's historical exploration of distant reading offers a fascinating glimpse into the evolution of literary analysis, showcasing how technology has reshaped scholarly research practices over time. I wonder how the fusion of technology and traditional scholarship will continue to influence the way we approach literary analysis today, especially as technology (particularly as it regards AI) continues to evolve and develop so quickly. How do scholars navigate the complexities of integrating computational tools while preserving the humanistic essence of their research? What implications does this integration have for the future of literary studies, and how can researchers effectively balance empirical data-driven approaches with nuanced human interpretation in their work?

James Sowerby

11:30 PM I thought Tahmasebi was a great lecturer, and really helped me visualize some of the ideas that we have been talking about. Just the fact that she had diagrams and visual allusions to her arguments was a great way to further convey the nuance around the research methodology. Others have mentioned it but the idea that there are better and worse datapoints to base an argument off of makes sense on its own, but to combine it with the image of the Eiffel Tower captured exactly what she meant. Certain holes in the picture were ambiguous, but others made it more obvious that it was of Paris. I also liked other examples she had, like the relative word frequency in Pride and Prejudice, although that is something has popped up in our articles before. Underwood's piece was also very relevant to my weekend's work—I mentioned this in a comment in Perusall but I was just working on a close reading essay for my Comparative Literature class so it was very cool to have the difference to "distant reading" explained. Although I found it a bit hard to follow the references to other articles, it was worthwhile to see the contrast between digital humanities, which is naturally implanted in my brain aas the paradigm name for the field, and distant reading. I have even seen in our work as a class so far that not all of it is necessarily digital, like in our discussions of a plain book as a technology in and of itself. Underwood's point—that this confusion may cause unnecessary arguments between people in the field—seemed to be pretty small, but admittedly nuanced. I think it's hard for me to completely understand the strife that intradisciplinary argument can cause as someone who hasn't really inhabited academia yet.

Yaashree Himatsingka

12:30 AM Tahmasebi suggests that reading like a human(ist) – I wonder if reading like a humanist is effectively the same as reading like a human – involves drawing meaning from texts in an immersive and subjective manner (similar to understanding from 'within' which we had encountered earlier in the course). This humanistic approach to reading is distinct from 'text mining', which I have understood (correctly?) as the mechanical extraction of 'relevant' or metricized data from a vast repository of texts. Both approaches have obvious (and less obvious) merits and demerits, which is why I really appreciated Tahmasebi's emphasis on bringing in different sets of expertise to understand and interpret source texts. So often scholars seem to privilege one disciplinary toolkit over another or debate different methodologies, so it's refreshing to think about combining various paradigms to expand, enrich, and complicate one's understanding of a research area or question. In light of this call to interdisciplinary collaboration, Underwood's history of distant reading really resonated – I pictured it like a Venn diagram unfurling more and more overlapping circles with the passage of time. This underscores Tahmasebi's argument that integrating diverse methodologies can lead to a more complex and comprehensive understanding of texts. But I guess the disclaimer is that not all methodological mixing is apt or particularly illuminating, so it's important to adapt and implement different 'windows' onto texts with care and intention in order to ensure a fruitful interdisciplinary approach. New

Helen Gao

12:54 AM I appreciated that "A Genealogy of Distant Reading" discussed the history of literary studies, and I was surprised to learn that computational methods did not impact the field of literature as much as I thought they would have. I also found the findings from Reading the Romance that were discussed in the article to be surprising, as they contradicted my own (mis?)conceptions about the effects of stereotypes in books. It was also interesting to read about the quantitative methods that Radway used in what is usually considered a qualitative field, though I had some doubts about the methodology, given societal attitudes regarding the romance genre (and readers of the genre). I liked how "The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies" went through the steps of analyzing text to answer research questions. Tahmasebi addressed many of the limitations of data intensive research, as well as how to evaluate it, and several of the points she mentioned were consistent with previous articles (or even our personal experiences from class – for example, the issues of OCR on older texts). I also liked that she mentioned a common NLP pipeline; it felt very relevant to modern computational practices, and helped put previous topics like data cleaning into a new perspective.

Emanuelle Sippy

1:38 AM Rather than distant reading and digital humanities being comprised of one or the other, both Underwood and Tahmasebi speak to the fact that linguistics work in conjunction with social science. Underwood writes: "Linguistic categories are just as important as the social categories" (Underwood 5). He, interestingly, notes literary scholars' readiness to draw on "social scientists' conclusions," while they are hesitant to "borrow their methods" (Underwood 5). I think the lack of training for literary scholars in the social sciences, as Underwood later gets into, helps to explain this, as well as the significant differences in approach and the way these differences are born out of and mapped onto "normative" ideas about the value of these methods. I appreciated Underwood's point about the understanding of distant reading as a normative counter to close reading being a misinterpretation (Underwood 8). I was also struck by Tahmasebi's ability to provide a streamlined version of distant reading processes, wherein the researcher

approaches the dataset with a hypothesis. Although it is not always the case, I think many scholars who employ close reading attempt to (and sometimes claim to) evaluate the text in its own right (i.e., if there is a hypothesis it comes out of the text rather than evaluating texts with hypotheses from the outset). This also seems like a semantic difference in large part; even when literary scholars do employ a hypothesis, they rarely use the language of a hypothesis when describing their research methods. Overall, I think Underwood is right to suggest that while distant reading has a distinct “genealogy” and should not be collapsed into DH, these fields are intertwined and deserve both critical and capacious thought, as he quotes Moretti, writing: “Since no one knows what knowledge will mean in literary studies ten years from now, our best chance lies in the radical diversity of intellectual positions, and in their completely candid, outspoken competition [Moretti 2000b, 227]” Underwood 7.