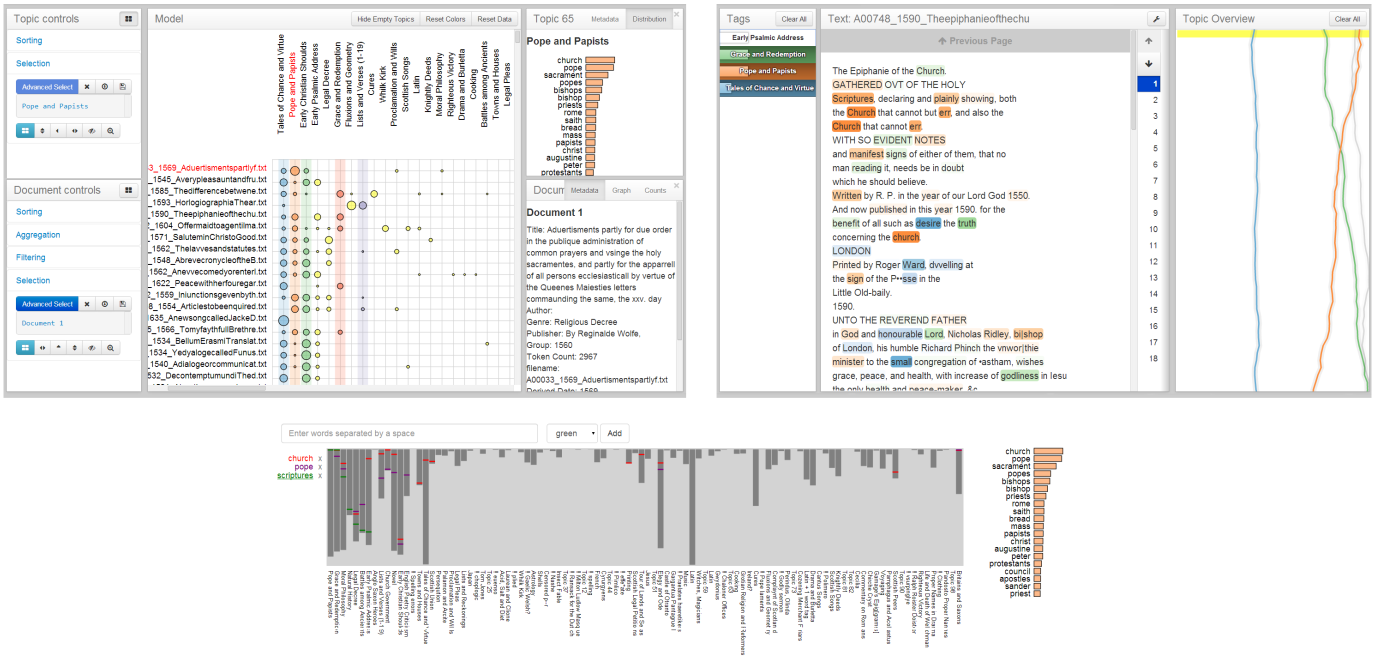
Serendip: Topic Model-Driven Visual Exploration of Text Corpora

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher, *Member, IEEE*



Document 1

Title: Aduertisments partly for due order in the publique administration of common prayers and vsinge the holy sacramentes, and partly for the apparrell of all persons ecclesiasticall by vertue of the Queenes Maiesties letters commaunding the same, the xxv. day
Author:
Genre: Religious Decree
Publisher: By Reginalde Wolfe,
Group: 1560
Token Count: 2967
filename:
A00033_1569_Aduertismentspartlyf.txt

The Epiphanie of the Church.
GATHERED OVT OF THE HOLY
Scriptures, declaring and plainly showing, both
the Church that cannot but err, and also the
Church that cannot err.
WITH SO EVIDENT NOTES
and manifest signs of either of them, that no
man reading it, needs be in doubt
which he should believe.
Written by R. P. in the year of our Lord God 1550.
And now published in this year 1590. for the
benefit of all such as desire the truth
concerning the church.
LONDON
Printed by Roger Ward, dvvelling at
the sign of the P••sse in the
Little Old-baily.
1590.
UNTO THE REVEREND FATHER
in God and honourable Lord, Nicholas Ridley, bi|shop
of London, his humble Richard Phinch the vnwor|thie
minister to the small congregation of •astham, wishes
grace, peace, and health, with increase of godliness in Iesu

Fig. 1. The three main views of Serendip: CorpusViewer, TextViewer, and RankViewer.

Abstract— Exploration and discovery in a large text corpus requires investigation at multiple levels of abstraction, from a zoomed-out view of the entire corpus down to close-ups of individual passages and words. At each of these levels, there is a wealth of information that can inform inquiry—from statistical models, to metadata, to the researcher's own knowledge and expertise. Joining all this information together can be a challenge, and there are issues of scale to be combatted along the way. In this paper, we describe an approach to text analysis that addresses these challenges of scale and multiple information sources, using probabilistic topic models to structure exploration through multiple levels of inquiry in a way that fosters serendipitous discovery. In implementing this approach into a tool called Serendip, we incorporate topic model data and metadata into a highly reorderable matrix to expose corpus level trends; extend encodings of tagged text to illustrate probabilistic information at a passage level; and introduce a technique for visualizing individual word rankings, along with interaction techniques and new statistical methods to create links between different levels and information types. We describe example uses from both the humanities and visualization research that illustrate the benefits of our approach.

Index Terms—Text visualization, topic modeling.

1 Introduction

Exploration and discovery in large text corpora can be a daunting task. Corpora can easily grow to thousands or more texts, ranging in length from short snippets to long books. The task is further complicated by the range of questions that can be asked of such corpora, broad both in subject (making comparisons across time, genre, author, etc.) and in level of detail (corpus, document, passage, even word). Discoveries must often connect multiple subjects and levels of inquiry. Fortunately, there is considerable information to aid these inquiries. Beyond the texts themselves, there are statistical summaries of content, document metadata, and analysts' explicit and implicit knowledge of the documents and their context. However, mixing these different types of information across scales of inquiry is challenging. The information types, and the existing tools that support their use, generally focus solely on a particular scale.

In this paper, we introduce a topic modeling tool for text exploration that is designed to address the issues of inter-mixing scales of inquiry and information types. Our core idea is that to enable fluent fusion, a system must provide not only a set of views for looking at the data from multiple viewpoints, but also connections between the different types of information allowing a reader to move smoothly across scales, data types, and research questions. To achieve this, we have had to adapt existing views to work with different types of text corpora data, develop new views that address some unmet needs, and introduce statistical methods that help connect between different object types. The resulting system enables users to explore questions about collections of texts, passages within texts, and sets of words that define topics, intermixing these types and scales in their inquiry. We have embodied our approach in a prototype system called *Serendip*.

- *E. Alexander, J. Kohlmann, and M. Gleicher are with the Department of Computer Sciences at the University of Wisconsin-Madison. E-mail: ealexand@cs.wisc.edu, kohlmannj@gmail.com, gleicher@cs.wisc.edu.*
- *R. Valenza is with the Department of English at the University of Wisconsin-Madison. Email: valenza@wisc.edu.*
- *M. Witmore is with the Folger Shakespeare Library in Washington, D.C. Email: mwitmore@folger.edu.*

Our motivations for Serendip evolved while working with literature scholars to enable the use of topic modeling as a tool in studying large historical text corpora. These readers have different emphases in their use of text analysis [14]; however we believe that the lessons we have learned from them apply more broadly. Over a period of months, we collaborated to understand their needs for text analysis, evolving a prototype and refining our understanding of needs. This collaboration included the system's primary developer spending a month at the Folger Shakespeare Library—an international center for literary and historical research—working with potential users to understand how the prototype might be adapted to meet their kinds of inquiry. The work identified four key goals in a text corpora exploration system.

First, the system must address issues in scale. Computational methods like topic modeling are attractive because they enable working with corpora that are too big to read. A system needs to scale to large numbers of large documents (e.g. books and plays). While scaling to large numbers of documents is commonly considered in corpus exploration systems, *long* documents present an uncommon challenge. A system must both present aggregate trends *across* documents, but also guide readers to where these trends are reflected *within* documents.

Second, a system must allow for inquiries across different scales. Inquiry might flow from top down: for example, identifying trends across sets of documents leads to identifying specific documents and passages that support them. This is critical for literary scholarship, where arguments are frequently won or lost on the basis of close analysis of exemplary passages rather than the distribution of patterns and charts. But it is also valuable in other domains, where readers must check to see if the statistical patterns really reflect the meaning of the passages [14]. Inquiries might also flow bottom-up, starting with a word or topic of interest and seeing how these things make up larger patterns across documents. Interaction between levels is important: a question about one scale inevitably leads to questions and answers at other scales. A system must provide clear starting points for exploration, as well as ways of using intermediate results to build to next steps at potentially different levels.

Third, a system must allow readers to pull together multiple sources of information, both statistical as well as human-curated metadata. Some of the metadata is explicit (e.g. the date of a book), while others is implicit knowledge of the reader. Inquiries often mixed these different types of information.

Finally, we wanted to promote *serendipitous discovery*: the act of finding something unexpected while looking for something else (or nothing in particular). There is a long history of thinking about serendipity as an expression of individual luck, statistical chance, or divine predestination, yet research points to practical ways in which it can be consciously fostered [36]. This sort of design can be seen in situations as commonplace as where to put books on the shelves of a library, but is relatively unexplored in visualization.

We have built a system called Serendip that addresses these goals, providing three tightly coupled main views (see Figure 1). It extends a reorderable matrix view with novel features that allow it to better address issues of scale, as well as to allow for multi-scale and multi-information fusion. For example, we introduce new orderings that connect between different data and new visual encodings that allow for effective aggregation. It uses a tagged-text-plus-overview encoding, adding features that direct users to key passages and convey the probabilistic nature of topic tags, improving the quality of user interpretations as well as their understanding and trust in the model. It provides a novel view of topic words designed to address specific questions that arise and connecting these inquiries to other scales. All of these are combined with interaction techniques that allow readers to follow branching paths of inquiry across multiple scales and units of analysis. We provide example use cases of Serendip, both on a visualization data set chosen to demonstrate features, as well as a real data set with explorations driven by a literature scholar.

2 Related Work

There is much from the field of text corpus visualization—particularly topic model visualization—that influenced our techniques and design.

2.1 Topic modeling

Topic modeling is a type of text processing that determines major themes of a collection of texts through statistical analysis [4]. While there is a broad and evolving range of available techniques, most produce results of a similar form: topics are represented by sets of commonly occurring words, allowing for documents to be assigned to topics by considering the words they contain. Most topic modeling techniques are probabilistic, so the assignments produced are weighted.

The most common topic modeling algorithm is Latent Dirichlet Analysis, or LDA [4], and there are many available LDA implementations. The work described in this paper is designed to allow for different types of topic models, although we only demonstrate it with LDA models. Similarly, our system is designed to be compatible with a variety of topic modeling tools. The examples in this paper were all created using the open-source Mallet software [28], but Serendip has also been used with models built by other tools.

In this work, we view topic model construction as a separate step. A topic modeling tool is used to construct a model from a corpus as a pre-process. A second, Serendip-specific pre-process combines the model data and the corpus to pre-compute much of the data used in Serendip. Once these pre-processing steps are complete, the Serendip tool is used to explore the processed corpus and model. In the future, we hope to integrate the model construction and exploration steps to better allow for tuning and adapting models.

2.2 Topic model visualization

The nature of probabilistic topic models makes them difficult to interpret, and the need for visual tools has been identified before [3]. In particular, the fact that the data tend to be noisy and variable makes direct interpretation difficult: indeed the strength of the models is that inferences are often built by combining many small things. Another issue is the range of tasks in working with topic models, ranging from evaluating and tuning models to observing trends in topics to finding thematically similar documents.

Most tools for topic model visualization focus on specific tasks and questions by providing specialized views. For instance, Dissertation Browser [9] uses models built on PhD dissertations to track inter-department collaboration. Other techniques are primarily concerned with tracking topic evolution through time, including [37, 22, 16] which use "river flow" layouts. In our work, we have sought to provide flexibility in the range of inquiry supported through the use of multiple linked views. Termite [8] is a tool for understanding topic models *themselves*, not using them as *tool* for exploring the corpora. While we draw several ideas from Termite, including the reorderable matrix of dots and word salience computations, our approach extends it significantly in order to work with the model and the corpora, allowing for multi-scale explorations.

A broad range of work has considered using visualization to explore text corpora beyond topic models. A common strategy is to abstract the texts as glyphs and position them in 2D as a scatterplot. Numerous approaches for organizing these layouts exist (see [30, 24, 20]), with recent work on focusing user control [19] and understandability [7, 21]. We apply the idea of flexible layout as a mechanism for using topic model data.

Existing tools have identified *individual* documents as an important unit of study as well—though rarely the same tools that visualize documents at the corpus level. Within most topic model visualization tools, single documents are either inaccessible or viewable only as plaintext. This is generally sufficient, as model corpora typically contain documents on the scale of abstracts, which can be easily skimmed. When modeling much larger documents like books, additional information from the model is needed to direct the user through the document's structure. (The motivation for this is described further in §3.) Others have employed tagged text displays for such overlay of information [14, 15], and they have been shown to allow users to make aggregate judgments without sacrificing readability [13]. Plaisant et al. have used colored tags to indicate metadata and user interest [32, 10]. Whereas such tags are typically binary in nature, we have applied them in a way that interactively conveys the probabilistic uncertainty of a topic model, reflecting not only which words belong to which topics, but which words are *important* to those topics (see §5, 6).

Most importantly, analysis of such corpora tends to be coordinated dynamically across levels—unsurprising, since language works in an integrated fashion, with small features contributing to broader narrative functions. Our tool needed to recognize and focus user interest upon that vertical integration of levels. Jigsaw [35] is one of few text visualization tools to consider multiple levels of inquiry. At distant levels, it relies on metadata and user selection; at closer levels, it provides plain text juxtaposed with statistics. In contrast, our approach fuses multiple sources of information (particularly topic models), allows and annotates longer texts in the detailed views, and has explicit support for inquiries that move from details up to broader summaries. PaperLens [27] similarly combines multiple views of clean metadata, but does not consider the challenges of topic models.

2.3 Fostering serendipity

The word "serendipity" was coined by Horace Walpole, referring to a story called "The Three Princes of Serendip," in which the protagonists are able, through chance observation, to uncover the nature of a camel they have never seen. The word has come to be associated with occurrences of happy accidents that lead to unexpected discovery. Though such instances are often attributed to fate or providence on one hand or extreme cleverness on the other, research has been done to determine how to make such "accidents" more likely. The most often used example is that of looking for a book in a library: a patron navigates the stacks to find a particular book, but because of the way books are organized, they end up stumbling upon an even better book sitting on the same shelf. Thudt et al. [36] provide a thorough survey of the research on promoting serendipity, and distill it into a concise set of principles that apply to the design of visualizations:

1. **Providing multiple access points.** Unlike physical books on a shelf, electronic documents can be arranged in many ways simultaneously. Users can make a more diverse set of findings if they can view the data from a variety of different angles.

2. **Highlighting adjacencies.** Serendipitous findings tend to occur *near* where the user is searching, and so it is important to visually emphasize these proximities.

3. **Offering flexible pathways for exploration.** While many systems offer data access through directed querying, encouraging open-ended exploration—with a variety of viewpoints and transitions between them—seems to enhance serendipity.

4. **Enticing curiosity and playfulness.** Finally, even when presented with surprising juxtapositions, the user must be in a creative state of mind to be able to make connections between them. An experience that engages their sense of fun has been shown to promote this state of play and exploration.

Thudt's Bohemian Bookshelf system, based on these principles, helps users find books in a library. We have aimed to use the principles in our approach for text exploration.

3 Exploring text corpora with Serendip

Our approach is designed to combine tenants of serendipity and multi-level exploration, dealing comprehensively with issues of scale. We incorporate three main views, each designed to serve as an access point to the data and support a different level of inquiry. At the *corpus level*, we provide a reorderable matrix to highlight adjacencies between documents and topics.At the *document level*, we use tagged text and overview displays to help readers find and analyze passages in large documents. Finally, at the level of *individual words*—a level we only observed the need for after watching users interact with our text level tool—we introduce a ranking visualization that shows how words are distributed across the topics.

Ultimately, it is the interactions *between* these levels that provide "flexible pathways for exploration" as laid out in the above principles of serendipity. There are many possible units of interest within the corpus: topics, documents, metadata, passages, words. A user may find him or herself entertaining one (or some) of any of these, either as their initial entry-point to exploration or as an intermediate step along the way. To provide for flexible information usage, we offer techniques for using any of these units to identify other units of interest, of potentially different types. For example, documents (or sets of documents) can be used to find other documents, metadata categories, topics, passages, or words. Providing linkages between the different combinations of units requires an array of different visual, interaction, and statistical techniques. User inquiry must be allowed to move across these levels in a flexible, sustained way; our users need access to multiple starting points and control over their own successive re-orientations, building inquisitive momentum as intermediate results drive next steps. These linkages are described in the following sections.

The centerpiece of the corpus level tool, CorpusViewer (§4), is a re-orderable matrix that connects documents to topics. To address the "many documents" and "many topics" issues of scale, the matrix supports filtering and selection, aggregation, and ordering. Ordering is a key tool as it not only addresses scale—by placing salient objects at the top—but also promotes serendipity by placing similar objects next to each other. CorpusViewer focuses on exposing patterns across documents and topics, and identifying specific items to explore more closely. Additionally, it provides ways to overlay other information types, such as metadata and words, and to link to other views.

TextViewer (§6) allows for detailed examination of how topics are reflected within a specific document. A tagged text visualization shows the topics and the text. To support long documents, a summary graph shows how the topics occur over the length of the document. This view's main role is to connect high-level trends to specific example passages, validate them for the user, and help build the user's understanding in the workings of the model. However, it is also important for identifying topics and words to explore in other views.

RankViewer (§7) allows users to examine specific words and see which topics use them. This tool is useful for relating topics and words, and comparing different topics and words. It can provide topics (and orderings of topics) to explore more closely in other views. Central to viewing words in topics (in both RankViewer, but also CorpusViewer) is a mechanism for ranking words (§5).

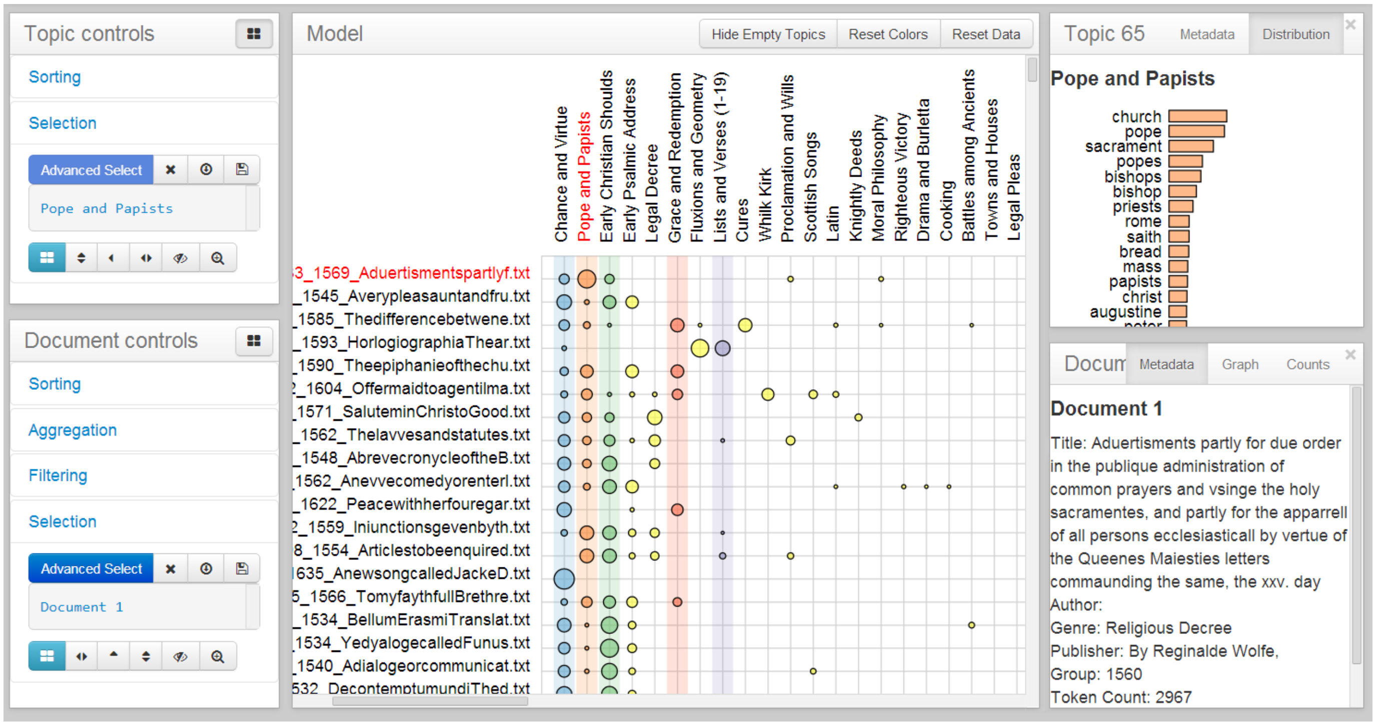
4 Viewing the corpus

CorpusViewer provides a high-level overview of the entire corpus. It is designed to help identify trends in documents and topics, and to use them to focus on more specific items (or sets of items). Its main view is a reorderable matrix that plots documents (rows) against topics (columns), encoding the values of the distributions as circular glyphs on the vertices of the grid (see Figure 2). We have supplemented this matrix with features to combat scale, connect outside information, and promote serendipitous discovery at the corpus level.

4.1 Filtering and selection of data

A simple but important way to combat the potentially vast dimensions of this matrix is to make sure users are able to focus on objects of interest. We provide a query-system that allows users to pick out documents and topics based on their metadata. Once selected, these sets can be hand-tuned, colored, moved to a more prominent position in the matrix (typically the top-left corner), used as a basis for reordering the matrix as described in §4.2, or saved to be explored later.

Selections (and set building) can also be done manually. This is sometimes useful for removing erroneous rows and columns from a query. More importantly, the ability to build sets provides a way for the user to express knowledge (e.g. of a known set of objects of interest), and to use intermediate results of prior steps to make new steps (e.g. using the top elements of a sorting, or surprising anomalies, to create a new ordering). First class selection sets are supported for both documents (rows) and topics (columns).



Document 1

Title: Aduertisments partly for due order in the publique administration of common prayers and vsinge the holy sacramentes, and partly for the apparrell of all persons ecclesiasticall by vertue of the Queenes Maiesties letters commaunding the same, the xxv. day

Author:

Genre: Religious Decree

Publisher: By Reginalde Wolfe,

Group: 1560

Token Count: 2967

Fig. 2. CorpusViewer centers around a re-orderable matrix that provides a variety of ordering, selection, aggregation and annotation features to help users find high-level patterns in the corpus and connect to specific documents and topics. Each row represents a document, each column a topic, and the circle size encodes the proportion. Here, colorings are applied to selected columns in order to connect to other views of topics. In the upper right, a topic is depicted by showing the proportions of its most salient words.

4.2 Reordering the data

Reorderable matrices have been around since they had to be rearranged by hand [2]. While many try to find the "optimal" order of rows and columns in a matrix [11, 23], people have been shown to find interesting attributes and patterns within the data if they are given direct control of the orders themselves [34]. Embracing this idea, we have created a number of ordering options designed to address the requirements of our tool.

Good orderings combat scale, concentrating the most salient items at the top or bottom of the list. Good orderings also promote serendipity by putting similar objects next to each other and providing different ways of looking at the objects. We identify three different types of orderings: **blind orderings** that just use the distributions, but are useful for starting an inquiry; **question-based orderings** that use other sources of information (such as sets of other units or metadata); and **similarity-based orderings** that order based on similarity to a seed set of like units. These orderings can serve as starting points of inquiries (with or without a specific question or object in mind), or serve to use an intermediate result or finding as a point for further inquiry. In our prototype tool, we have provided orderings of each type for both the rows (documents) and columns (topics) of the matrix.

4.2.1 Document orderings

The number of documents (and therefore rows) contained within a corpus is perhaps the biggest scaling issue in these models, and therefore good techniques for document ordering are crucial.

Blind: As a blind entry-point into the documents, we offer a way of sorting by their topical complexity. This was inspired by [17], which used a form of entropy to distinguish between documents that contain single topics and those that contain multiple topics. Rather than using entropy, we allow the user to sort documents by the proportion of their n^{th} strongest topic. This can provide an approximation of topic entropy. For example, sorting by the third highest topic proportion finds documents with (at least) three strong topics. Sorting by small numbers of topics (especially 1) finds documents that are strongly dominated by a single topic.

Question-based: Question-based orderings attempt to answer user queries by pulling the most relevant documents to the top of the list, where relevancy can be based on information other than the documents. The most commonly used is ordering based on the strength of a selected topic (or set of topics) of interest (e.g. "Which documents are highest in topic *x*?"). Documents can also be sorted by metadata fields, which can be useful for exposing topic trends in a particular group of documents or for aiding in search. (Some of these same tasks can also be achieved by using metadata for searching, filtering (§4.1), and aggregation (§4.3).)

Similarity: Creating orderings of documents by similarity of their topic data—e.g. finding documents that resemble an example document or set of examples—is a particularly important ordering tool. It provides an initial entry point when the user has familiarity with even a few of the documents, but also provides a follow-up step when interesting documents are found. There are two key questions in building similarity metrics: how to compare document vectors, and how to calculate distance to a group.

The documents are represented as vectors of their topic proportions. In general, meaningful distance metrics in high-dimensional spaces are challenging (see [38] for a survey of the issues). Document vectors are also sparse and (nearly) convex (discounting truncation issues). By convention, we generally use the cosine similarity metric rather than Euclidean distance, but both metrics are mainly meaningful when documents are close: as documents become farther apart, the metrics become less interesting. We use Euclidean metrics for performing clustering. Our system also provides *weighted* variants of similarity metrics, where different topics are weighted differently. The weights can come from various forms of topic selection and ordering criteria (below), providing a simple form of distance metric adaptation.

The primary way that our system measures distance to a set of documents is by computing the "center" of the set (by computing the mean), and measuring the distance to this point. This approach has two key flaws: first, averaging vectors damages the sparse and convex structure; second, the center may not adequately capture a multi-modal (or oddly shaped) distribution. To combat these issues, we provide an alternative. We generate multiple centers for the set by k-means clustering, and define distance as the minimum to any one of the centers. As a special case, choosing the number of clusters equal to the set size guarantees that items in the set have zero distance to it. The multi-center approach, in principle, improves performance because the distances are smaller and therefore are better approximations. In practice, the errors in the single center approach often create serendipitous accidents: while the closest documents to a set are often not in the set, they are nonetheless often interesting and/or surprising. When computing distance to a center, we can use the variance of the set to provide a weighting so that higher variance topics contribute less.

4.2.2 Topic orderings

While there are generally considerably fewer topics than documents, it is still often impractical to scan through the entire list. Column orderings are also important for promoting serendipity: not only for identifying other topics of interest, but for seeing that documents combine multiple topics in different ways. Column orderings can also be useful for suggesting smaller sets of relevant topics so as to have a more focused distance metric for sorting rows.

Blind: For the topics, we offer a variety of blind entry points based on statistical metadata. Most prominently used among these is sorting the topics by the number of documents containing them, giving the user a sense of the most prevalent topics in the model. Another valuable ordering is the variance of a topic's proportions when it is present within a document: does it tend to dominate documents, or is it more briefly mentioned? Other blind orderings include maximum proportion, minimum proportion, and mean value.

Question-based: The most common tool for ordering topics is based on a document (or set of documents). The topics are sorted by proportion in the document (or average in the set).

A second set of question-based orderings pulls in outside information in the form of document metadata. These orderings use statistical measures to determine how well different topics correlate with metadata distinctions. The more general of these orderings is the ANOVA ordering. This tool uses a categorical metadata element (such as genre) and performs a one-way ANOVA for each topic to determine how likely the different categories are to have different mean values. Sorting by the F-value ranks the topics by their ability to distinguish the different categories. A second such ordering tool is contrast ordering. This tool takes two sets of documents and computes the t-statistic for each topic, testing that the two sets have different mean values. Again, ranking by this statistic orders the topics by how well they distinguish between the two classes.

We note that topic data does not meet the assumptions for the statistical tests applied to produce orderings. However, since our goal is to assess the relative values for ranking, rather than using the precise values to determine significance, we feel these approximations are justified. Alternate statistics, such as the non-parametric Kruskal-Wallis H, could be applied instead (see §10).

Similarity: Much like documents, topics can be sorted by their similarity or weighted similarity to a particular topic or set of topics. The metrics compare how the topics are used in the documents, rather than the words they contain. By default, cosine distance is used to compare vectors; however, an alternative uses the Spearman rank correlation coefficient to measure how similarly the topics rank the documents. In practice, these seem to provide similar results.

4.2.3 Other Proximity Displays

Ranking, often using distance metrics, is an important method for creating serendipity by putting similar things next to each other in a list. However, the confines of a 1D ranking may not adequately capture nearness, and other visual encodings of similarity may promote serendipity in different ways. For these reasons, our system also generates scatterplots of two dimensional embeddings of the distance functions. The default is to use a spectral embedding as it captures the near-neighbor behaviors that are most likely to be interesting, and ignores larger distances that are less likely to be meaningful. Nonlinear manifold embeddings, such as IsoMap, have similar properties. Our implementation uses a standard library (scikit.learn [31]) that provides a number of embedding techniques. Scatterplots are colored by metadata, as shown in Figure 7.

To create a very different view of proximity, our system performs a k-means clustering and presents the results with each cluster being a list ordered by distance to the cluster center. This view emphasizes neighborhoods of similar objects to promote the serendipitous discovery. It also provides a sense of the diversity in the corpus, as the cluster centers provide a sampling of dissimilar documents.

4.3 Aggregating the data

While a variety of ordering metrics combats scale by concentrating important documents together, this is not always enough, especially when trying to compare *groups* of documents. This connects to our goal of pulling outside information into the analysis: comparing collections of documents—especially those grouped by categorical metadata like genre or conference, as seen in §9—is a very common use case. We enable such comparison by allowing the user to aggregate documents into sets based on arbitrary fields of metadata. This can dramatically reduce the size of the matrix to be explored.

When aggregating, we average document rows into single vectors that display the mean value of each topic's proportion using filled circular glyphs. Our encoding also reflects the *variance* of topic proportions within these groups of documents. On top of the filled circles, we add three thinner, unfilled circles that encode the first, second, and third quartiles of the aggregated values (see Figure 3). In other words, if a set of documents varies dramatically in its proportion of a particular topic, that glyph will resemble a bulls-eye of concentric circles. If the documents all share similar proportions of the topic, the concentric circles will fall roughly on top of one another, approaching a glyph that looks like a single circle.

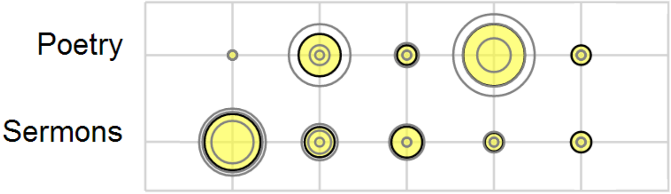
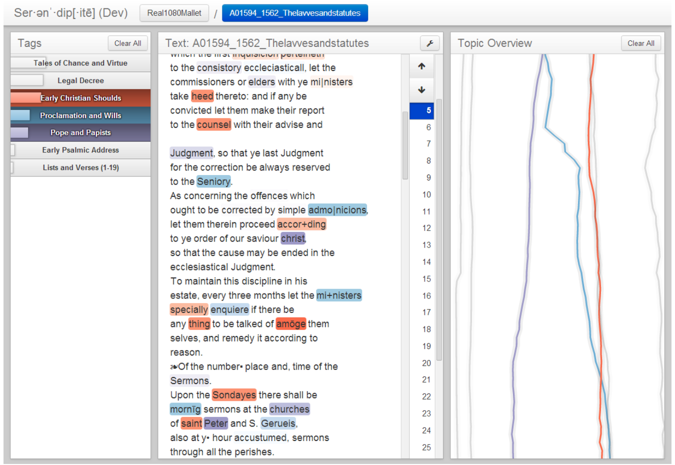


Fig. 3. When aggregating documents, CorpusViewer uses glyphs that give a sense of both mean and variance. The filled circle represents the average proportion for a given topic for all documents in the aggregation. The three non-filled circles indicate the first, second, and third quartiles for this topic's proportions across the documents in the aggregation.

4.4 Other features for exploration

There are a number of features for annotation within the tool. The user can label documents with any field of metadata; title is a common choice. While there is no such associated metadata for topics, the reader can rename topics with arbitrary strings of their choosing (see Figure 2), creating meaningful labels based on their observations across words and documents. By combining multiple sources of information, users can come up with names that are much more descriptive and interesting than what can be generated algorithmically (see §9.2). Readers can also assign colors to sets of documents and topics (see Figure 2), either manually or by queried selection. These colors are retained across all levels of Serendip (see §6, 7).

CorpusViewer also provides extra details on demand in the two windows on the right that give statistical information and metadata about selected topics (top) and documents (bottom) (see Figure 2). The topic view is particularly useful for viewing the topic's highest ranked words, as described in §5. Finally, these windows act as jumping off points into the other levels of Serendip. Double-clicking the document's heading will open the document within a new TextViewer



which the first inquisicion perteineth
to the consistory eccleciasticall, let the
commissioners or elders with ye mi|nisters
take heed thereto: and if any be
convicted let them make their report
to the counsel with their advise and

Judgment, so that ye last Judgment
for the correction be always reserved
to the Seniory.
As concerning the offences which
ought to be corrected by simple admo|nicions,
let them therein proceed accor+ding
to ye order of our saviour christ.
so that the cause may be ended in the
ecclesiastical Judgment.
To maintain this discipline in his
estate, every three months let the mi+nisters
specially enquiere if there be
any thing to be talked of amōge them
selves, and remedy it according to
reason.
❧Of the number• place and, time of the
Sermons.
Upon the Sondayes there shall be
mornīg sermons at the churches
of saint Peter and S. Gerueis,
also at y• hour accustumed, sermons
through all the perishes.

Fig. 4. TextViewer combines tagged text with a line graph overview for navigation. The line graph can be used to navigate to passages with varying densities of topics. The tagged text is ramped so that words with higher ranking (see §5) are darker and more salient. The column on the left displays allows toggling topics on and off.

window (§6) while clicking on a particular word in the topic view will display that word's rankings within a new RankViewer window (§7).

5 Representing Topics

Representing topics as a list of words requires a good metric for ranking them, as only the topmost words can be displayed at once. Both CorpusViewer (§4) and RankViewer (§7) show lists of top-ranked words, and the metric determines color-ramping in TextViewer (§6).

The most commonly used ranking metric is frequency: the percentage of a given topic accounted for each word. This has a distinct bias toward words appearing in *many* topics. Models typically factor out pervasive stop words such as articles and pronouns, but the most frequent words in a topic are often still pervasive enough to be rather uninformative for *distinguishing* topics.

The other extreme is to rank words within a topic by the information they gain toward identifying that topic. Information gain can be computed by using the Kullback-Leibler divergence [26] between the probability an arbitrary word k was generated by topic T *given* $k = w$, and the probability k was generated by T *without* that knowledge. Intuitively, ranking words by information gain in this way will pick out words that best distinguish the topics. However, this metric has a large bias toward very rare words that appear only a handful of times within the corpus, and are therefore uninformative in their own way.

We create a metric that combines the benefits of both frequency and information gain by multiplying them together. This *saliency* metric is similar to that introduced by Chuang et. al. for finding salient words across an entire model, not just within a topic [8].

6 Viewing documents

TextViewer not only allows the viewer to see specific documents, but also to see how various topics are reflected within them. This lets well known passages serve as entry points to the model (by suggesting topics for exploration at a higher level) as well as being a way of providing exemplary passages for high level trends. The need to trace trends down to the passage level is particularly prevalent among humanities scholars, for whom textual examples are a required part of their rhetoric [14]. However, providing low-level examples can also help readers in other domains, both to explain high-level trends and to build trust in the model [14].

6.1 Intra-document navigation

Topic model visualizations that *do* give access to the documents typically present the raw text, unannotated. This is often sufficient since the documents being modeled are on the scale of abstracts. If the model assigns a particular topic to a given document, abstracts can be quickly skimmed as a sanity check. Such is not necessarily the case when modeling documents on the order of novels and books. Just as themes and subject matter will come and go throughout the course of a story, so do the occurrences of a topic vary in density. As such, readers may require a navigational aid to find exemplary passages of high-level topical trends within longer documents. We use overview visualizations to direct readers in this manner.

A variety of existing techniques for representing document structure were more complicated than necessary for the task [33, 25, 12]. Our overview is simple: a line graph displaying densities for each topic with adjustable smoothing. These graphs can show many topics at once, with users being able to toggle topics on and off, giving "on" topics a qualitative color encoding (pulled from ColorBrewer [6]) that is consistent across the different levels of Serendip—making comparisons of topic trends perceptually clear and affording smooth transition of the user's exploration across levels.

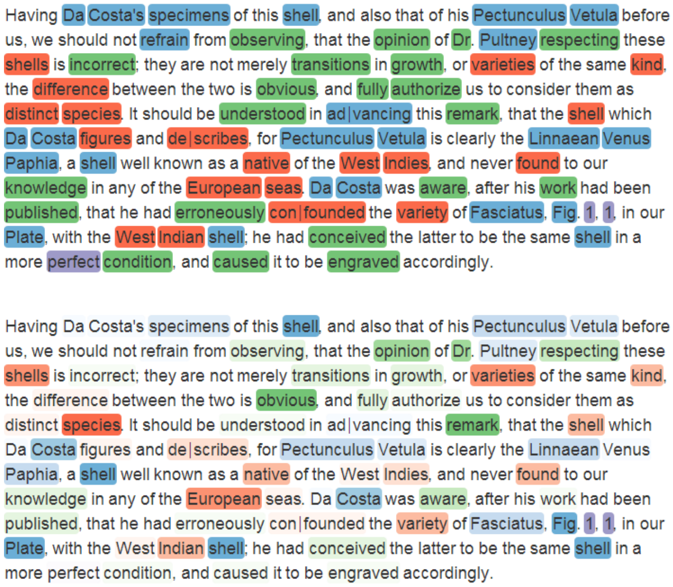
In TextViewer, these graphs become a useful aid to navigation within a document. It is easy to determine from the peaks and valleys of the topic lines which passages are high or low in a topic, which contain a *mixture* of topics, etc; by clicking a particular position on the overview, the reader is able to easily scroll to any passage of potential interest within the pages of the document.

6.2 Text tagging

Once the reader has arrived at a passage, their question becomes: which words matter? Providing the raw text offers some utility, but we can provide more by annotating the text with data from the model. In TextViewer, we do this using colored backgrounds to highlight individual words. Since LDA labels each word with a topic, the easiest approach would be to just assign each topic a single color. However, tagging *all* of the words in a document—even discounting stopwords—tends to result in displays that are overwhelming and often uninformative (see Figure 5). Worse, tagging all words equally can sometimes be *negatively* informative. For instance, there may be some words that have too low a frequency for LDA to "know what to do with," yet must get *some* tag, likely just the most common one around them. Researchers accustomed to the practice of close reading may read too much into these relatively less "meaningful" tags. As such, we need to sparsify the display to give greater perceptual weight to words that the model deems more "important." And for those readers who are particularly interested in individual words, we need to give them an idea of what *else* any given word might relate to.

There are many ways to define which words are important. Within the toggled-on topics, we deem the "importance" of a given word to be its ranking within its topic, using whatever ranking scheme is currently enabled (saliency being the default—see §5). We divide words into bins along a single-hue ColorBrewer ramp based on their ranking, giving darker tags to higher ranked words and vice versa. On a white background, this has the double benefit of drawing user attention to meaningful words as well as greatly sparsifying the visual display, making the text easier to read (see Figure 5).

Apart from letting readers focus on the most salient words, this method of tagging also conveys the inherent uncertainty associated with probabilistic methods like topic modeling. This is often difficult for readers to accept. Such was our observation from our work with humanities scholars. Our collaborators would often focus in on one surprising word, perhaps exclaiming: "Why is *that* in Topic 3?!" Sometimes the answer might be meaningful: that word is actually associated with others in Topic 3 in an interesting and surprising way. However, the answer might not be meaningful. Maybe the word appears in *every* topic at some point, or perhaps it is seen so infrequently that the model did not have enough context to informatively tag it, or it may just be the luck of the draw. By using saliency-based color ramps, readers focus much more on the *meaningful* words, as determined by the model. Additionally, conveying this element of uncertainty within the model lets new users begin to appreciate the inexact nature of the algorithm. This decreases their tendency to accept each aspect of a model as gospel, or to dismiss entire models out of hand when they



Having Da Costa's specimens of this shell, and also that of his Pectunculus Vetula before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in ad|vancing this remark, that the shell which Da Costa figures and de|scribes, for Pectunculus Vetula is clearly the Linnaean Venus Paphia, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously con|founded the variety of Fasciatus, Fig. 1, 1, in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

Having Da Costa's specimens of this shell, and also that of his Pectunculus Vetula before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in ad|vancing this remark, that the shell which Da Costa figures and de|scribes, for Pectunculus Vetula is clearly the Linnaean Venus Paphia, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously con|founded the variety of Fasciatus, Fig. 1, 1, in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

Fig. 5. Top: Traditional color-coded tagging creates an overwhelming view that is difficult to read and more difficult to interpret correctly. Bottom: Using our system of ramped tags, the most important words stand out, and the entire passage is easier to read.

exhibit some unexpected property.

Given our readers' predisposition to focus on single words or groups of words, we needed to go one level deeper beyond the level of passage. Clicking on any individual word, then, takes the reader into the deepest level of Serendip, RankViewer, allowing them to see how that word—and any others in which they might be interested—are dispersed throughout the topics (see §7).

7 Viewing Words

Individual words are not typically the focus of exploration within a topic model, but they are frequently the objects of study within the humanities. They can also serve as an accessible entry point to a model within *any* domain. Even barring thorough knowledge of topic modeling, most any researcher will be able to come up with a few words whose behavior they would be interested to track within a corpus.

Single words also offer an additional form of adjacency within the topics, and thus another opportunity for serendipitous discovery. While watching our collaborators use our tools, we saw that their interest was often piqued by "surprising" words—words appearing in a topic the user thinks he or she understands, but which do not immediately fit with that understanding. As described in §6, there are many reasons why such surprises may occur—some interesting and some not—so it is important to filter out the less meaningful ones using saliency ranking. But for the salient surprises, the user's immediate question tends to be "What *else* is that word in?"

RankViewer (Figure 6) was created to answer this question. A simple list of topics containing a word is insufficient, because the saliency of a word within a topic (*where* it falls within the topic) is important for determining its relevancy. Instead, our tool shows where a word—or group of words—appears within topic word rankings using lines in an inverted bar chart. Gray bars indicate the relative size of the topics, which can vary dramatically within a model. Color coded lines within these bars correspond to the ranking of individual words that the user has indicated for analysis, either by clicking on them in TextViewer or CorpusViewer, or by manually searching for them within RankViewer. Topics can be sorted and rearranged based on the prevalence of a particular word or set of words, and clicking on a given topic creates a fuller display of its rankings on the right.

Giving users this level of depth lets them confirm the importance of words within a topic, improving the validity of their interpretations

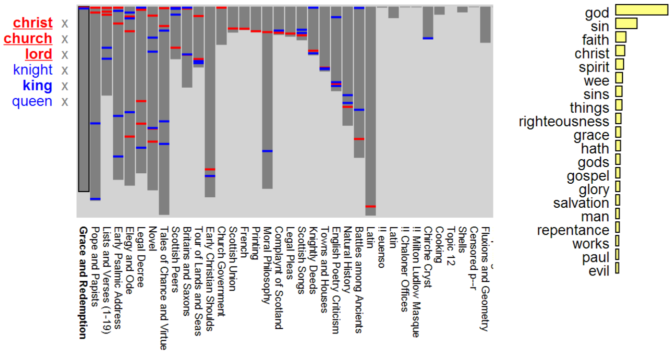


Fig. 6. RankViewer shows where words fall in the rankings of individual topics. Topics are represented by gray bars and can be sorted by any combination of words being searched for (which are underlined). Individual lines indicate each word's ranking within the topics, color coded to match the list on the left. The view on the right displays the top-ranked words of a selected topic.

and strengthening their understanding of the model. By juxtaposing topics in a different way, RankViewer also opens up new pathways for exploration at higher levels. When users see that an interesting word is present in another topic, they can move quickly to that new topic in CorpusViewer and explore it in depth, examining other documents which contain it. This cascading effect is productive, allowing for bottom-up exploration, and ensuring that inquiries don't necessarily "bottom-out" at the lowest (word) level.

8 Implementation

Our prototype for Serendip is web-based to make it easily shareable and accessible to a variety of users. Serendip operates on a back-end written in Python with the Flask framework and a front-end written in Javascript and D3 [5], with Twitter's Bootstrap providing UI elements. Topic model data is stored on the server as CSV files, and texts are pre-processed into paginated HTML. Though the tool is designed to be model agnostic, all of the models described in our use cases were generated using Mallet [28].

9 Use Cases

We describe here some initial experiences using Serendip on various corpora. The first use case was performed by visualization researchers, and is intended to illustrate the capabilities of our techniques on a familiar dataset. The second use case was performed by a literature scholar with some experience using the tool and provides an example of serendipitous findings on real data. Other initial use cases with domain researchers (not reported here) include a collection of over 600 plays, and a large collection of novels.

9.1 Vis Abstracts

To demonstrate Serendip's features, we describe its use on a familiar corpus: a collection of abstracts from select IEEE sponsored visualization conferences from 2007-2013, including SciVis, InfoVis, VAST, BioVis, and PacificVis. We standardized the conference names and used various heuristics to remove bibliographic entries for "non-papers." The corpus consists of 1127 abstracts, ranging from 30 to 389 words. The discussion below is based on a 30-topic model. The findings on a familiar data set allow us to sanity check that system features provide reasonable results.

We started with a common question: are the content differences between the conferences reflected in the model? For an initial look, we chose a 2D spatial embedding (using Spectral Embedding), coloring the scatter plot by conference (Figure 7). This provided a picture with the thematic conferences being relatively distinct, while the general conference (whose topics span the range of the others) is more spread out. To see which topics create the distinction, the ANOVA ranking

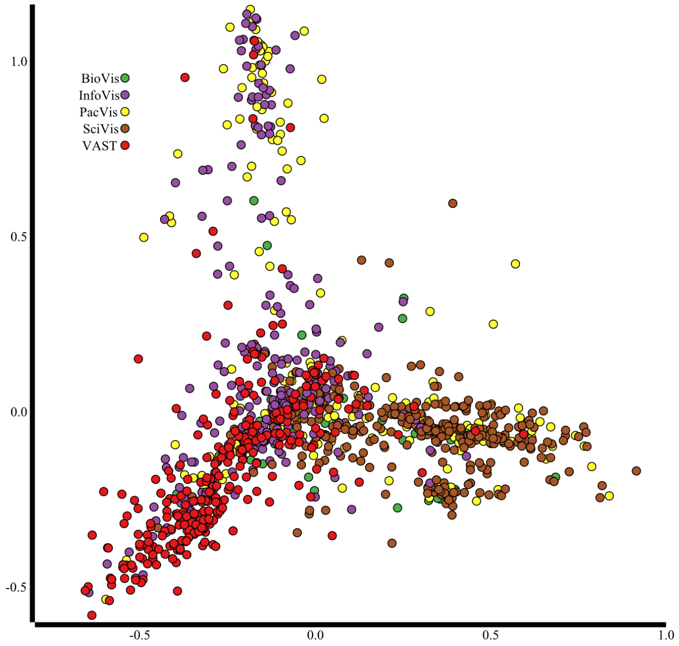


Fig. 7. A scatterplot of an embedding of the documents in the Vis-Abstracts corpus. Spectral embedding was applied to the document vectors. Each point represents a document, and is colored based on the venue of the document. The plot shows, at a glance, that the topic data is capturing some sense of the distinctions in the venues. Venues with more focused themes (VAST, InfoVis, SciVis), tend to group more closely together, while general venues (PacificVis) are more diverse.

was used to order the topics. The top topic has terms *visual, analytics,* and *analysis* among its top terms, and showed that many VAST papers self identify themselves in their abstracts. The second ranked topic also identified VAST papers, using terms related to the VAST challenge contest. The third topic featured the terms *volume* and *rendering.* The lowest ranked topics for distinguishing venue featured generic terms, such as *problem* and *approach.*

Next, we re-ranked the topics based on their ability to contrast VAST and InfoVis papers. Amongst the least distinctive topics were not only topics with generic terms (e.g. *problem, approach, . . .*), but also a topic featuring *time* and *dynamic* and one with *space* and *dimensions,* both suggesting common topics at the venues.

Ranking topics by their ability to contrast 2007 and 2013 indicated topics that have changed. The two highest ranked topics for contrasting these years were *studies, significant, evaluate, . . .* and *fast, gpu, . . .* While many of the lowest ranked topics were generic terms, some recurring challenges, such as *imaging, diffusion* also appeared.

For a more specific exploration, we looked for documents to enhance the related work section of this paper. We started with an initial similar paper, ParallelTopics [17]. This paper had only a single salient topic, one that was clearly relevant ("text, search, learning"). To create a broader query, we searched for all papers with the string "topic" in their title, and ranked by distance to the resulting set of 5 papers. Using the "distance to group center" ranking, the 5 papers were *not* closest to the center—in fact one of the papers was not in the top 30. The top ranked documents shared some aspect of the topic visualization problem, such as handling uncertainty, but not all discussed text visualization. Many of the top documents were relevant, including two of which we were not previously aware but now are discussed in the related work section of this paper. Using the minimum-distance-to-set ranking did place the group at the top of the ordering, and put a slightly different, but still relevant, set of papers next to them.

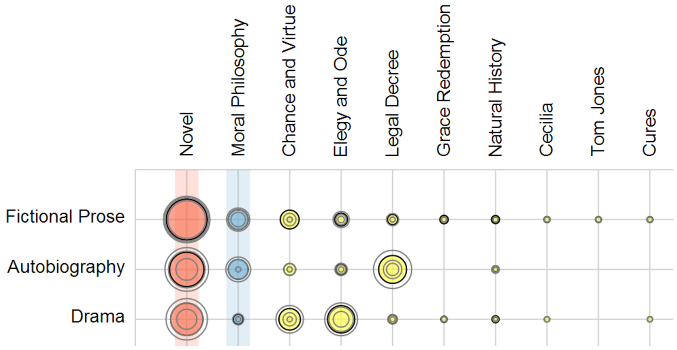


Fig. 8. Sorting topics by the aggregate genre "Fictional Prose" creates an unexpected juxtaposition of topics concerning the novel and moral philosophy.

9.2 Early Modern Literature

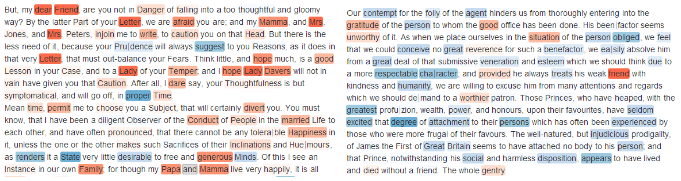
Our collaborators have developed a corpus of 1080 digitized texts published between 1530 to 1799. The corpus was built by randomly sampling 40 texts per decade from a larger archive, in an attempt to provide a less biased view than just using well-known texts. However, this means that the corpus has significant diversity and is unfamiliar to most who work with it. With documents ranging from a few hundred words to hundreds of pages, the corpus is too large for any researcher to read manually, and so we've been interested in seeing how the task of exploring it scales within Serendip. This documents have been run through the VARD 2 modernizer [1] and annotated with metadata such as year, genre, author, publisher, etc.

A literature scholar[1] spent an extended period of time exploring the model over a number of weeks, working between the tool and the texts themselves. He began with the topics themselves, assigning names based on his sense of the texts containing the topics and the distribution of topic words. Naming required extensive switching between levels and combining various sources of information. It also led to some surprises. For example, using the topic lists and RankViewer, he observed that there were a handful of topics containing long lists of numbers. Some contained numbers from 1 all the way up to 100. Another only contained the numbers up to 20 or 30. Examining the documents containing this latter topic revealed that many were written by Protestants. Drilling into the passages revealed the reason for the cutoff: the authors were giving references to Bible passages to support their arguments, and the numbers 20-30 were distinctively biblical (as opposed to 1-20, which were spread across the corpus more broadly). The scholar named the topic "Grace and Redemption."

After building familiarity with the model, the scholar continued to explore it, combining multiple information types and levels of scale in ways that not only answered posed questions, but led to serendipitous discoveries as well. For example, he was able find support for the argument—advanced by scholars of the novel—that the English novel was the literary expression of English moral philosophy (ethical works designed to guide the conduct of citizens). This exploration began by aggregating documents by genre and honing in on a particular one labeled "Fictional Prose" (by a human bibliographer). After sorting the topics by this genre, the top two were topics with which he was familiar from earlier explorations and had labeled "Novel" and "Moral Philosophy" respectively—an interesting juxtaposition (see Figure 8). Sorting the topics by similarity to one another confirmed that these two were very closely related. By drilling into the prose fiction genre within CorpusViewer, the user identified a few texts in which to look for examples of this overlapping or convergence. He began with Samuel Richardson's *Pamela*, one of the first English novels.

From a passage of the novel, he was able to assess how words from both topics were interacting: words from the "novel" topic were in-

[1] The tool was extensively iteratively refined based on his experience. His contributions were so significant that he is a co-author of this paper.



But, my dear Friend, are you not in Danger of falling into a too thoughtful and gloomy way? By the latter Part of your Letter, we are afraid you are; and my Mamma, and Mrs. Jones, and Mrs. Peters, injoin me to write, to caution you on that Head. But there is the less need of it, because your Pru|dence will always suggest to you Reasons, as it does in that very Letter, that must out-balance your Fears. Think little, and hope much, is a good Lesson in your Case, and to a Lady of your Temper; and I hope Lady Davers will not in vain have given you that Caution. After all, I dare say, your Thoughtfulness is but symptomatical, and will go off, in proper Time.
Mean time, permit me to choose you a Subject, that will certainly divert you. You must know, that I have been a diligent Observer of the Conduct of People in the married Life to each other, and have often pronounced, that there cannot be any tolera|ble Happiness in it, unless the one or the other makes such Sacrifices of their Inclinations and Hue|mours, as renders it a State very little desirable to free and generous Minds. Of this I see an Instance in our own Family; for though my Papa and Mamma live very happily, it is all

Our contempt for the folly of the agent hinders us from thoroughly entering into the gratitude of the person to whom the good office has been done. His been|factor seems unworthy of it. As when we place ourselves in the situation of the person obliged, we feel that we could conceive no great reverence for such a benefactor, we ea|sily absolve him from a great deal of that submissive veneration and esteem which we should think due to a more respectable cha|racter; and provided he always treats his weak friend with kindness and humanity, we are willing to excuse him from many attentions and regards which we should de|mand to a worthier patron. Those Princes, who have heaped, with the greatest profu|zion, wealth, power, and honours, upon their favourites, have seldom excited that degree of attachment to their persons which has often been experienced by those who were more frugal of their favours. The well-natured, but injudicious prodigality, of James the First of Great Britain seems to have attached no body to his person; and that Prince, notwithstanding his social and harmless disposition, appears to have lived and died without a friend. The whole gentry

Fig. 9. Passages from the novel *Pamela* (left) and the *Theory of Moral Sentiments* (right). The topic associated with novels is shown in read, while the "moral philosophy" topic is shown in blue.

troducing concrete characters whose actions were the subject of moral evaluation, while the words from the "moral philosophy" topic were applying abstract concepts to those actions, concepts that are the main subject of more argumentatively rich ethical writing (see Figure 9). That convergence of these two kinds of topic words made sense to the user, since the novel must not only analyze actions (involving the reader in a parallel exercise of active moral evaluation), but also render those actions in a rich narrative. In a work of moral philosophy, Adam Smith's *Theory of Moral Sentiments*, the scholar was then able to explore the pattern in reverse (see Figure 9). In this document, words associated with the novel were interwoven with argumentation about moral sentiment and conduct, a finding that also made sense, since moral philosophy must—perhaps unlike metaphysics or logic—take its cues from concrete human action. In other words, there can be no novel, nor any moral thinking, without a concrete and specific situation of personal action and deliberation. Focusing in on the word "situation," the user then transitioned into RankViewer and found that this word, which is often used to shift readers away from their immersion in the story into a more explicitly evaluative cognitive frame, was highly rated on both topics.

The analysis had thus progressed unexpectedly through four levels of abstraction: a ground truth had been correlated with algorithmically generated topics (a topic, "novel", tracked reliably with works aggregated as prose fiction); that ground truth was then extended into an unexpected juxtaposition (the close relation of the "novel" and "moral philosophy" topics); exemplary works were identified and their narrative techniques evaluated on the level of passages and individual words (novel and moral philosophy words intermix on the page); and finally, topic words were found that sit at the intersection of these two narrative forms ("situation"). Having opened up a new level on which to explore a current critical debate about the novel, our user then returned to the matrix view to rate the existing genres according to their scores on the "novel" topic—an exploration that was also suggestive, since "novel" captured not only prose fiction, but texts classified as autobiography, drama, travelogue, and biography. Each of these subgenres has also been related to the novel in literary studies, and so the user was able to begin generating hypotheses about how novelistic language might have developed from, or be shared with, these types of writing, many of which pre-date the novel as literary forms.

10 Discussion

Serendip was intended to support and promote scalable, multi-level serendipitous discovery in text corpora, something it appeared to do among the users who tried it. We believe that the tool was able to achieve these initially promising results because it multiplies the angles from which users can enter and then transition through a corpus—in effect, minimizing "roads not taken." Our use cases show investigations that combine topic models with other sources of information to reveal discoveries at multiple levels of detail. Our methods for addressing scale seem to apply for the corpora with over 1000 documents, texts on the order of full books, and up to 100 topics in the model. The various starting points and ways to use intermediate results and questions to springboard to next steps provide a fluency of exploration that keeps users engaged.

At present, we have only seen the tool in the hands of a limited audience, and have not performed a formal evaluation. We need to see how our methods work across a broader range of corpora, and to see if other users can effectively make use of the tool in their work. While a formal evaluation of specific elements of our approach may be interesting, and help to refine them, the more interesting and challenging question is to confirm whether our techniques meet their goals of fostering insight and serendipitous discovery across multiple scales of data and abstraction. Evaluating for insight generation is an open problem, and a difficult one [29]. Evaluating for serendipity is perhaps even more so, due to the unexpected nature of the findings it allows. Beyond our case studies, we hope to perform more longitudinal study of the technique's successes in this direction.

One drawback to our approach is the static nature of the models upon which it operates. Though we believe model agnosticism is important for the tool to be accessible to users unfamiliar with topic modeling, their input into the tuning process is nonetheless invaluable. We are currently exploring ways of using users' interactions with Serendip as a way of harvesting their expertise for the model training process.

Scalability to larger corpora will be important. Apart from some current implementation bottlenecks, there is a question of how to better handle scale in the corpus view. Even with filtering, sorting, and aggregation, there is a need to assess orderings and see patterns at an even broader scale. We foresee a need for methods that "zoom out" on the matrix, as well as incorporating other kinds of corpus overviews. While many visual summary methods exist in the literature, one challenge will be developing interaction techniques that couple them with our current views.

A more technical challenge is developing a more rigorous mathematical toolkit for working with document vectors. As mentioned, our current distance metrics, averaging methods, and statistics do not preserve the sparse, convex structure of the vectors. Local neighborhood graph distances seem promising potential improvements given the success of local distance-based embedding. While our prototype offers a rich set of ordering metrics, an improved set would offer better opportunities for achieving scaling through placing relevant objects first, and serendipity by bringing different things together. One promising avenue is to apply distance metric learning approaches to allow users to craft ordering functions based on sparse sets of examples.

In the future, we would like to add improved methods for finding passages, e.g. through similarity to exemplars. We also hope to support exploration that uses multiple topic models, either to compare them or make use of differences in what they uncover. Support for a richer array of model types, such as hierarchical or multinomial topic models would also be an interesting extension [18].

By providing a set of visual encodings that address multiple levels of exploration, interaction techniques for coupling these views, and statistical techniques for linking different sources of information, Serendip allows users to use topic models and other information to guide the exploration of text corpora. The methods are designed not only to address scale, but also to promote serendipitous discovery. Our initial experience using the tool for literary scholarship suggests that it engages users and helps them make discoveries.

Acknowledgments

This work was supported in part by NSF award IIS-1162037 and a grant from the Andrew W. Mellon Foundation.

References

[1] A. Baron, P. Rayson, and D. Archer. Quantifying early modern english spelling variation: change over time and genre. In *Conf. New Methods in Historical Corpora*, 2011.

[2] J. Bertin. *Semiology of Graphics: diagrams, networks, maps*. Esri Press, 2011.

[3] D. Blei. Probabalistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE TVCG*, 2011.

[6] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: a catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32, 2003.

[7] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *IEEE Visual Analytics Science and Technology*, pages 83–92. IEEE, 2012.

[8] J. Chuang, C. Manning, and J. Heer. Termite: visualization techniques for assessing textual topic models. In *Proc. Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[9] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. ACM Human Factors in Computing Systems*, pages 443–452. ACM, 2012.

[10] T. Clement, C. Plaisant, and R. Vuillemot. The story of one: Humanity scholarship with visualization and text analysis. *Relation*, 10(1.43):8485, 2009.

[11] S. Climer and W. Zhang. Rearrangement clustering: Pitfalls, remedies, and applications. *J. Machine Learning Research*, 7:919–943, 2006.

[12] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009.

[13] M. Correll, E. Alexander, and M. Gleicher. Quantity estimation in visualizations of tagged text. In *Proc. ACM Human Factors in Computing Systems*. ACM, 2013.

[14] M. Correll and M. Gleicher. What Shakespeare taught us about text visualization. In *IEEE Visualization Workshop Proceedings, Interactive Visual Text Analytics*, 2012.

[15] M. Correll, M. Witmore, and M. Gleicher. Exploring Collections of Tagged Text for Literary Scholarship. *Computer Graphics Forum*, 30(3):731–740, 2011.

[16] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE TVCG*, 17(12):2412–2421, 2011.

[17] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *IEEE Visual Analytics Science and Technology*, pages 231–240. IEEE, 2011.

[18] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE TVCG*, 19(12):2002–2011, 2013.

[19] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proc. ACM Human Factors in Computing Systems*, pages 473–482. ACM, 2012.

[20] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *IEEE Visual Analytics Science and Technology*, pages 121–130, 2011.

[21] M. Gleicher. Explainers: expert explorations with crafted projections. *IEEE TVCG*, 19(12):2042–51, 2013.

[22] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing theme changes over time. In *Proc. IEEE Information Visualization*, 2000.

[23] N. Henry and J.-D. Fekete. MatLink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007*, pages 288–302, 2007.

[24] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato. Local Affine Multidimensional Projection. *IEEE TVCG*, 17(12):2563–2571, 2011.

[25] D. A. Keim and D. Oelke. Literature fingerprinting: A new method for visual literary analysis. In *IEEE Visual Analytics Science and Technology*, pages 115–122. IEEE, 2007.

[26] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[27] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1969–1972. ACM, 2005.

[28] A. K. McCallum. MALLET: A machine learning for language toolkit. 2002. http://mallet.cs.umass.edu.

[29] C. North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, 2006.

[30] F. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *Conf. Information Visualisation*, pages 245–251. IEEE, 2006.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Machine Learning Research*, 12:2825–2830, 2011.

[32] C. Plaisant, J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord. Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In *Proc. ACM/IEEE Joint Conf. Digital Libraries*, pages 141–150. ACM Press, 2006.

[33] R. Rohrer, D. Ebert, and J. Sibert. The shape of Shakespeare: visualizing text using implicit surfaces. In *Proc. IEEE Information Visualization*, pages 121–129. IEEE, 1998.

[34] H. Siirtola. Interaction with the reorderable matrix. In *Proc. IEEE Information Visualization*, pages 272–277. IEEE, 1999.

[35] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[36] A. Thudt, U. Hinrichs, and S. Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proc. ACM Human Factors in Computing Systems*, pages 1461–1470. ACM, 2012.

[37] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proc. ACM Knowledge discovery and data mining*, pages 153–162. ACM, 2010.

[38] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012.