Jessica Shainker

History & New Media

Dr. Trevor Owens

Topic Modeling Enslavement Narratives

Two hundred and ninety-four. That is how many text files are present in DocSouth's North American Slave Narrative archives, an incredibly rich and underutilized resource that is freely available online.[1] As simple text files downloadable in a compressed folder, these narratives are already optimized for data analysis. And as component parts of such a large collection, without big data methods they are unparsable by any individual researcher within a reasonable amount of time. As a result, this archive was perfect for my first foray into topic modeling. Topic modeling, a form of computational analysis, is intimidating at first - and second, and third - glance. But even for historians who have no experience with data analysis or programming, topic modeling is a useful exercise for finding patterns in historical data and generating potential research questions.

In topic modeling, the researcher inputs their data into a computer program - in this case, MALLET - which then runs that data though an algorithm.[2] The algorithm recognizes and lists words that often occur near one another. Depending on additional parameters set by the researcher, the program can also rank these groups of words, or "topics," based on how prevalent they are in the data set.

The algorithm doesn't understand any of the words that it processes. Individual words are parsed as simple strings of characters, meaningless apart from their relationship to nearby

---

[1] "North American Slave Narratives," *DocSouth,* University of North Carolina, https://docsouth.unc.edu/neh/.

[2] "Mallet: MAchine Learning for LanguagE Toolkit," https://mimno.github.io/Mallet/index.

strings. This “dumbness” has both drawbacks and benefits. On the one hand, the computer cannot recognize misspellings or alternate spellings of words. One typo may not make much of a difference in the final model, but in some cases, MALLET’s inability to assign meanings to words might leave gaps in the analysis. Data sets composed of, for instance, transcribed oral histories with multiple spellings of the same term may produce a skewed model.

On the other hand, MALLET’s inability to recognize meaning also limits biases that can skew models in other ways. As BIPOC and feminist scholars have pointed out, “technology is not neutral.”[3] Automated data analysis always risks amplifying the programmers’ biases, the researcher’s biases, and biases embedded in the data themselves. Charlton McIlwain, for instance, has written extensively on how policing algorithms that were ostensibly objective actually amplified racist outcomes across the country.[4] Unlike programs that are trained on previous, potentially racist, data sets, however, MALLET can be used without training the program ahead of time. It doesn’t understand the texts that it is reading, so while data set and researcher bias may still be a problem, there is less of a chance that biases embedded within the code will have an effect. In this case, MALLET’s dumbness may actually preclude some of the pitfalls of other big data methods. Some black feminists have even singled out topic modeling as a tool of recovery, positioning it as a potential solution to race and gender bias in historical archives.[5]

---

[3] Donna Haraway, "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century," in *Simians, Cyborgs and Women: The Reinvention of Nature* (Oxfordshire: Routledge, 1991),149-181.

[4] Charlton D. McIlwain, *Black Software: The Internet and Racial Justice, From the AFRONET to Black Lives Matter* (New York City: Oxford University Press, 2021).

[5] Nicole M. Brown, Ruby Mendenhall, Michael L. Black, Mark Van Moer, Assata Zerai, and Karen Flynn, “Mechanized Margin to Digitized Center: Black Feminism's Contributions to Combatting Erasure within the Digital Humanities,” *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 110-125, https://www.euppublishing.com/doi/full/10.3366/ijhac.2016.0163.

Within the past ten years, other scholars have also begun integrating MALLET and similar programs into their research. Cameron Blevins, for instance, has done extensive work topic modeling Martha Ballard's Diary.[6] Matthew McClellan has done some preliminary research on modeling enslavement narratives with a particular focus on Olaudah Equiano's diary.[7] Sarita Alami, Moya Bailey, Katie Rawson, and Sara Palmer used MALLET to analyze sermons given on the occasion of Lincoln's assassination.[8] And with perhaps the best funding of any projects of this sort, Robert K. Nelson created Mining the Dispatch, a project which analyzes Civil War Richmond through a newspaper archive.[9] Historians are beginning to explore the possibilities that MALLET presents.

Some historians have even begun doing research with the same data set I am using. Laura Tilton's analysis of racialized dialect in Federal Writers' Project enslavement narratives references DocSouth's archive among other sources.[10] Jed Dobson explored MALLET as one part of his larger book project.[11] And Jim Casey similarly dabbled in using MALLET to analyze these narratives, even going so far as to create a Gephi visualization.[12] Unfortunately, most of these projects - my research included - are only surface level explorations of MALLET's

---

[6] Cameron Blevins, "Topic Modeling Martha Ballard's Diary," Cameron Blevins, April 1, 2010, http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.
[7] Matthew McClellan, "Project Overview," Topic Modeling Equiano and the Slave Narrative Genre (Harvard University), accessed April 30, 2022, https://hist1993-15.omeka.fas.harvard.edu/exhibits/show/modeling-equiano/project-overview.
[8] Sarita Alami et al., "Topic Analysis With MALLET," Lincoln Logarithms (Emory University, April 6, 2015), https://lincolnlogs.digitalscholarship.emory.edu/mallet/.
[9] Robert K. Nelson, Mining the Dispatch (University of Richmond), accessed April 30, 2022, https://dsl.richmond.edu/dispatch/.
[10] Lauren Tilton, "Race and Place: Dialect and the Construction of Southern Identity in the Ex-Slave Narratives," Current Research in Digital History (Roy Rosenzweig Center for History and New Media, George Mason University, August 23, 2019), https://crdh.rrchnm.org/essays/v02-14-race-and-place/.
[11] Jed Dobson, "Introduction," accessed April 30, 2022, https://jeddobson.github.io/textmining-docsouth/.
[12] Jim Casey, "Topic Networks of Slave Narratives, Part 1," Jim Casey, December 14, 2014, https://jim-casey.com/posts/topic-networks-of-slave-narratives/.

potential. Few historians have so far developed the technical expertise necessary to fully incorporate MALLET into their research methods.

Ultimately, to utilize MALLET to its fullest potential and address biases embedded in their data, scholars will need to open up the "computational black [box]" that is MALLET's algorithm.[13] They will need to understand its inner workings so that they better understand why and how it creates specific topics. In the short term, however, MALLET can be used safely - e.g., without too much risk of amplified bias - if it is simply used as a brainstorming tool in the preliminary stages of a research project. To this end, I modeled my data set multiple times to get a better sense of what topics recurred, representing highly prevalent subjects, and which subjects appeared unexpectedly in individual models, representing previously unrecognized research possibilities.

Methods and Results

In data analysis, quantity matters just as much as quality. I quickly realized that the only way to gather any sort of meaningful data was to run a gamut of tests. The richest analysis and best pattern recognition, it seemed, came from comparing the results of multiple models. To this end, I first ran through an excellent tutorial from the Programming Historian.[14] Then, substituting my own data for the tutorial's sample data, I created a MALLET file containing all of the texts

[13] Mark Sample, "Part V Chapter 43: The Black Box and Speculative Care," Debates in the Digital Humanities 2019, accessed April 30, 2022, https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/3aa0b4f4-bd72-410d-9935-366a895ea7a7.

[14] Shawn Graham, Scott Weingart, and Ian Milligan, "Getting Started with Topic Modeling and Mallet," Programming Historian, September 2, 2012, https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

from DocSouth's North American Slave Archive. Once I had the resulting file, I was ready to create my topic models.

I ran the first cycle of tests, aimed at determining the optimum number of topics, using an optimization parameter of 20, which is the number suggested in the tutorial.[15] After running models set to identify 10, 20, 30, 35, and 40 topics, I ported the resulting files into Google Sheets, where I was able to sort by topic weight and tentatively assign names to each topic. I decided that 20 topics was, for my purposes, the ideal number, avoiding on one hand topics that were too vague or repetitious and on the other hand topics that were too specific, only relevant to specific texts.

To be clear, there is no established best practice for how many topics one needs for each data set.[16] Scholars all have their own ideas about what the ideal number of topics are, and professionals who have the expertise to parse MALLET's inner workings probably have more quantitative ways to determine this number, but it ultimately depends on the specifics of your particular data set and research goals.

While I decided on 20 topics for the reasons listed above, I actually found comparing a range of differently-sized models to be more useful for recognizing patterns and potential research subjects. For instance, I found that every model I ran recognized the prevalence of topics related to the Haitian Revolution, Religion (in three dimensions - spiritual, institutional, and community), and the State. In most of the models, though not the smallest ones, I found that racialized dialect, maritime life and the international slave trade, family, and food and agriculture

---

[15] Jessica Shainker, "SlaveNarr Topic Number Test," https://docs.google.com/spreadsheets/d/1f_UaE9VhFlhZ5TCBANzKDYBNl4iNyHPeb_-C4EuhWsc/edit?usp=sharing.

[16] Mark Needham, "Topic Modelling: Working Out the Optimal Number of Topics," Mark Needham, March 24, 2015, https://www.markhneedham.com/blog/2015/03/24/topic-modelling-working-out-the-optimal-number-of-topics/.

remained prevalent. And in the largest models, the topics of women, the military, poetry/prose, and individual narratives repeated themselves. The fact that MALLET recognized these topics as prevalent throughout multiple models points to their importance within the data set.

While the topics listed above were common, some of the most interesting topics that arose were uncommon. Commerce, the plantation landscape, and the English empire, for instance, only showed up once or twice across all five models. Additionally, while topics like "the state" and "family" are accurate - they certainly play a large role in enslavement narratives - they are too broad for fruitful research. To address this concern and hopefully reveal more focused topics, I ran a second series of tests aimed at determining the best optimization parameters.

When topic modeling with MALLET, the optimization parameter determines how often the algorithm reevaluates the prevalence of each topic. I do not understand how this works. I do know, however, that using optimization parameters allows MALLET to assign "weights" to each topic, effectively telling the researcher which topics are more prevalent than others. I ran five models of 20 topics each, with the optimization interval set to 5, 10, 20, 30, and 50.[17]

Similarly to my previous test, I cannot fully determine which optimization parameter is the best, but comparing the models to one another yielded interesting results. The same topics that were most prevalent in the previous test remained prevalent in these new models: Haiti, the State, and various iterations of religion, among others. But with varying optimization intervals, some unusual topics rose to the surface as well: Education, Abolition, Domesticity, Ranching,

[17] Jessica Shainker, "SlaveNarr Optimization Tests (Topic20)," "https://docs.google.com/spreadsheets/d/1nY3k4rELDMJN59GdaLK8F-G_zp-byPS1R5pzE3FrZ8M/edit?usp=sharing.

and Women & Embodiment, for instance. These topics, while still broad, offer slightly more focused research pathways.

Conclusions and Future Research

In the future, I would like to model how these topic models present over time. Measuring, for instance, the relative presence of the Haitian Revolution in enslavement narratives between the time of the revolution itself and the Civil War would be fascinating. Similarly, I would also like to visualize the relative frequencies of different approaches to religion in these texts. Throughout multiple models, religion reliably presented as three separate topics: spirituality or metaphysical truth; religious institutions and their relationship with the state; and religious communities, including neighbors, family, and friends. Charting the changing ways that these topics present in enslavement narratives over time would, I predict, yield interesting results.

I would also like to conduct tests on the drawbacks and benefits of modeling this data paragraph-by-paragraph, rather than memoir-by-memoir. A huge portion of my research efforts were focused on this question, and I went so far as to write a Python script that split the memoirs into their component paragraphs. Unfortunately, answering this question was simply outside of the scope of what I could accomplish this semester. This was largely because of my rusty programming skills - a short script took me five times longer to write than it should have. I was deeply disappointed that I wasn't able to come up with any definite conclusions on the pros and cons of modeling shorter text snippets. As a result of my efforts, though, I am more equipped to manipulate text files with Python in the future, and I am more determined than ever to continue this research.

In conclusion, I plan on staying up-to-date on topic modeling in humanities scholarship. In particular, I hope to . Now that I have familiarized myself with this technology, I feel more confident in my own ability to learn the technical side of modeling thoroughly enough to use it in historical research. The next and more important challenge to overcome is to ensure that I am using it in a way that amplifies and honors the voices of the formerly enslaved and their descendents.

Topic modeling is often intimidating to humanists who have never used quantitative methods before. In some cases, this hesitance is warranted - historians are notorious for misusing or misinterpreting MALLET results. Many of the pitfalls of MALLET are avoided, however, when it is only used as a jumping-off point for more traditional research methods. MALLET's greatest strength is for in-depth pattern recognition and quantitative analysis, but for graduate students and people who are new to quantitative methods, employing MALLET as a brainstorming tool is a great way to ease into digital tools and quantitative analysis.