The Topic Browser
An Interactive Tool for Browsing Topic Models

Matthew J. Gardner
Department of Computer Science
Brigham Young University
mjg82@byu.edu

Joshua Lutes
Department of Computer Science
Brigham Young University
jlutes@byu.edu

Jeff Lund
Department of Computer Science
Brigham Young University
jefflund@gmail.com

Josh Hansen
Department of Computer Science
Brigham Young University
jjfresh@byu.edu

Dan Walker
Department of Computer Science
Brigham Young University
danwalkeriv@gmail.com

Eric Ringger
Department of Computer Science
Brigham Young University
ringger@cs.byu.edu

Kevin Seppi
Department of Computer Science
Brigham Young University
kseppi@cs.byu.edu

Abstract

Topic models have been shown to reveal the semantic content in large corpora. Many individualized visualizations of topic models have been reported in the literature, showing the potential of topic models to give valuable insight into a corpus. However, good, general tools for browsing the entire output of a topic model along with the analyzed corpus have been lacking. We present an interactive tool that incorporates both prior work in displaying topic models as well as some novel ideas that greatly enhance the visualization of these models.

1 Introduction

Proper visualizations are essential to extracting information from and identifying trends in data, especially large, high dimensional data. Large text corpora are particularly difficult to visualize, as they may include many thousands of documents and millions of words. Topic modeling [1] is a method of reducing the dimensionality of a corpus into a set of meaningful topics. Topic models have been used with some success as an aid to corpus visualization [3, 9]. However, in many papers the presentation of topic models has been limited to hand-selected, coherent topics, when often there are many meaningless topics to sift through. These presentations leave the viewer wondering what else the topic model discovered and provide little help in gaining a deep understanding of the model.

To aid in the visualization of topic models and in pattern discovery in document collections, we present The Topic Browser, a web-based tool for interactively exploring both the output of a topic modeling algorithm and its attendant corpus. Our topic browser incorporates many visualizations of topic models previously published as well as some innovative ideas of our own. The Topic Browser

is an aid both to those who wish to browse through a corpus and for those who wish to analyze the topic model itself. We believe that this interactive approach to visualization in browsing reveals topics and topical trends in large document collections in a way that is not possible with static visualizations. The rest of this paper describes the Topic Browser in an attempt to demonstrate that this is true. We display our tool with a topic model run using MALLET's implementation of LDA [6] with 150 topics on a collection of about 460 campaign speeches from the 2008 presidential primary and general elections retrieved from http://2008election.procon.org.

2 Extra Information

Aside from the documents and the topic model themselves, our browser incorporates three other pieces of information: attributes (metadata) associated with each document, topic metrics, and document metrics. Here we give a brief description of these kinds of information, deferring a discussion of their use in visualization to the subsequent sections.

Attributes of documents have been used heavily in recent topic models, including the Author-Topic model [11], Topics over Time [12], and Dirichlet-Multinomial Regression [7]. While we currently do not have specialized visualizations for these topic models, we do include document attributes in our browser and have some of our own visualizations that include the attributes.

In order to browse more effectively through topics, we introduce topic metrics that give information about the topic. These range from simple metrics, such as the number of word tokens and types labeled with the topic, to more complicated metrics such as how dispersed the topic is across documents, or how coherent its words are [10]. We also use pairwise topic metrics as similarity measures to show similar topics.

Similar to topic metrics, one can also compute document metrics. Beyond simple metrics like token count in the document, these include things such as the entropy of the topic distribution of the document [8]. And as with topics, we make use of pairwise document metrics such as topic correlation [3] to show similar documents.

3 Sidebar

The sidebar in our browser is the main navigation tool. To facilitate navigation through the large amount of information in the browser, the sidebar lists the items of the type currently being explored. The new capability introduced by this browser is the ability to sort and filter these lists as the user desires.

When presenting the results of a topic model analysis, papers will often hand pick particularly good topics to display, leaving out a large number of meaningless topics. While such presentations may help to highlight the benefits of a new topic model, they hide the fact that finding good topics is often a laborious process. With our topic and document metrics, the sorting and filtering capabilities of the sidebar list allow the user to go quickly to meaningful topics that are relevant to his questions.

When browsing through topics, the user can filter the topic list by coherence to eliminate from the view those topics that are mostly meaningless and sort by document entropy (a measure of the dispersion of the topic across the documents) to find topics that were used widely throughout the corpus. For example, Figure 1 shows that "the family" was one of the most consistent themes in the speeches. We use the top two words to name a topic in the sidebar and other places, to avoid clutter, though we are actively investigating better ways to name topics.

The user can also filter by document or attribute to show only topics that were used in a particular document or by a particular author. Documents can similarly be filtered and sorted, showing only documents that contain tokens with a particular topic or with a particular value of an attribute.

4 Topic Browsing

The results from a topic model are typically presented as static lists of words, often simply showing the top ten words from each topic in a list [1]. While this practice is often sufficient to give a reader

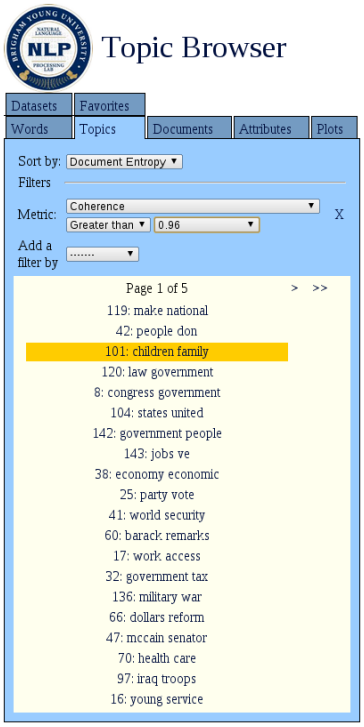

Figure 1: The navigational sidebar showing topics. In this example the topics are sorted by dispersion throughout the corpus and filtered by coherence

a basic idea of what the topic captured, it is largely incapable of conveying a deep understanding of the context of that topic in the corpus. In order to answer significant questions about a corpus one must browse the topics discovered, and previous methods have failed to adequately allow the user to explore topics. Our browser presents new ways to gather information about a topic in the corpus, both in terms of how the words in the topic are displayed and in terms of what is presented about the topic.

4.1 Showing Top Words

Instead of showing a list of the top ten words, we show a word cloud of the top 100, where the size of each word is determined by the probability of seeing that word in the topic. We also show a re-weighted word cloud as determined by Blei & Lafferty's Turbo Topics method [4].

While these visualizations are modest improvements over typical visualization methods, our most useful display of the words is showing them in their context. The topics in a topic model cannot be fully interpreted when completely separated from their context. Thus in addition to the word cloud we also display the top ten words in the topic inside of their context. We select a random token of each word type labeled with the topic and show up to 50 characters on either side of the token, keeping words intact. We also allow the user to cycle through contexts to gain a broader view of how each word is actually used in that topic. When viewing the topic about troops in Iraq, for example (see Figure 2), one can see that most often when the word "troops" is used in this topic, it is in reference to bringing the troops home, and when "end" is used, it is not used in the context of "end the war" as often as one might expect.

4.2 Getting More Information

The browser provides two main ways of getting more information about topics in the model. The first is showing documents and attribute values which have either the highest number of tokens labeled with that topic or the highest proportion (by percentage) of the given topic. This allows

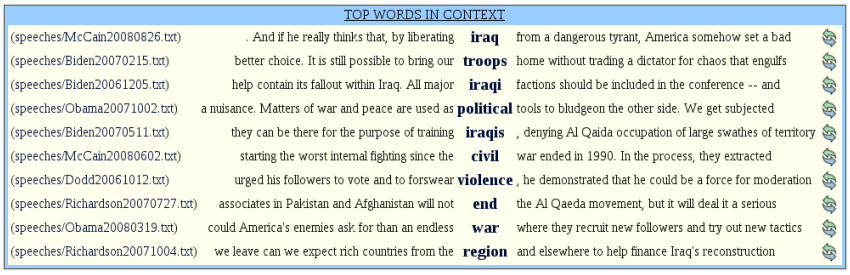

Figure 2: The top ten words in the topic shown with a random context selected from the corpus.

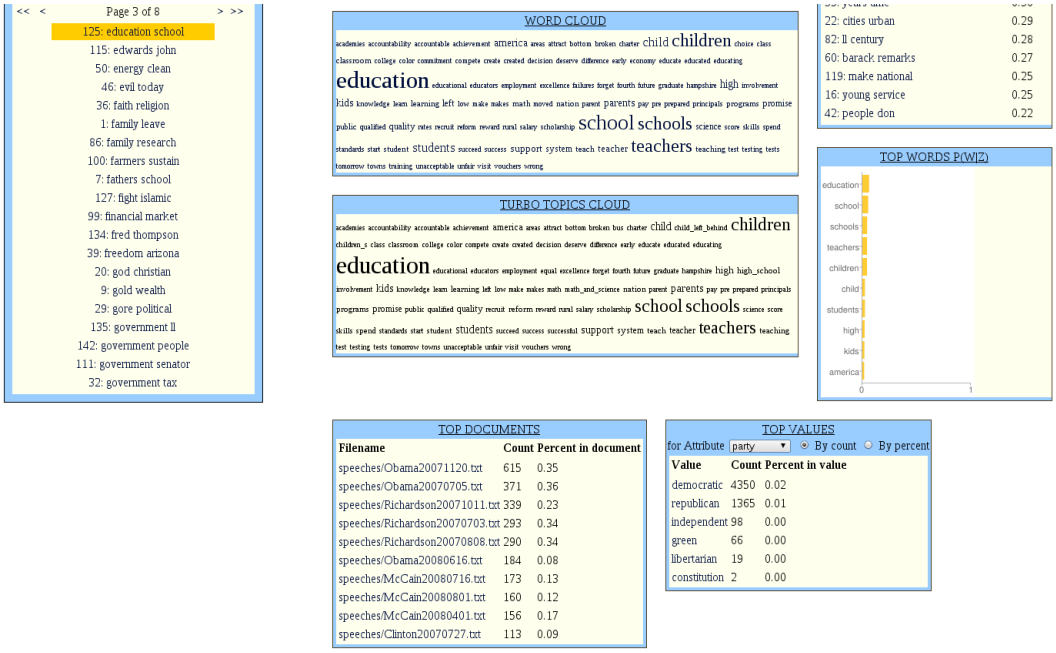

Figure 3: Part of the overall topic view, looking at a topic about education. Note particularly the top documents for the topic and the top values for the attribute "party," shown at the bottom.

the user to quickly find documents that best demonstrate what the topic captured. Showing the top values for a given attribute (such as authors for the attribute "Author") gives the user an idea of how focused the topic is and often reveals interesting information about the corpus being browsed. For example, in Figure 3, we see that Democrats spoke more often about education than Republicans did, at least in the topic shown.

The other way that we give more information about topics is by showing similar topics. We find similar topics either by looking at the distribution of documents that contain the topic or by the topic's distribution over words. Looking at similar topics by document shows topics that are commonly used together in the corpus, and finding similar topics by word distribution shows topics that have similar words, possibly because the two topics really should have been one topic. When looking at a topic about health care, the user can see that other topics used together with the health care topic include topics about the cost of medication and the quality of care (see Figure 4).

5 Document, Word, and Attribute Browsing

In addition to enabling the user to browse the topics in the topic model, we provide means for browsing the documents, words, and attributes in the corpus. These facilities give users the ability to explore the corpus itself in the context of the topic model.

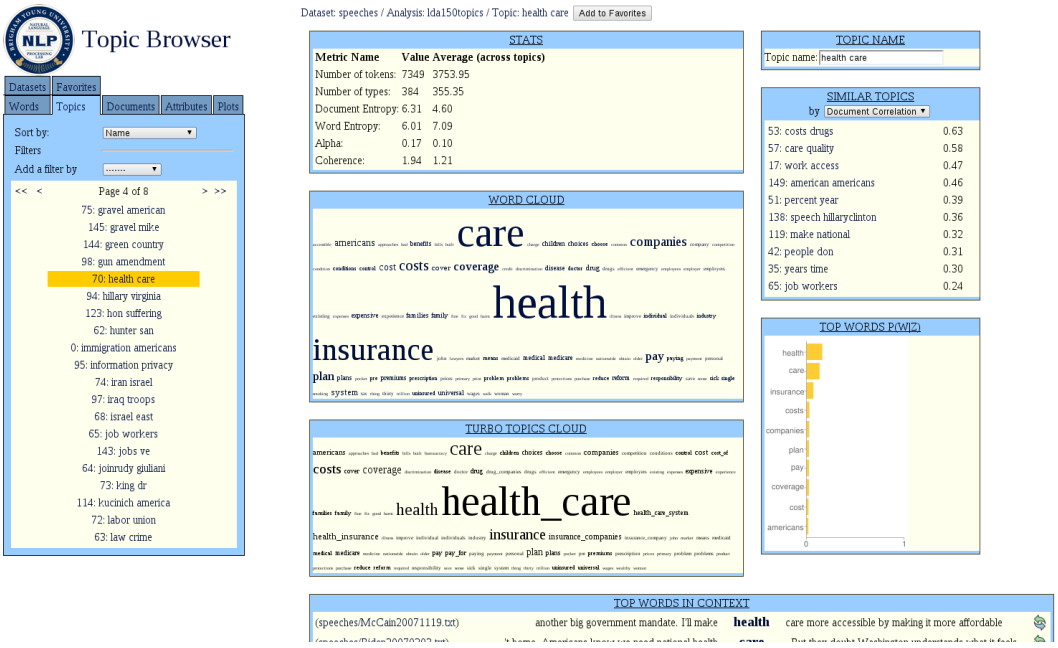

Figure 4: Another look at the overall topic view, featuring different parts of the view. Note particularly the similar topics box near the top right.

When simply browsing through the documents, we provide sorting and filtering methods on the list of documents, as mentioned previously. When looking at a particular document, we show basic information about the document, its text, the topic distribution in the document, and similar documents based on that distribution. When looking at the document in the context of a topic, we also highlight the tokens in the document that were labeled with that topic. For example, the user might be curious to see a document that best demonstrates the health care topic mentioned above. Clicking on the top document in the document list (as in the bottom of Figure 3) brings the user to the view shown in Figure 5.

The document visualizations reported here were drawn from the work of others [1], and in fact the view in Figure 5 constitutes the entirety of most previous corpus browsers based on topic models (e.g., [2]). Our browser substantially goes beyond the existing functionality reported by others.

We also provide views of individual words in the corpus. When viewing a word in the context of a topic, the user can see all uses of the word in that topic in the corpus, with context taken from their corresponding documents. The user can also view words independently with a search-like interface, seeing topics and documents in which the word appears most frequently. The user examining the health care topic might be curious where else the word "health" was used in the topics and in the corpus. Figure 6 shows the result of using our word search to answer that question. While providing basic functionality, however, the search interface leaves much to be desired, as only single words can currently be searched for. We plan on expanding that to phrases.

The user can also look at aggregated information for the values of an attribute (i.e., a particular candidate or party), combining all of the topic and word counts for all documents with the given attribute. This view is also at present somewhat limited, showing only what topics and words are used most frequently by the collection of documents with that attribute. Figure 7 shows us, among other things, that one of Barack Obama's top topics was "change in politics."

6 Plots

We currently include two kinds of plots in our topic browser, with plans to implement many more. The first shows trends for topics over the values of an attribute (such as date, or candidate), useful for corpus browsing. This kind of plot has been used in visualizing topic models almost since their

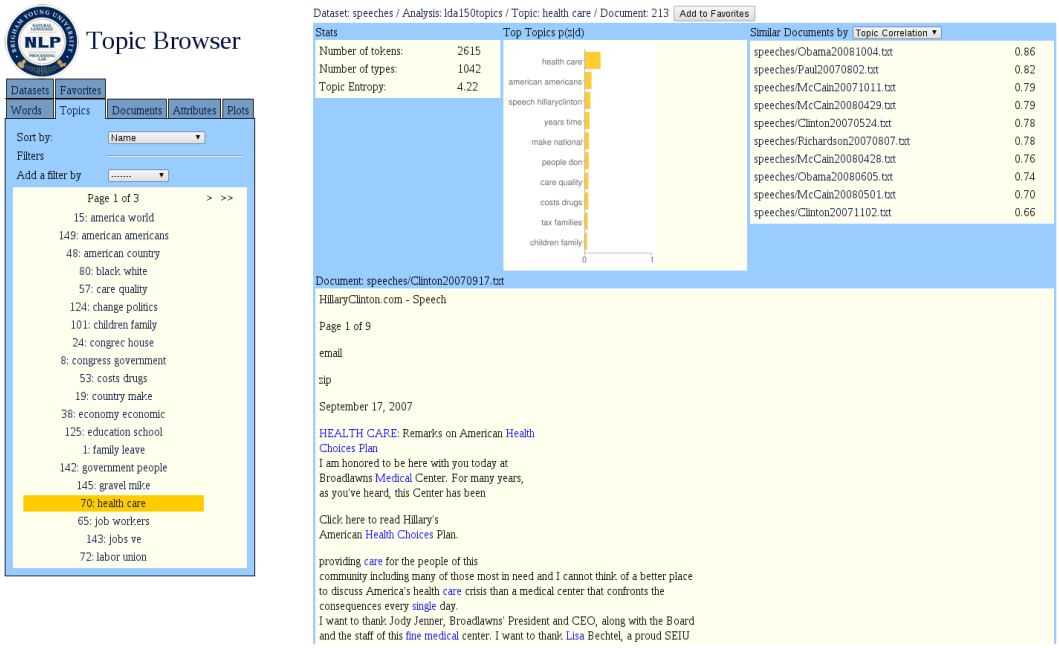

Figure 5: The document view, looking at a speech by Hillary Clinton in the context of a topic about health care (the rest of the document is cut off in this screenshot, and sadly, the colored tokens in the document are not very visible in black and white).

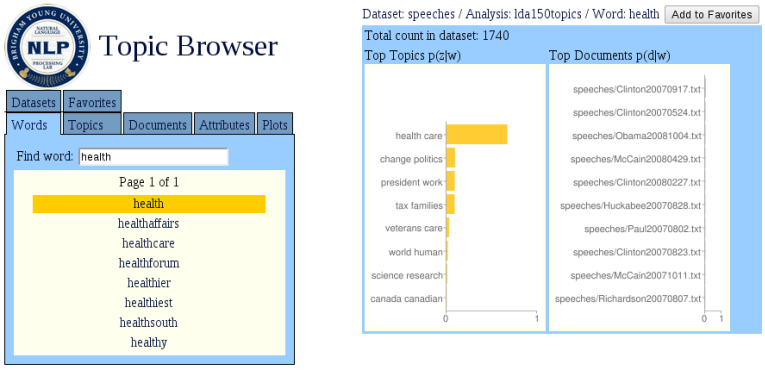

Figure 6: The word search functionality, showing the results for searching for the word “health.”

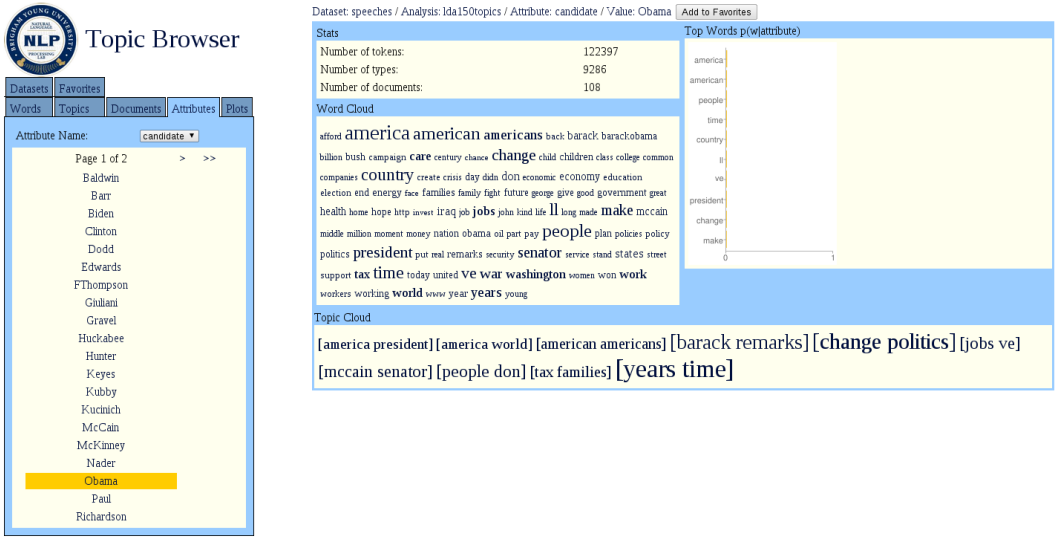

Figure 7: The attribute view, showing aggregated information for all speeches given by Barack Obama.

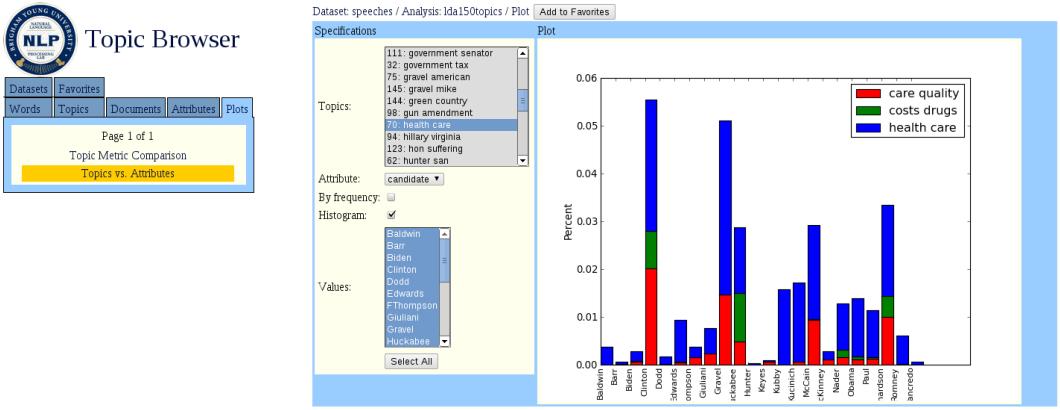

Figure 8: A plot of topics over attributes, showing the use of three health care-related topics across candidates.

introduction [5]. Our topic browser allows the user to interactively generate these trend plots over any attribute for any topic or combination of topics in the corpus. Our user exploring health care topics may want to view how much each candidate spoke about health care. We saw already that there were three related topics that mentioned health care, so the user might view all of them together in a histogram, shown in Figure 8.

The second kind of plot that we include is more useful for analyzing and understanding the behavior of the topic model itself. We allow the user to plot two topic metrics against each other and compute a linear regression. This allows the user to see some interesting properties of the topic metrics, such as the fact that document entropy seems to correlate with the logarithm of the number of tokens in the topic, and that coherence does not seem to correlate with any other topic metric. The user can also find outliers, such as topics with low document entropy but a high token count, that can then be examined in the topic page.

An interesting application of these topic metric plots occurs when the metrics include how consistently each candidate spoke about each topic. Figure 9 shows a plot of topics, comparing John McCain's use of each topic to Barack Obama's use of the topic. Topics in the upper left were topics unique to Obama, and topics in the lower right were unique to McCain. Topics in the middle (near the value of 2 for each candidate) were somewhat shared between the two.

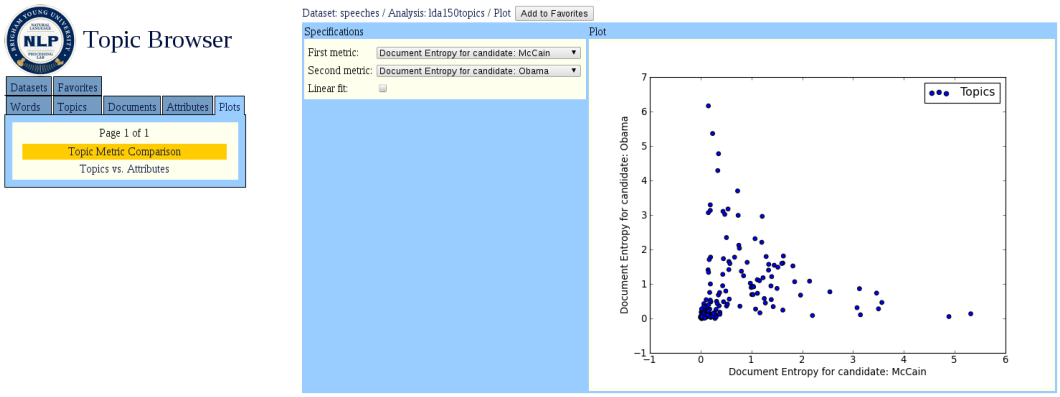

Figure 9: A topic metric comparison plot. The metrics plotted are how consistently Barack Obama and John McCain used each topic. The plot shows topics that were unique to Obama or McCain and topics they shared.

7 Conclusion

We have presented the Topic Browser, an interactive tool for browsing both the output of a topic model and the corpus that was modeled. We have shown that our tool incorporates many previously published visualizations of topic models, including basic corpus browsing functionality, plots of trends over attributes in the corpus, the use of coherence to ignore meaningless topics, and Blei & Lafferty's Turbo Topics method of finding significant phrases for each topic. We have also presented many novel ways to mine information from a topical analysis of a corpus in an interactive browsing experience. The Topic Browser is an effective tool both for those wishing to browse through a corpus in the context of a topic model, and for those wishing to better understand topic models and develop new models.

While our tool is still under development, a description of the Topic Browser, a working demo, and the current version of the code are available at http://nlp.cs.byu.edu/topic_browser. Our tool currently supports any topic model that labels individual tokens in the corpus with topics, and is built to import data directly from MALLET input and output files [6], or files similarly formatted. We have plans to include specialized visualizations for more complicated topic models, such as Topics over Time, sentiment-topic models, hierarchical topic models, and others.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] D.M. Blei. 50-topic-browser of latent Dirichlet allocation fit to the 2006 arXiv. http://topics.cs.princeton.edu/arxiv/browser50/. Accessed 10/21/2010.

[3] D.M. Blei and J.D. Lafferty. Topic Models. In Ashok Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*. Taylor and Francis, 2009.

[4] D.M. Blei and J.D. Lafferty. Visualizing Topics with Multi-Word Expressions. arXiv:0907.1013v1 [stat.ML], 2009.

[5] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228, 2004.

[6] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.

[7] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2008.

[8] H. Misra, O. Cappé, and F. Yvon. Using LDA to detect semantically incoherent documents. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 41–48. Association for Computational Linguistics, 2008.

[9] D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2010.

[10] D. Newman, J.H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL HLT*, 2010.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, page 494. AUAI Press, 2004.

[12] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.