Serendip: Turning Topics Back to the Text

Eric Alexander*
University of Wisconsin-Madison

Joe Kohlmann†
University of Wisconsin-Madison

Robin Valenza‡
University of Wisconsin-Madison

Michael Gleicher§
University of Wisconsin-Madison

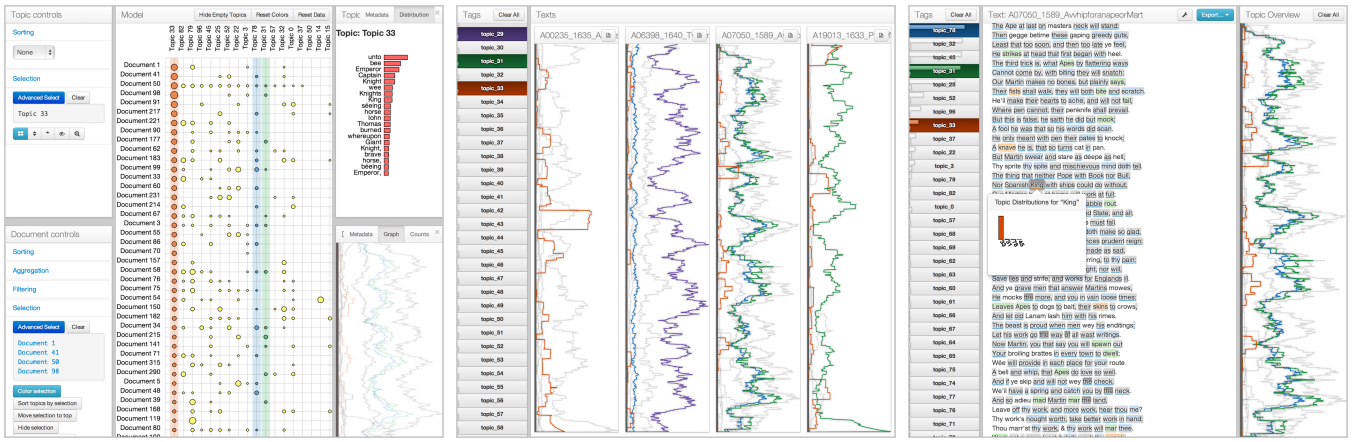

Figure 1: Serendip's three levels: CorpusViewer, MesoViewer, and TextViewer.

Abstract

Statistical topic modeling is an increasingly popular approach to text analysis. Many existing visualization tools focus on analyzing the model itself, distinct from the documents upon which it was trained. In contrast, we seek to treat the model as a lens through which to *view* the original documents. This would enable the reader to observe trends and build hypotheses at multiple scales—ranging from across a corpus to within a single text—and find both algorithmic data and textual examples to defend these hypotheses. Supporting this workflow requires a multi-tiered framework that affords comparisons at three levels: the entire corpus, small sets of documents, and a single document. We provide such a tool in our implementation of Serendip, a web-application that combines view-coordinated reorderable matrices, small multiples displays, and tagged text in order to allow readers to develop insight at multiple levels and carry that insight into their analysis of the others.

Keywords: Text visualization, topic modeling.

1 Introduction

A large corpus of text can be explored at many scales. At a high level, statistical models can be used to infer global trends in the corpus with what has been called "distant reading" [6]. In contrast, *close* reading considers individual documents word by word, paying careful attention to specific passages. At an intermediate level lie a number of largely underexplored properties of text, such as the patterns exposed by the rise and fall of a document's plot or argumentative structure. Though these levels are typically explored using distinct tools (ranging from computer programs to ink-and-paper books), investigation into them is inherently interrelated and there is value in using them together: high-level statistical trends may suggest interesting sets of documents or require specific exemplary passages to verify; specific passages of a document may suggest broader trends worth exploring across a corpus. Such multi-scale explorations require the smooth flow of information up and down the ladder of abstraction so that readers are able to form their analysis with insight from several levels.

Statistical topic modeling is a method of deriving latent topics from a corpus of texts. Specifically, algorithms like Latent Dirichlet Analysis generate distributions across a vocabulary for each topic and distributions across the topics for each document [2]. While there are some visualization tools for understanding these models [3, 5], our core idea is that topic modeling can serve as a guide for exploration of the *texts themselves* at all scales. This idea is realized in our prototype system, Serendip. To display how topics are distributed over a corpus, Serendip employs a reorderable matrix extended with aggregation and filtering functionality that enable it to scale to large corpora. To show how properties are distributed within a text or small set of texts, we provide "meso-scale" visualizations that are small multiples of graphical summaries. Finally, to show how the dense and potentially overlapping distributions of topic models manifest themselves over the text, we extend the tag encodings in tagged text displays to display multivariate data. These displays are visually linked through shared design elements, direct juxtaposition, and multi-view coordination to support interactive information flow.

2 The Tool

Serendip uses three different representations of the text, each at a different level of abstraction. At the furthest level, topic distributions of individual texts are represented as glyphs within a re-

*e-mail: ealexand@cs.wisc.edu
†e-mail: jkohlmann@wisc.edu
‡e-mail: valenza@wisc.edu
§e-mail: gleicher@cs.wisc.edu

orderable matrix. At an intermediate level, document structure is represented in color-coded line graphs showing the rise and fall of individual topics. At the closest level, topic assignments are represented as colored tags overlaid upon the raw text. These three encodings form the core visualization elements of the three components of Serendip: CorpusViewer, MesoViewer, and TextViewer.

2.1 CorpusViewer

Our highest level encoding of the texts is a reorderable matrix [1] which plots documents (rows) against topics (columns). The values of the document distributions are encoded as circular glyphs located at the vertices of the grid. It has been shown that people can find interesting attributes and patterns within the data of such matrices if they are given direct control of the orders themselves [7]. We extend this control to include a set of meaningful metrics by which readers can rearrange their data, including similarity distances, statistical properties of the distributions, and human-generated metadata. We give readers access to all of these features with which to aggregate, rank, filter, and annotate the data as they see fit. Rows and columns can also be moved manually.

2.2 MesoViewer

At the intermediate level, we sought to represent document structure as inferred by the topic model. Just as themes and subject matter will come and go throughout the course of a story, so do occurrences of a topic vary in density, especially in longer documents. We reflect these variations using line graphs displaying densities for each topic. MesoViewer provides a small multiples display of these line graphs, allowing easy comparison between sets of documents. Individual topics can be toggled on and off, in which case they are color coded to facilitate comparison.

2.3 TextViewer

At the level of passages, we annotate the raw text with data from the statistical model using colored backgrounds to highlight individual words. These sorts of "tagged text" displays allow us to adorn the text with additional information without sacrificing readability [4]. In our encoding, we use such tags to indicate the topics that our model predicts to have generated the words in a given document.

The process we use does not generate a single topic for each word, but rather a *distribution* across topics, such that words are potentially labeled with overlapping tags, each with an associated probability. We provide a number of options for displaying these probabilities (see Fig. 2). One option highlights words from multiple topics using gray backgrounds to indicate the overlap—they are then underlined with their most probable topic's color. Another option indicates the corresponding probabilities of the tags by representing topics as single-hue color ramps rather than single colors (e.g. a word with high likelihood of being in a topic might be dark blue while a word with low likelihood might be light blue). This gives greater perceptual weight to the tags about which the model is most confident. Finally, clicking on an individual word displays a popup showing its full distribution.

TextViewer's tagged text display is juxtaposed with another instance of the topic distribution line graph described in Section 2.2. This line graph operates as a navigational tool for longer texts. Simply by clicking on a section of the graph with an interesting feature (e.g. a peak or valley of a particular topic), the reader can navigate to the corresponding passage of text to see how said feature is reflected in the semantic content of the words.

2.4 Cross-Viewer Interaction

Interaction techniques across the three Viewers allow readers to identify interesting documents or sets of documents in one level and open them in another level. Shared topic colors and cross-viewer brushing aid in transferring insight from one level to another.

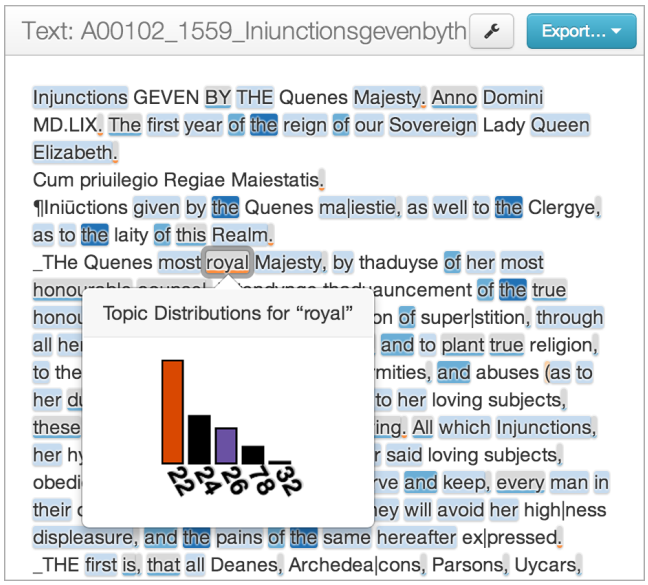

Figure 2: In TextViewer, topic assignments are indicated using colored tags. Probabilities can be shown using varied-hue color ramps. Individual words' topic distributions are displayed within popups.

3 Conclusion

As part of a project to understand the development of English language literature after the introduction of print, our collaborators used Serendip to explore a corpus of 1080 documents sampled from 1530 to 1799. Already familiar with the dataset, they were able to quickly spot unexpected topic occurrences in CorpusViewer, especially when documents were aggregated by metadata such as genre and decade. The real advantage of Serendip's workflow was evident in our collaborators' ability to make comparisons amongst these aggregations in MesoViewer and drill down to explanatory passages in TextViewer. These lower level views provided the context and examples needed to make sense of high level trends.

In addition to the literature domain, we used Serendip to explore models built on corpora of online hotel reviews and news articles. In the future, we will examine Serendip's utility within other disciplines and extend its use to more general statistical models of text.

Acknowledgements

This work was supported in part by NSF award IIS-1162037 and a grant from the Andrew W. Mellon Foundation. Thanks to Ce Zhang for topic modeling support and Mattie Burkert for corpus curation.

References

[1] J. Bertin. Semiology of graphics: diagrams, networks, maps. 1983.
[2] D. Blei et al. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
[3] J. Chuang et al. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. 2012 ACM Human Factors in Computing Systems*, pages 443–452. ACM, 2012.
[4] M. Greco et al. On the portability of computer-generated presentations: The effect of text-background color combinations on text legibility. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(5):821–833, 2008.
[5] S. Liu et al. Interactive, topic-based visual text summarization and analysis. In *Proc. 18th ACM Conf. Information and Knowledge Management*, pages 543–552. ACM, 2009.
[6] F. Moretti. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso Books, 2005.
[7] H. Siirtola. Interaction with the reorderable matrix. In *Proc. 1999 IEEE Int. Conf. on Information Visualization*, pages 272–277. IEEE, 1999.