**Organisciak — Remembering**

<3 figures under 2 figure numbers [one is in two pieces]>

**Remembering Books: A Within-Book Topic Mapping Technique**

Organisciak, P., Auvil, L. and Downie, J. S.

Applying clustering techniques in modeling text collections is effective for surfacing high-level themes at scales that humans cannot match with close reading. Within-book topic modeling is more difficult to perform, due to noise inherent to small frame sizes and the expectation that a single book can simply be read. In this work, we argue that this form of text analysis is effective as a research aid, presenting a method and tools to visualize the progression of topics through the course of a long text.

This approach focuses on improving human readability of machine-modeled topics and is notable for (a) a sliding-frame topic inference technique that smooths noise and (b) a visualization approach that uses sorting and filtering to help focus attention on clearer topics and their narrative order.

The goal is to use topic modeling as an aid to a reader’s understanding of a work as well as the ability to communicate that understanding. In particular, it appears to have promise as a mnemonic tool, helping past readers recollect plot points and progression. As this work progresses, it will explore better ways of surfacing important topics and measuring the quality of its visualizations. To this end, a tutorial and scripts have been released, enabling researchers and instructors to evaluate this technique for their own uses.

**Data**

This study was performed using the extracted features dataset from the HathiTrust Research Center (HTRC). HTRC is the research arm of the HathiTrust, developing tools for research access—and particularly large-scale access—to the holdings in the HathiTrust Digital Library. For our purposes, the HTRC’s feature extraction dataset is used less for its breadth than its features. The extracted features dataset includes page-level feature information, including part-of-speech tagged token counts with headers and footers removed. Our approach will work similarly on any clean text documents with page-level information or comparably sized sections—for example, blocks of TEI marked-up sections or paragraphs.

**Approach**

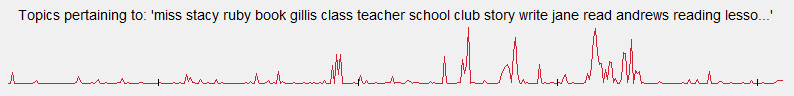
This study’s method for within-book topic modeling is notable for its approach to training and inference: where the training data is composed of individual page texts, the data it is inferred against is a sliding frame of pages. This lends a coherence to the resulting topic models by offering clean input data while assuming that actual occurrences of topics in a book occur in broader spans of text. We take a conservative view of what a theme in the text is when training, and a liberal view when inferring.

While pages are succinct for training, most of the time they are a physical artifact, disconnected from the content of the book. Language also deviates from page to page: inferring their topics in groups of pages allows for themes at that point in the book to emerge more clearly. Since we do not know where topics start and end, nor can we reliably assume that any particular part of a book is about any one topic, we use a sliding frame of groups of pages.

Latent Dirichlet Allocation (LDA) is used in this approach (Blei et al., 2003). LDA is a generative mixture modeling approach used to estimate probability distributions over correlated data. When used with text unigrams as features, these distributions are often interpreted as ‘topics’. One can think of a topic distribution as a word generator, outputting different words with varying frequencies. A topic about ‘Valentine's Day’, for example, is more likely to generate the terms ‘love’ and ‘February’ than one would normally see in the English language; to guess how likely a document is to be about Valentine’s Day, we can look at the likelihood of the Valentine’s Day topic generating the particular words in the document.

In training a generative model, it is ideal to have each input document represent a minimal number of concepts. While LDA is robust at differentiating different components of a training document, a clearer input improves the possibility of coherent clusters. For this reason, training a classifier on a large chunk of text such as a book can lose the nuance of the various themes within that text. For this study, we train on individual pages of a work. This page-level information is offered in the HTRC Feature Extraction dataset, and modeled using the LDA functionality of the MALLET toolkit (McCallum, 2002).

After a work is modeled against its pages, the resulting topic distributions are then applied to infer the most probable topics of sliding frames of pages in the same work. For example, for a frame size of 5, the work is processed into sections representing pages 1–5, pages 2–6, 3–7, and so on. Figure 1 demonstrates the difference in readability that the sliding frame allows.



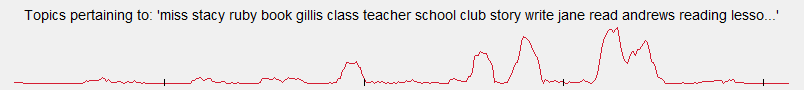


Figure 1. Example of topic for *Anne of Green Gables* inferred without (top) and with (bottom) sliding page frame.

In pursuing a technique for aiding an individual’s understanding and recall of a work, inferring coherent within-book themes is only as important as their presentation. For this reason, this study pursued visualization of within-book topic models with sorting and filtering.

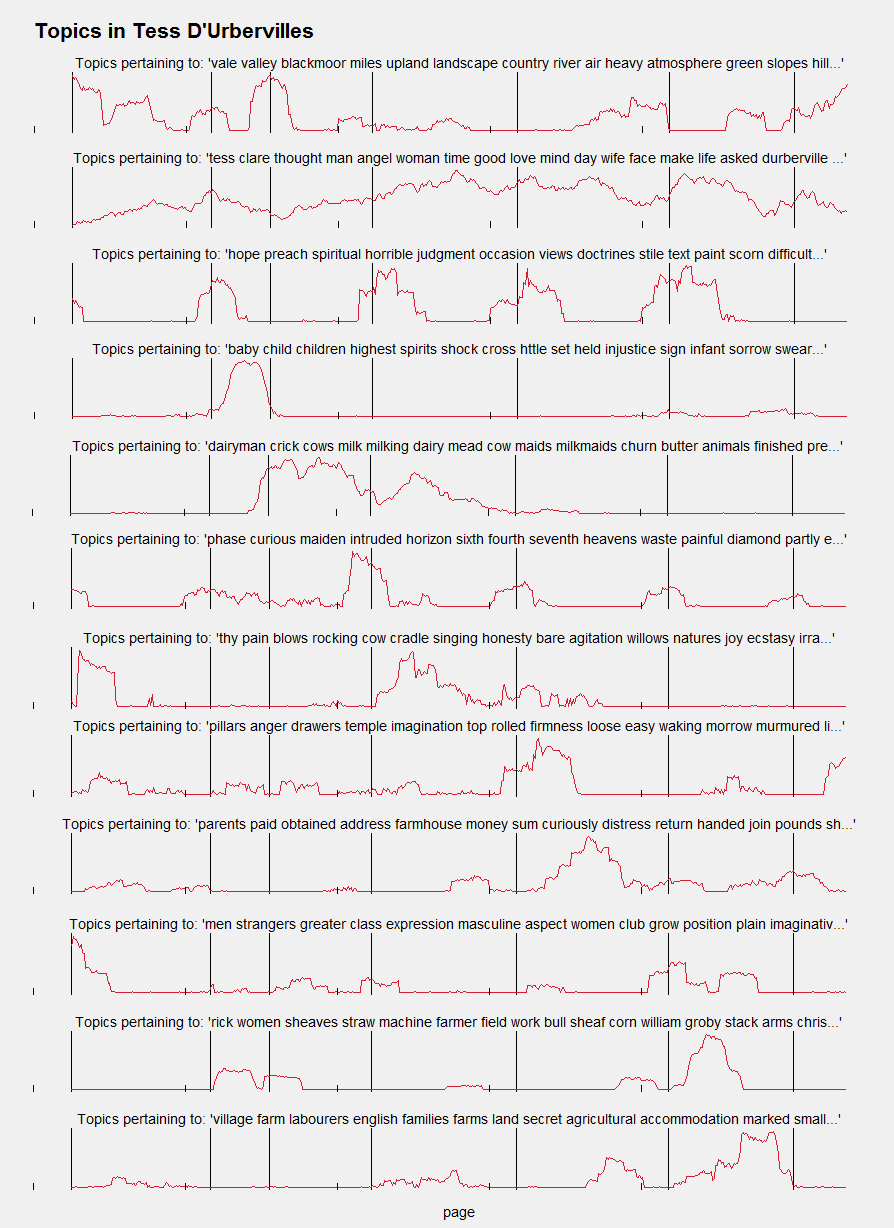


Figure 2. Topics in *Tess of the D’Ubervilles*, shown with a sliding frame of 10 pages. Every third of 30 topics shown, to demonstrate topic sorting with brevity.

For sorting, topics are tagged by their most representative page, and subsequently visualized in chronological order of these pages. This sorting technique has proven to be effective at showing their progression through a work.

Not all topics are equally insightful. Particularly, there are usually a few topics that serve as catch-alls: attracting probability mass for difficult-to-assign terms. These ‘noise topics’ serve a useful function, but not for an individual’s analysis or understanding. We attempted to identify these topics and filter them through a number of methods, so that they could be ignored in visualization. This included looking at high-variance distributions, distributions with high peakedness, and topics with disproportionately large pieces of the overall probability mass. Unfortunately, each of these techniques would filter out some useful topics, so we thus far have not found a filtering approach that is worth performing. Instead, the sorting performed at visualization leads with the most likely ‘noise topics’, making them easy for a person to visually assess and ignore if necessary.

Figure 2 demonstrates topic visualization for topics seen in *Tess of the D’Ubervilles* by Thomas Hardy. For brevity, only every third topic is shown, showing the sorting progression and the grouping of potential but not necessarily noisy topics at the start. In addition, vertical lines are included to denote the ‘phases’ in the book, allowing a comparison of topics to parts of the book. For example, Phase the Second clearly contains most of the discussion around the protagonist’s birth and subsequent loss of her child, and the language representative of this topic does not recur again in the book.

**Next Steps**

The code for this project is available online.1 As this is a work-in-progress, we hope this tool encourages others to use it and provide feedback. Thus far, this study has pursued qualitative improvements based on the judgments of the authors. As we move forward, we hope to evaluate this approach against the satisfaction of domain experts, compared to previous techniques.

Additionally, the goal of filtering uninteresting or noisy topics is still interesting, even if our given approach was not tractable. A more deliberate approach might be more effective in the future; specifically, human judges will be asked for their opinions on the most insightful topic distributions. This will allow us to compare their responses to a number of statistical metrics about the distribution, potentially building a classifier for useful topics.

**Note**

1. *HTRC-Book-Models*, Github, https://github.com/organisciak/htrc-book-models.

**References**

**Blei, D. M., Ng, A. Y. and Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* **3** (2003): 993–1022.

**McCallum, A. K.** (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.