Moriah L Kirdy
Computers and Writing 2018

Topic Modeling and/as Genre Study

As my title indicates I'm going to be talking about topic modeling and/as genre study. I'll preface by saying that my research has involved rhetorical genre studies for some time, and that while I've long enjoyed experimental modes of reading and making to get myself out of my own head while engaging with texts, I've only entered the topic modeling fray in the last couple of years. I'm very much a novice programmer, and feel I can read it better than I can write it. My dissertation research is on pen-and-paper notebook-based information management systems like the commonplace book, a genre of reading log popularized in the early modern era, and the contemporary productivity and mindfulness practice, the Bullet Journal. So it might seem counterintuitive that I chose to pursue computer-assisted methods for the study of pen-and-paper genres, but much of the talk around these genres occurs across social media and contemporary lifestyle blogs, webzines, and discussion forums. In addition, these "paper programs" as I'm calling them, are organized by and discussed through a shared set of conventional keywords, headings, and contemporary topoi (such as "self-care," and "mindfulness").

I took on the task of topic modeling the top lifestyle blog entries related to bullet journaling using the popular toolkit MALLET from the time the system was "released" on the bulletjournal.com website by its creator Ryder Carroll up through November 2017. In his original depiction of it, the Bullet Journal is a task-management system, "the analog system for the digital age," whereby at the start of each month and then each day through what he calls "rapid logging" you compile all of the tasks, events, notes, and other items on your mind to think about or get done in a bulleted list which you organize by assigning different kinds of "signifiers" or icons. For example, tasks might be the traditional bullet, a note is signaled by a dash, and events by an open circle. At the end of the day, you process your list by assigning a different set of icons. An "X" to identify tasks that have been completed, an eye to signal something to do more research on, and a greater than symbol to indicate that a task is incomplete and needs to be represented in the next day's list. Carroll calls this act of processing "migration."

I wanted to look at how this formal discourse used by Carroll to describe the system had evolved and changed once in the hands of a population of practitioners, practitioners who have taken the basics of this system and have customized and evolved it in ways Carroll could not have anticipated. Without going into too much detail, it's worth showing you an image from a popular lifestyle blog post on Bullet Journaling, also capturing daily lists. The central mechanic of the itemized bulleted list is still present, but you can see how at least this practitioner has taken to more complex and artistic design, using color-coding and other materials like the sticky tabs on the far right. I'll just say for the moment, this looks much more like what most people these days call a "bullet journal" than Carroll's original concept.

I only looked at 106 blog posts, so as to not repeat posts by popular authors (and thus flood the model with a particular author's discourse). In approaching the work of this chapter I hypothesized that given the Bullet Journal's clear origin story and "official" discourse, as a computer assisted-method based on the prevalence of associated terms, topic-modeling might reveal something about the ways in which that sanctioned discourse has persisted and/or transformed in the short amount of time since the practice has been circulating and gaining popularity. Genre stabilization and change are notoriously challenging aspects of genre work to track. So to be clear, I'm not working with bullet journals themselves, I'm working with meta-genres, a term I borrow from Janet Giltrow, that is, the talk around genre practice that describes it and instructs others how to do a genre's work.

So I had some speculative theories and some ideas about what topic modeling might teach me about this set of meta-genres advocating for the Bullet Journal system, but I had little knowledge of topic modeling itself beyond that it was a way to identify clusters of associated words (topics) that have some prevalence across a collection of texts. In my hair-pulling attempts to understand Latent Dirichlet Allocation, the algorithm MALLET uses for modeling a corpus of texts, my understanding of it was unlocked by an aspect of topic modeling that is often not discussed as central.

Ted Underwood's advanced use of topic modeling for literary study is expansive and his long-standing blog has been an incredible resource to many. In a post aptly titled "Topic modeling made just simple enough," he describes topic modeling as follows: "Say we've got a collection of documents, and we want to identify underlying 'topics' that organize the collection. Assume that each document contains a mixture of different topics. Let's also assume that a 'topic' can be understood as a collection of words that can have different probabilities of appearance in passages discussing the topic…Of course, we can't directly observe the topic; in reality all we have are documents. Topic modeling is a way of *extrapolating backward from a collection of documents to infer the discourses ('topics') that could have generated them*." My close reading tendencies were buzzing. What is with Underwood putting topics before documents in inventional order?

As I dug in even further, I finally took on the task of reading (and rereading, and annotating, and rereading) David Blei's original introduction to LDA in "Probabilistic Topic Models" wherein he writes: "We formally define a *topic* to be a distribution over a fixed vocabulary… We assume that these topics are specified before the data has been generated" after which is a footnote that reads: "Technically, the model assumes that the topics are generated first, before the documents."

That the statistical algorithm was written with this assumption sparked an interest in me I wasn't expecting. According to the members of my dissertation committee and others I've shared the method with, the topic-first assumption of LDA is perhaps one of the most challenging aspects to grasp. After all, even with rhetorical genre study it is most common to begin with texts, and even if we don't mean to or want to, the texts are treated as points of origin, even while we're fully aware of the complex processes of invention, the ways in which any text is informed by socio-cultural influences, not to mention the material, technological, and further influence of media networks and nonhuman actors. I was perhaps a little better equipped to *see* the topic-first orientation as a generative perspective having come at topic modeling from a genre perspective.

Here I make my key observation of this talk. And it is an observation -- not a claim, or an argument. But it is an observation, I feel, that has potentially important rippling effects for future uses of topic modeling and/as genre study.

In Amy Devitt's seminal work *Writing Genres,* she contributes to the long and complex definition of "genre as social action" (from Carolyn Miller's important article of that name) the context of genre as an important factor in genre work. That is, she pleads to those attending to genre to acknowledge that genre knowledge at once precedes and follows any particular genre instance or practice, that "[O]ne never writes or speaks in a void. What fills that void is not only cultural context and situational context, but also generic context, the existing genres we have read or written or that others say we should read or write…genres are always already existing… As opposed to an abstract concept of genre, the context of genres is the existence of particular genres, the already existing textual classifications and forms already established and being established within a given culture, the set of typified rhetorical actions already constructed by participants in a society" (27-28).

It occurred to me that those key contributors to the construction of topic modeling algorithms were operating from the context of genre, that the topic-first orientation is akin to the knowledge that discourse and subject conventions precede and outlast the documents within which they manifest. While often, as both Underwood and Blei do, the starting point of the inquiry is that you want to discover the latent topics underlying a collection of texts, the assumption that *precedes that assumption* is that the collection must also cohere in some way for the topics to be meaningful. That collections cohere often through genre is an unremarkable statement.

[Slide with a few topic model studies reliant on collections that cohere as a genre]. Here are just a few topic modeling studies proving my point. Devitt's position is not only that genre knowledge and experience exists *a priori* to any genre instance, but that the context of genre is the "past in the present," as a condition that continues regardless of any particular instance. That topic modeling often attempts to capture information about a collection of texts at scale, often at a much larger scale than my own small study, offers a unique glimpse into not only subject and discourse trends across a corpus, but into a sense of genre stabilization and change, not necessarily over time, but within the context of genre.

The study of meta-genre through topic modeling, in particular, offers an even clearer sense of what precedes and persists in advance of any particular instance, as it looks at the language *about* rather than the language *within*, and is thus more likely to capture information like the discourse used to describe genre convention, customization, and adaptation, including, for example, language describing conventions from other genres.

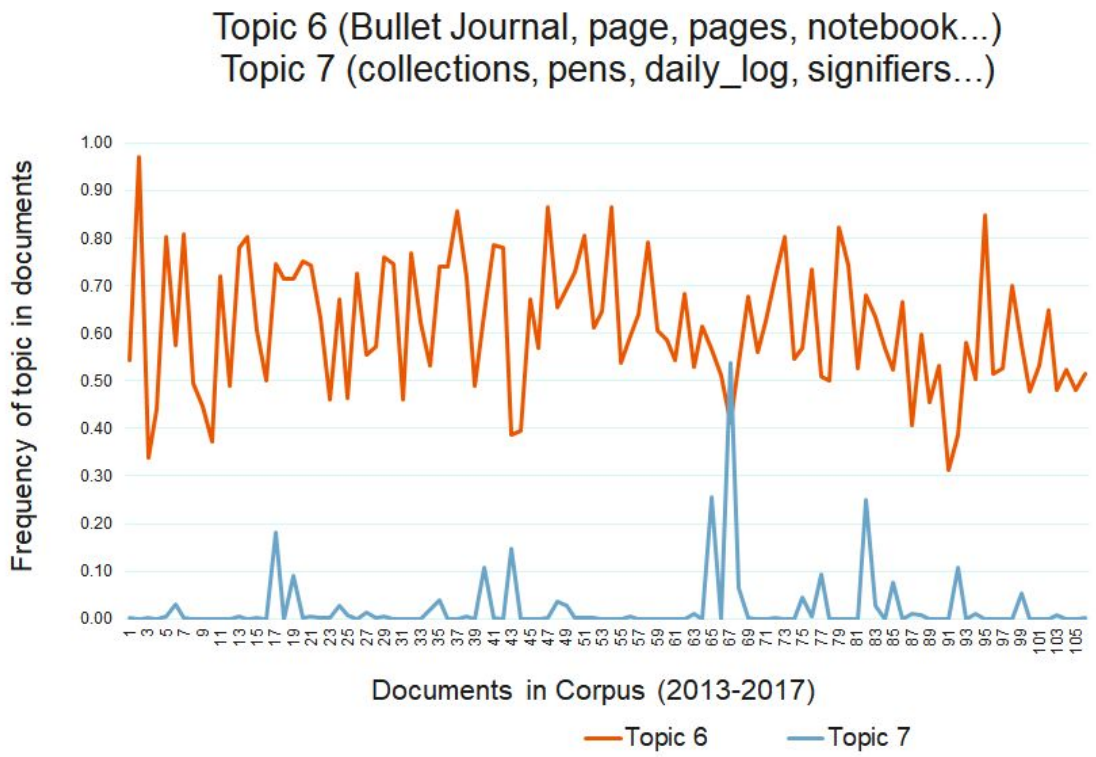
These are two topics out of ten in one of the models I generated for the bullet journal-related blog posts.

Topic 6	2.92662	Bullet_journal page pages notebook list day journal things time month system make love work planner daily monthly find post write
Topic 7	0.09472	Collections pen daily_log signifiers task entries future_log collection add index simply monthly_log bullets migrate leuchtturm entry log write migration fountain

Topic 6 has a very high distribution across the corpus, and Topic 7 has nowhere near as high distribution across the corpus. Were I to label these topics descriptively, I would describe Topic 6 as practitioner-dominant

discourse and Topic 7 as formal discourse (as in, the topic that best represents the terms Carroll himself designed to describe the system). The differences at first might seem fairly nuanced, but it's clear that by and large bloggers who advocate for the utility of Bullet Journaling use discourse that is closer to the context of genre related to planning and planners than terms Carroll designed to describe it, which prioritizes the *logging* action and the assignment of *signifiers* to tasks and the processing mechanic called *migration.*

Here is a representation of the statistical likelihood of each topic to appear across the corpus .



For segments of the corpus where the formal discourse is stable, the planning discourse is not, and visa versa. While very many posts are statistically likely to include topic 6, comparatively few are statistically likely to use Carroll's original formal discourse, represented by topic 7. So even while there's a point of origin for the Bullet Journal system, and a set of terms designed specifically to describe its component parts, its advocates who blog about it are more likely to rely on the context of the planner and planning genres than the formal terms designed by its creator. In part perhaps because of this interpretation of the system as a kind of DIY planner, the original task-management aspect of the system is minimized in favor of aesthetic and creative output. There are many potential implications for this, for instance that the language of planning and planners is more legible to potential audiences, by which I mean both human readers and search algorithms, than the formal terms Carroll came up with. As many of the bloggers who authored posts in this dataset are aiming to or already make money if not a living off of their blogs, that legibility is paramount for their own personal success on the web. Topic 6 is also an indication that the much longer history of planning genres makes the introduction of closely adjacent time- and task-based discourse an uphill sort of battle. It also suggests a way in which the context of

other genres, or the rhetorical move of comparison, is a part of the ways in which individuals describe or introduce emergent genres – because after all, we're not looking at context here, but from having read them, I know that the language of planning often comes up in the sense of, "the bullet journal *is kind of like a* planner..."

This is just one example from this chapter where the context of genre can offers a possible interpretive strategy for somewhat surprising topic clusters, and I hope it reads to you as a strong case for how modeling meta-genres, in particular, can give us a glimpse into the otherwise elusive concerns associated with genre uptake, stabilization, and change. By and large topic models are used as a content-mapping strategy, but this observation that the LDA algorithm was written in the context of genre, that language patterns discoverable through topic modeling precede (and continue beyond, and exist without) the documents that make up a collection of texts, suggests that topic modeling can offer insight well beyond subject trends.