Serendip: Turning Topics Back to the Text

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Gleicher
University of Wisconsin-Madison

The problem

Topic modeling---an increasingly popular method of large-scale textual analysis being employed in many fields across research and industry---is ripe for visualization tools that will make it more accessible. Some such tools already exist. However, efforts to visualize the output of these algorithms tend to focus on analyzing the *model itself*, distinct from the documents upon which it was trained. While these visualizations can be useful for forming high-level arguments about statistical patterns and trends, they often offer no way to ground said arguments in the text, as is expected in the rhetoric of many disciplines.

Our solution

We seek to treat the model not as an end in and of itself, but rather as a *lens* through which to view the original documents. Serendip is our prototype application for a multi-level approach to this problem. It is a three-tiered tool that allows readers to peruse a topic model at the level of the entire corpus, a single document, and a specific passage (described to the right).

Why the tiered workflow?

There are three main reasons why we believe it is necessary to have a multiple-tiered workflow that ultimately bottoms out in the text:

1. Though the focus tends to be on corpus-level trends, we believe there are **interesting hypotheses to be explored at each level**.

2. Within many disciplines, **textual examples are a required part of the rhetoric**. A tool that can direct readers to exemplary passages of high-level trends will be better suited to help researchers prove their case to computation-skeptical peers than one that just shows the trends alone.

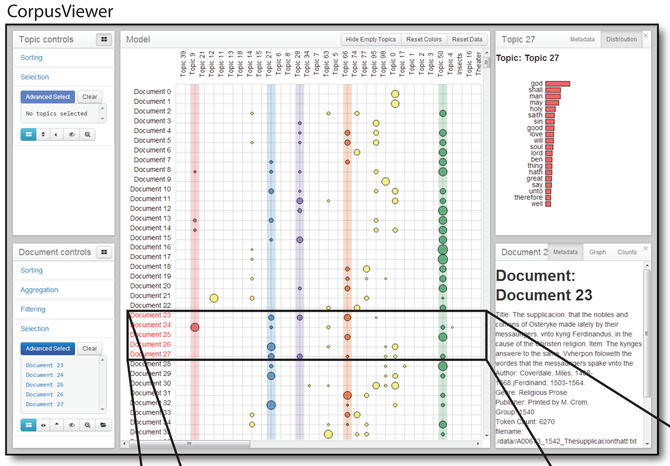
3. Drilling down into the **raw data can help build user trust in the model**. Seeing topic tags in context allows the reader to decide for herself whether she is seeing a hitherto unknown property of the corpus, a junk topic, or a bug in the implementation.

Interaction between views

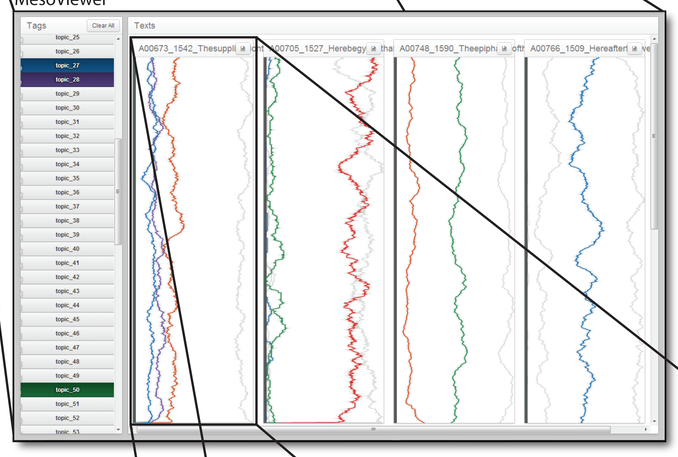
In Serendip, readers can identify documents or sets of documents of interest and open them in lower levels for further inspection. Doing so opens a view in a new browser window, making it easy to refer back and forth between levels—though we provide some redundant information across views so the user need not switch too frequently.

Topics of interest can be annotated with color in CorpusViewer or toggled on in MesoViewer and TextViewer. Toggling or annotating topics in any of these Viewers does so for all open Viewer windows for a given corpus. The colors for these topics match across windows, making it easy to study topics of interest at multiple tiers.

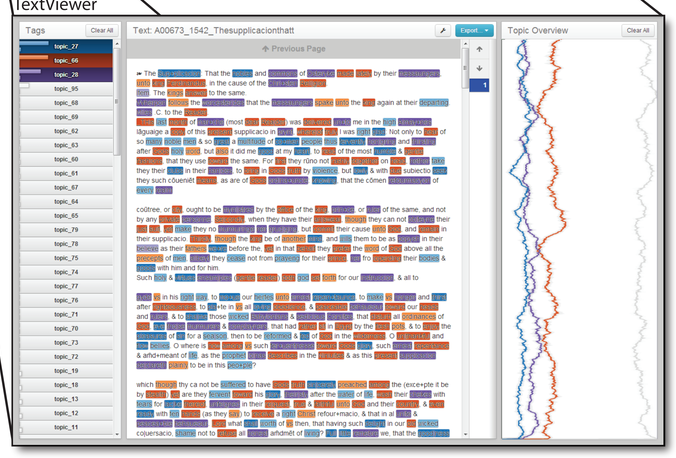
CorpusViewer



MesoViewer



TextViewer

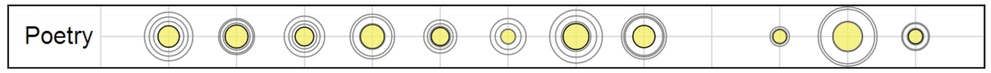


CorpusViewer

Our highest level encoding of the texts lets readers search for **corpus-wide patterns and trends**. Its main component is a reorderable matrix 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. Users are able to sort, filter, aggregate, and annotate this matrix using the statistical properties of the distributions as well as bibliographic metadata like author, genre, or date.

Sorting and Filtering: There are three kinds of data with which can be used to sort the rows and columns of the matrix, or to filter away rows and columns that are not of interest: metadata, both statistical and bibliographic; cosine similarity; and the rankings of the topic distributions themselves. All of these can be applied to both individuals and sets of documents and topics.

Aggregation: To help deal with the scale of large corpora, users can aggregate collections of documents into groups based on bibliographic metadata. When doing so, the circles indicating topic proportions are replaced with glyphs resembling targets, as seen below. The three extra circles in these glyphs indicate the first, second, and third quartiles of the proportions averaged to create the aggregate glyph.



Details on Demand: Windows to the right contain multiple tabs providing additional detail to the reader about selected topics and documents of interest. The topic window displays the selected topic's most frequent words, as well as statistical information about its distribution throughout the corpus. The document window shows bibliographic metadata, the document's topic distribution, and a line graph displaying changes in topic densities throughout the document (as seen in MesoViewer).

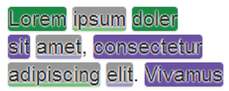
MesoViewer

At the intermediate level, we seek 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 relationships 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.

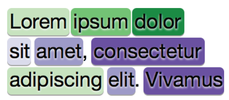
TextViewer

To enable **close reading of individual passages**, we annotate the raw text with data from the statistical model using colored backgrounds to highlight individual words. This sort of "tagged text" displays allows us to adorn the text with additional information without sacrificing readability. In our encoding, we use tags to indicate a word's predicted topic.

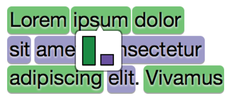
The Tags: The data we derive from the topic model do not generate a single topic for each word, but rather a *distribution* across topics: words are potentially labeled with overlapping tags, each with an associated probability. We provide a number of options for displaying these probabilities:



Overlapping tags: If a word has multiple potential topics, this option uses a gray background to indicate the overlap and underlines words with their most probable topic's color.



Ramped Tags: This option indicates topic probabilities using single-hue color ramps, giving greater perceptual weight to the tags about which the model is most confident.



Distribution on demand: Clicking on an individual word displays a popup showing its full topic distribution.

Navigation: The line graph on the right displays the relative densities of topics across the text. Clicking the graph navigates the reader to the corresponding section of text, making it easy to find exemplary passages of observed topical trends.

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.