Reflections on Topic Modeling, part 1

Readings

- Froehlich, Heather. *Moby Dick Is About Whales, or Why Should We Count Words?*.
- Blei, David M. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities*, vol. 2, no. 1, 2012.
- Nelson, Robert K. *Mining the Dispatch*.

Raphaela Gold

11:18 AM (For Tuesday the 19th) I really appreciated Froehlick's examination of the typical method of using word-frequency to extrapolate important themes and topics in a book, and her assessment that there is "nothing terribly groundbreaking" in discovering that certain words are central to Moby Dick. Her assessment that this model is so popular because "both humans and computers can handle the saturation of these words" really stuck with me. In our exploration of digital humanities so far, I think we tend to focus on what computers can do that is different from humans, but the "Moby Dick is About Whales" models centers around similarities rather than differences; humans and computers both enjoy rules, logic, algorithms. The contrast between function words and content words was also very interesting – I wonder if there are some points of overlap between these words, and ow a computer model might deal with that similarity when that is the case. Although according to Zipf's law content words tend to be lower-frequency than function words, we tend to pay a lot more attention to the content words due to their variation and meaning in the book. According to Froehlich, the words we find the most meaningful are high-saliency content words like "whales" in Moby Dick, occurring frequently enough that we notice them but not so frequently that they blend into the background and become mere function words.

Seeing this model as the "start of a conversation" rather than the ultimate destination made a lot of sense to me, and I believe that this could be applied to many different aspects of the digital humanities. For example, when we've counted word frequency and explored datasets, they often provide us with valuable but somewhat shallow information. I have come to associate this quality with distant reading, whereas closer reading allows people to access the deeper themes of a novel (i.e. isolation, homosociality, and self-discovery). One can certainly be aware that those themes exist in the novel simply by being told, but the understanding of those themes and how they play out over the course of the novel, in all their intricacies and particularities, can really only come from reading the book. This is why I found Froehlich's point toward the end of her paper that, "computers are good at finding patterns and people are good at interpreting patterns," very resonant. While frequency-mapping can further enhance understanding and reveal similarities between humans and computers, it is important to maintain awareness of their differences as well – I very much agree with Froehlich's point that quantification is not everything, and I do think some of the datasets we've observed have managed to somewhat push past mere quantification, although that is still what they center around.

Blei's "Topic Modeling and Digital Humanities" also picked up on the algorithmic nature of computers and their tendency to quantify. The simplicity of "LDA" seems to be exactly what Froehlich was cautioning against, as it makes the immediate assumption that word/topic frequency can be taken as an indicator of what a book is about. That being said, LDA's ability to reveal hidden topical structures within text through forms of distant reading is a very appealing tool. Identifying the process as a conceptual one is also helpful,

as it suggests that probabilistic modeling is meant to show us possibilities of what could be, not necessarily of what is. As Blei put it, probabilistic modeling is not meant to generate its own literary interpretation, but rather to “give scholars a powerful language to articulate assumptions about their data and fast algorithms to compute with those assumptions on large archives.” While I’m still not entirely sold on fast being equivalent to better, I do understand Blei’s point about probabilistic modeling as a tool toward more efficient interpretation, rather than being the interpretation in and of itself.

Exploring the “Mining the Dispatch” database was a really helpful way to observe these theories in action and see how probabilistic modeling can truly benefit the humanities. While the database was of course very quantitative, it was interesting to read the text that went along with it and note the compiler’s admission that they had excluded three topics which struck them as “less than substantive.” To me, this choice infused the database with the humanistic quality of interpretation, allowing me a glimpse of the “rosy vision” Blei had laid out about how these tools might be put to proper use in the humanities. (edited)

James Sowerby

6:46 PM I really enjoyed the way that all three of these week's sources added something different to the topic of word frequency and thematic modeling. In particular (like Raphi), I thought that Froelich's article had a lot of important nuanced concessions—the most common words in Moby Dick don't necessarily reveal anything meaningful about a text and mentioning "sea" "ship" "captain" or "whale," for example, only construct a C-average reading of a text. Thus Zipf's law's visualization is important to understand: the most frequent words generally only have deictic meaning and have no interpretable content on their own. Of course, that leads to a common point made in the readings for the course generally. Humanists have to work together with data and programs in order to move beyond facile conclusions. Context is everything for the deictic words like "the" "a" "you" etc, but I would also argue for every word. One of the most rich parts of textual analysis lies in understanding polysemy, or changing meanings or use of words throughout a text. Froehlich showed this to a degree. "New" is used often with other nouns in order to construct place names, but she found an interesting parallel when put in conjunction with the use of the word "old" that shows a depth of meaning in the phrase itself that can only be shown through quantitative analysis. The formal definition of a topic in Blei's article is also interesting. If a topic is a probability distribution over terms, maybe a humanist can tell the most rewarding places to analyze a text based on how unusually written or formulated a passage is—which a humanist can tell is still about the same topic whereas a computer might not.

Andrew Huo

2:51 PM Before reading any of the articles/websites, I was a bit skeptical about the application of topic modeling, but immediately when I began to read Froehlich’s “Moby Dick is About Whales, or Why Should We Count Words?” I was enticed. Firstly, I found Zipf’s law interesting as a visual representation of the inverse relationship between the frequency of words and their rank in the frequency table and, moreover, the distribution of function words compared to content-driven words. His description of the importance of words that are “high-frequency enough to be noticeable to a linear reader, but low-frequency enough to be content-driven” was fascinating to me. But I am wondering what if you take out the function words? How would that affect Zipf’s law and the data collection process? Another aspect that interested me was how topic modeling can answer questions about the context of synonymous and antonymous words. For example, Froehlich’s examples discuss the contextual differences between ship/boat, old/new, and ocean/sea. In a sense, topic modeling seems like pattern recognition in a large text, which I realize can

answer many important research questions about that text. Blei's explanation of topic modeling in digital humanities was very helpful in summarizing this subject and making it clear to me – someone who has very little experience in this field. Finally, Nelson's case study of "Mining the Dispatch" was a very clear and thorough application of this process that I enjoyed digging into.

Melissa Woo

8:30 AM Froehlich writes, "the interpretive argument that Moby Dick is About Whales is pretty dependent on words that are high-enough frequency to be noticeable to a linear reader, but low-frequency enough to be content-driven." The part of this statement that really resonates with me is the relativity of arguments, and the significance of measuring the appearance of words in relation to each other, not just by magnitude in isolation. This method requires a lot of contextual analysis and consideration of how the words are phrased and used in the text itself. At first, I thought of it as somewhat of a shortcut or a more fully quantitative method, but it seems like it must be combined with more careful and nuanced analysis to gain more insightful conclusions. Then, perhaps we can infer that an obvious finding from topic modeling is not indicative of the lack of usefulness of topic modeling, but rather a need for a change in methodology to dive deeper into the analysis. I wonder if there is a formulaic approach to achieving this or if it requires critical thought on a case-by-case basis. It seems that expertise and familiarity with the text, author, and topics discussed in the book could be key, therefore rendering more traditional humanist scholars not obsolete but rather key to this development. I thought the technical explanations from the Blei reading were very interesting – but perhaps not necessarily accessible or particularly accessible to readers with less mathematical or probabilistic background. I wonder if there is a way to make such details more intuitive, perhaps by examples or practical case studies even though the argument is mostly theoretical (like on the Mining the Dispatch interface". I liked the quote: "the statistical models are meant to help interpret and understand texts; it is still the scholar's job to do the actual interpreting and understanding." To a certain extent, I think that this is true for both textual and other types of analysis.

Colin Brown

12:40 PM Reading about topic modeling really brought to the forefront how computational methods can complement and enable the more traditional close reading methods. Each reading here not only described their computational ideas in detail, but then they made a point to assert that the output of these models is only just the beginning of the analysis for scholars. Researchers must use these trends as starting points to start digging deeper. In this perspective, I feel like classical literary scholars should embrace the opportunity to push their research to new areas and acknowledge that this does not diminish the close-reading insights that they are specially trained to find. These articles were also interesting to read in light of currently being in an intro to machine learning class, as I started seeing shades of subjects from that class show up in the descriptions of topic modeling. Topic modeling seems to be a blend of natural language processing, probabilistic modeling, and k-means clustering, essentially creating a form of unsupervised learning specifically tuned for humanities research. Moreover, Nelson also mentions that his Dispatch model received an update in 2020 that reduces the topic diversity in each article with a certain hyperparameter, and this is likely a form of regularization, a common method in machine learning to reduce overfitting. I enjoyed seeing the concepts from that class be applied in this unique setting.

Pippa LaMacchia

5:39 PM I was particularly struck by this week's readings because I still typically understand the influx of technology into different humanities-based fields as a threatening change and I am typically wary of these

tools. What these articles illuminated to me is that computational analysis can in fact deeply contribute to literary study without entirely replacing the scholar. I am realizing how much can be missed by “linear readers” and the distant reading that computers can actually fill in the gaps of humanistic research. In Blei’s “Topic Modeling and Digital Humanities” he explains the importance of, “developing modeling components and algorithms that are tailored to humanistic questions about texts.” This highlighted the value of taking advantage of the new research tools we have access to. I was also very appreciative of Robert Nelson’s “Mining the Dispatch” because it provided a concise and impactful example of how this technology truly be utilized to further specific research fields. In the introduction to the project they write, “Topic modeling uses statistical techniques to categorize individual texts and, perhaps more importantly, to discover categories, topics, and patterns that we might not be aware of in those texts.” It is the idea that computational modeling and analysis can actually uncover patterns in documentation that scholars would otherwise be unaware of. This reframes the entirety of Digital Humanities for me to be honest — it turns the field into a collaborative one that maintains the integrity of a “humanistic” project.

Pia Bhatia

11:07 AM In the Blei piece, they write: “Note that the statistical models are meant to help interpret and understand texts; it is still the scholar’s job to do the actual interpreting and understanding. A model of texts, built with a particular theory in mind, cannot provide evidence for the theory.[5] (After all, the theory is built into the assumptions of the model.)” I was curious about how researchers go about finding these assumptions when using topic modeling. Does it involve a kind of reverse-engineering, or is it merely a question of using the tool of a scholarly analysis as separate from an online tool?

“There are around 15 chapters in Moby Dick in which nobody talks about whales at all and we discuss boats in great detail; this is something we may not notice as linear readers but it is something that computers are very good at showing us.” In Heather Froehlich’s essay she illuminates one of the greatest benefits that can be accrued by simply counting words, which is the ability to view a corpus non-linearly. This struck me as extremely important because as readers, our reading of a chapter tends to be influenced by what the text that has just preceded it. Essentially, our view of a work wherever we are reading from is cumulative, and these softwares actually introduce one solution to what appears to be a biased way of viewing a piece of literature.

Clay Glovier

4:36 PM I enjoyed reading Froelich’s article about how using computers to analyze texts can be useful. I was intrigued by the author’s focus on “less visible” words such as “the, and, he, she, I, or” etc, as it is true that I do not pay much attention to these while reading due to their high frequency. However, they do have value that is important for understanding texts. In Macbeth for example I was surprised to learn that the word “she” is used less frequently than in Shakespeare’s other novels even though Lady Macbeth is the most important figure. This fact alone would provide an Interesting paper topic! In Mining the Dispatch I appreciated the author’s demonstration of the ability of computers to enable historical research. The frequency of “For Hire and Wanted Ads” exploded during the end of 1862 and start of 1863. This jump in the chart would offer another interesting data point from which to begin an essay or scholarly work. I also found Blei’s piece on Topic Modeling and Digital Humanities intriguing. I was particularly interested in the fact that the data analyzed by the topic model was sourced from the New York Times. Given the newspaper is suing OpenAI for training its model on its data (representing articles written by its journalists) I am curious whether permission was granted to the researchers or if they merely scraped this off the web? This

question of groups online using other people's published data to further their research is interesting and I am curious how it will develop.

Anya Kalogerakos

5:11 PM I enjoyed the ideas presented in Froelich's article, especially that there is some balance between the frequency of a word and its value. The less occurrent a word is, the more noticeable it is, but at the same time, it takes a certain frequency to be considered important. I was interested when Froelich discussed the trends in the frequency of words and how that held meaning. For instance, in Moby Dick, "old" in context was used as a descriptor for people, while "new" in context was used as a descriptor for location. While this is an interesting find, it did make me question how valuable this information is, as this is probably true for a large number of books, and therefore doesn't do much to increase our knowledge about a specific book. I also found topic modeling to be interesting for its reliance on statistics and its apparent accuracy in the Nelson article, but despite its merits, I found myself a little suspicious of topic modeling. I think the Blei article, while it did a great job of describing the workings of topic model algorithms, it also made me realize just how abstract and subjective the idea of a topic is. Furthermore, words within a topic can sometimes feel really unrelated or unenlightening unless a lot of conjecture is made.

Talia Goldman

10:01 PM I enjoyed this group of readings, especially considering the ideas of the Froelich and Blei articles in the context of the "Mining the Dispatch" website. The Froelich reading paired nicely with the ideas of close and distant reading we were discussing before break, but I especially appreciated the emphasis on how topic modeling reads the obvious, which can in turn provide avenues for human interpretation. Using quantification as a tool is also discussed in the Blei article, but I found that Blei delves more deeply into the base assumptions that inform topic modeling. I particularly liked the idea that probabilistic models provide "discovered structure[s] as a lens for exploration." This connects to how revising models for topic modeling creates better ways for tools to read the obvious, allowing distant reading to yield results most closely representative of the actual text. Relating to the "Mining the Dispatch" website, I looked at "Patriotism and Poetry," which was a good example of the paths to research questions that topic modeling opens up. On the graph, there was a spike in July 1861 and another in January 1864. One route of exploration from this data might be, for example, how patriotic poetry played a role in social and political life in Richmond at the beginning and end of the Civil War. Considering this tool made me feel more convinced in the usefulness of topic modeling as a tool, and overall, I feel more clear about it potential (and limitations) after engaging with these readings and the project.

Alison Fortenberry

10:05 PM While I find the topic modeling tools used in these articles interesting, I think it did give me pause at first. The idea of using a tool to pull out themes from a work at first seemed like it was imposing too much onto the humanities; what is a scholar's role if a tool is doing the work of pulling themes out for them? While this is similar to other tools used in distant reading, like word frequency generators, this feels more like it's superseding the role of a scholar. Counting the word frequency in Moby Dick would be tedious by hand, but automation makes it much quicker. Pulling out themes, however, requires some level of analysis– it asks readers to digest the text and understand its broader meaning. Word counting feels like it's taking the role of tedious data collection, but topic modeling feels like it's taking some of the analytical work that a scholar should be doing. That said, I really enjoyed the authors' insistence in the readings that topic modeling needs to be paired with analysis or some kind of "so what?" about the model's findings. Blei frames topic modeling

as a tool to find evidence, but argues that the scholar's role is still to analyze it. That approach makes me more open to the idea of topic modeling, but I wonder if there is a way to guarantee that it is being used responsibly/ not usurping the scholar's role in the analysis. If we get to a point where analysis is generated by a model, then what is the purpose of humanistic scholars?

Layla Williams

11:28 PM While the week of March 6th's readings were helpful because they explored some of the biases found within the data models that people create, this week's readings contributed to my working definition of the digital humanities as a field. I think explanations of topic modeling within Moby Dick, for example, illustrate that the "digital" in digital humanities are the tools that facilitate humanistic inquiries. Topic models track patterns in language, and the scholars then determine the actual significance of those patterns. The specific instances of naming (such as Ahab v Captain Ahab or "she" in Macbeth) lay the foundation to ask more questions about the results of topic modeling. It can be a point of inquiry used to make a path into a research project or potentially a database in itself. In addition, I found the "Mining the Dispatch" website helpful because it was like an "ngram" created based off of the topic models. It made me realize that some of the products of the digital humanities are tools I interact with in other areas. And its uses can be extended beyond literary analysis and into historical reconstruction. How can topic models be used in the archives? Will the digitization process make topic modeling easier for libraries?

Helen Gao

1:22 AM The "Moby Dick" article was a great explanation of how techniques as basic as word counting can be used to reveal deeper meanings in the text and even generate possible research questions. I liked that Froehlich included examples from both Moby Dick and Macbeth, and I appreciated the specific comparisons and discussions of word choice. In "Topic Modeling and Digital Humanities", I thought it was interesting to read about the definition of a 'topic' from such a mathematical perspective – I usually consider 'topics' to be a relatively vague and qualitative term, but Blei defined it clearly as co-occurring terms. I did feel like Blei was trying to describe the "conceptual process" of topic modeling in the simplest possible way, but I still found his explanation to be a little bit confusing at times – I think a diagram or a specific example (beyond the list of New York Times topics) would've really helped in this article. Specifically, I thought how Blei wrote about "choosing" the topics, weights, assignments, and observations was rather opaque – how does this choosing process work? "Mining the Dispatch" answered one of these questions, explaining that the topics are based on the patterns from the algorithm and not selected by the researcher. However, the fact that the number of topics is an input from the researcher reveals that some element of this process is still somewhat arbitrary. It was really interesting (and sad) to learn about how the topics in the Daily Dispatch advertisements reflected the cycle of "slave hiring". The "Mining the Dispatch" introduction also demonstrated how topic modeling, distant reading, and exploratory data analysis more broadly can be used to help formulate questions.

Ethan Haque

2:13 AM I think that topic modeling is a cool tool, but not a very powerful one on its own. In fact, I think the Blei article is wrong when it describes an idealized version of how methods like LDA might be used. For reference, here is the short paragraph I take issue with:

> What does this have to do with the humanities? Here is the rosy vision. A humanist imagines the kind of hidden structure that she wants to discover and embeds it in a model that generates her archive. The form of the structure is influenced by her theories and knowledge — time and geography, linguistic theory, literary theory, gender, author, politics, culture, history. With the model and the archive in place, she then runs an algorithm to estimate how the imagined hidden structure is realized in actual texts. Finally, she uses those estimates in subsequent study, trying to confirm her theories, forming new theories, and using the discovered structure as a lens for exploration. She discovers that her model falls short in several ways. She revises and repeats.

The way I read this implies the author makes two assumptions about these kids of NLP methods. First, that there is objectivity in downstream analysis based on these results. Second, and this is a little less concrete, that we should search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses. That is, introduce confirmation bias into our research.

Any sort of topic modeling tool like LDA is based on a set of inductive biases that are necessary to make them good predictors. LDA doesn't work well if there are tons of topics in the corpus. It doesn't work well if the ordering of specific words is extremely important. Likewise, state-of-the-art topic modeling techniques make certain assumptions about the inputs that may not be true for exotic corpora. This is to say there is no objectivity with these tools. The choice of tool is as consequential as the downstream analysis. It's not safe to take the results of these models at face value without really examining their outputs closely and possibly comparing them with other analysis techniques. As for the second point, I think this idealized process they've described is not quite how these hypothesis-based research questions should be approached. It sounds like bad science. In my own experience, topic modeling makes for cool visualizations and pretty graphs, and it also might help consolidate ideas, but generally doesn't do much on its own. However, because it's a computational tool people put a lot more stake in it than it deserves.