Increased Comprehension of Topic Models and Corpora Using the Topical Guide

Abstract

Although topic models can be employed for dimensionality reduction, their token-level output is as voluminous as the original dataset. Thus, humans' ability to manually assess model quality is limited. Moreover, as the quantity of digitized documents worldwide continues to expand, institutions correspondingly find an increased need for tools such as topic models to make information in these texts discoverable and thereby usable. In response to these twin needs, we present the Topical Guide, an open-source web application for interactive, topic-centric exploration and visualization of topic model output. We explain why such a tool is warranted, what it is capable of, and how to use it to explore the corpus or topic model of your choice.

1 Introduction

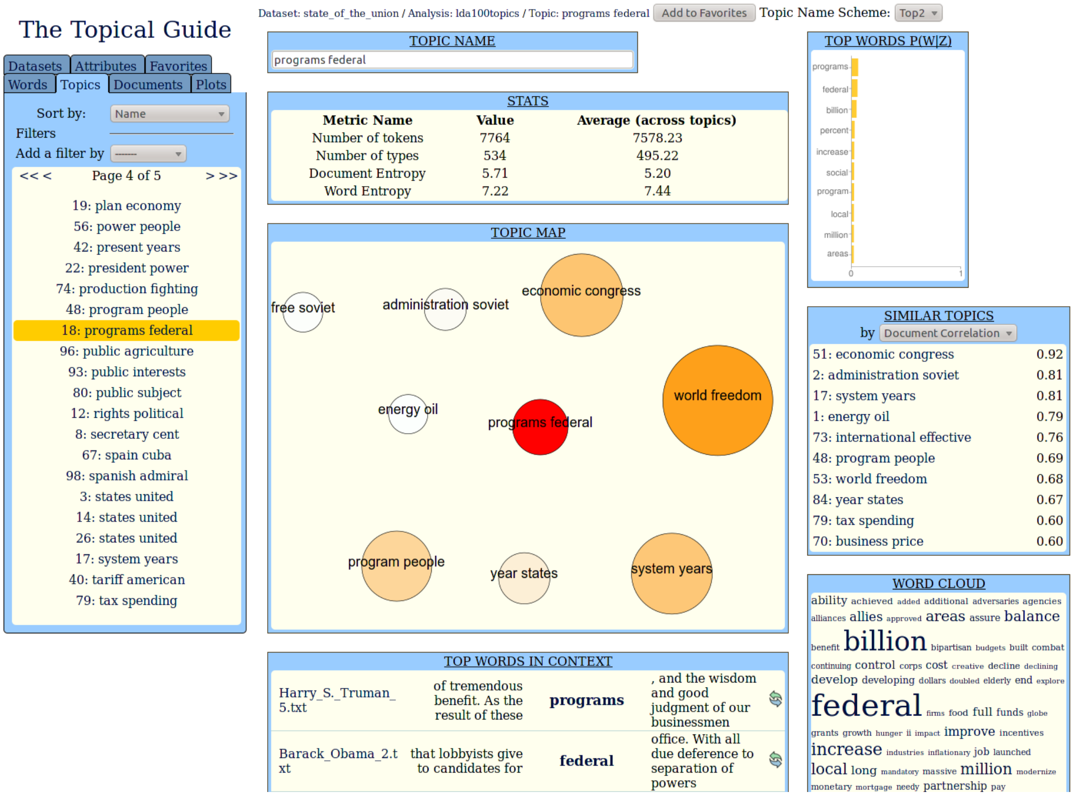
Since its introduction, LDA-based topic modeling (Blei et al., 2003) has become standard fare for those wishing to automatically distill large text collections into something more immediately useful to humans and computers alike. The usefulness of this sort of dimensionality reduction is widely acknowledged, and topic models continue to be extended into exciting new territory (Wang et al., 2008; Mimno et al., 2009; Brody and Elhadad, 2010). However, in their most verbose realization, these models output a topic assignment per token, resulting in an output that is a corpus unto itself. Though the range of possible values is substantially reduced, output size remains on the same order as input size. Notwithstanding, papers introducing new topic models often lean heavily on the hope that displaying a few hand-picked topics will convince others of the model's effectiveness. But the very vastness of token-level topic model output renders this sort of presentation unconvincing and of limited utility. Unfortunately, tools for deep understanding of topic model output have thus far been lacking.

Simultaneously, massive digitization efforts by libraries and other institutions are generating gargantuan quantities of electronic text, increasing demand for tools such as topic models. Because of their ability to characterize the contents of even large, unknown corpora, topic models seem to hold great promise for archival institutions wishing to make their newly-digitized collections accessible to researchers and the public. However, readily available tools for exploration of document collections in terms of topic models have been either unsatisfactory or nonexistent.

In response, we present the Topical Guide, an open-source[1] web application for interactive, topic-aware exploration and visualization of both document collections and the topic models inferred on them. The Topical Guide 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, topic-aware approach to corpus and model exploration exposes meaningful patterns to human comprehension in a way that static visualizations cannot.

In this paper, we will discuss: browsing of topic model output using the Topical Guide user interface;

[1] Licensed under the terms of the Affero General Public License, version 3.


The Topical Guide
Datasets
Attributes
Favorites
Words
Topics
Documents
Plots
Add to Favorites
TOPIC NAME
STATS
TOPIC MAP
free soviet
administration soviet
economic congress
world freedom
energy oil
programs federal
program people
year states
system years
TOP WORDS P(W|Z)
programs
federal
billion
percent
increase
social
program
local
million
areas
0
1
SIMILAR TOPICS
WORD CLOUD
ability achieved added additional adversaries agencies alliances allies approved areas assure balance benefit billion bipartisan budgets built combat continuing control corps cost creative decline declining develop developing dollars doubled elderly end explore federal firms food full funds globe grants growth hunger ii impact improve incentives increase industries inflationary job launched local long mandatory massive million modernize monetary mortgage needy partnership pay
TOP WORDS IN CONTEXT

Dataset: state_of_the_union / Analysis: lda100topics / Topic: programs federal

Topic Name Scheme: Top2

programs federal

Sort by: Name

Filters

Add a filter by ------

<< <   Page 4 of 5   > >>

19: plan economy
56: power people
42: present years
22: president power
74: production fighting
48: program people
18: programs federal
96: public agriculture
93: public interests
80: public subject
12: rights political
8: secretary cent
67: spain cuba
98: spanish admiral
3: states united
14: states united
26: states united
17: system years
40: tariff american
79: tax spending

Metric Name	Value	Average (across topics)
Number of tokens	7764	7578.23
Number of types	534	495.22
Document Entropy	5.71	5.20
Word Entropy	7.22	7.44

SIMILAR TOPICS by Document Correlation

51: economic congress	0.92
2: administration soviet	0.81
17: system years	0.81
1: energy oil	0.79
73: international effective	0.76
48: program people	0.69
53: world freedom	0.68
84: year states	0.67
79: tax spending	0.60
70: business price	0.60

Harry_S._Truman_5.txt	of tremendous benefit. As the result of these	**programs**	, and the wisdom and good judgment of our businessmen
Barack_Obama_2.txt	that lobbyists give to candidates for	**federal**	office. With all due deference to separation of powers

Figure 1: The topic view. Visual and textual means combine to convey topic properties in a multimodal fashion, while hyperlinks connect users to statistics about specific words and documents.

document metadata; document and topic metrics; a topic-space visualization; topic name schemes; and the data import system. Throughout the discussion, examples are taken from a corpus of State of the Union messages delivered by United States presidents from 1790 to 2010. This constitutes 223 documents totalling 757,823 tokens. Though this corpus is admittedly small, the Topical Guide has also been employed on corpora of hundreds of thousands of documents and millions of tokens.

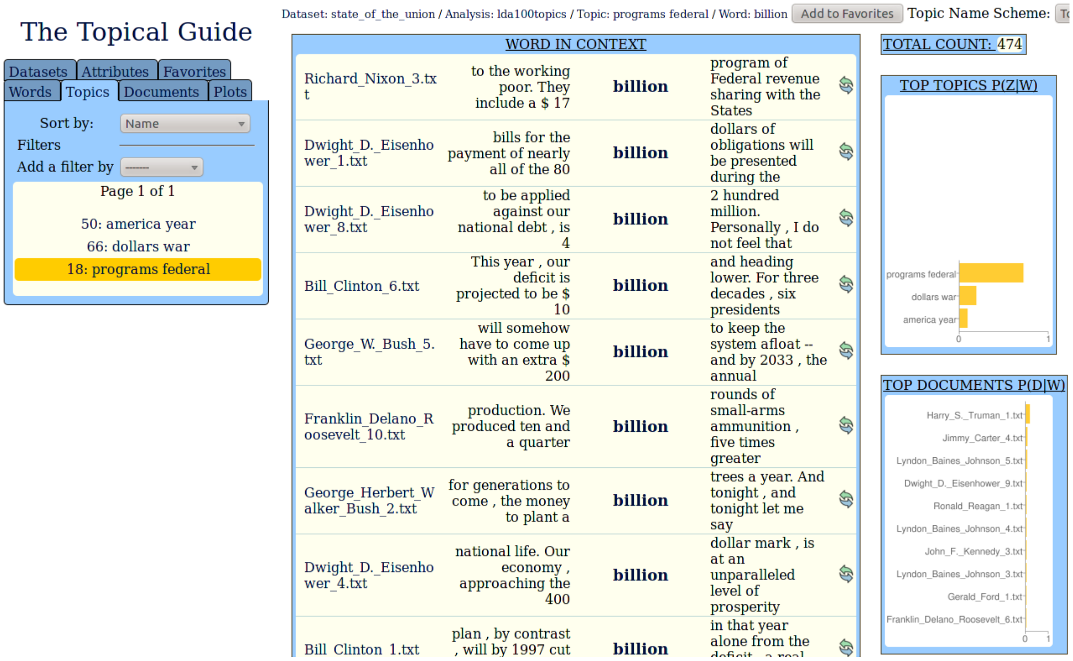
2 Browsing

All entities explicitly modeled by a basic topic model—topics, documents, and words—are first-class citizens in the Topical Guide, meaning that the user interface provides specific views for each. The topic view is central to the user experience. A view of the "programs federal" topic is rendered in Figure 1. On the left is a navigation sidebar listing other topics, with tabs providing links to other views such as Attributes, Documents, and Plots. The remainder of the page shows statistics about the topic (`STATS`); chart, word-cloud, and keyword-in-context representations of top words (`TOP WORDS P(W|Z)` / `WORD CLOUD` / `TOP WORDS IN CONTEXT`); and both textual and graphical representations of similar topics (`SIMILAR TOPICS` / `TOPIC MAP`).

The word view is shown in Figure 2. It presents the user with a list of contexts in which the word appears, along with the total number of occurrences, and lists of topics and documents in which the word occurs most often.

The document view can be seen in Figure 3, showing the content of the document along with prominent topics and any document metrics that have been computed. By means of pairwise document metrics, a list of similar documents is shown as well.

Intuitive hyperlinks facilitate an interactive experience. Click on the "economic congress" node in the topic map and be sent to its topic view to see how presidents have discussed economic issues. Click on



The Topical Guide

Datasets | Attributes | Favorites | Words | Topics | Documents | Plots

Sort by: Name

Filters

Add a filter by ------

Page 1 of 1

50: america year

66: dollars war

18: programs federal

Dataset: state_of_the_union / Analysis: lda100topics / Topic: programs federal / Word: billion | Add to Favorites | Topic Name Scheme:

WORD IN CONTEXT

Richard_Nixon_3.txt	to the working poor. They include a $ 17	**billion**	program of Federal revenue sharing with the States
Dwight_D._Eisenhower_1.txt	bills for the payment of nearly all of the 80	**billion**	dollars of obligations will be presented during the
Dwight_D._Eisenhower_8.txt	to be applied against our national debt , is 4	**billion**	2 hundred million. Personally , I do not feel that
Bill_Clinton_6.txt	This year , our deficit is projected to be $ 10	**billion**	and heading lower. For three decades , six presidents
George_W._Bush_5.txt	will somehow have to come up with an extra $ 200	**billion**	to keep the system afloat -- and by 2033 , the annual
Franklin_Delano_Roosevelt_10.txt	production. We produced ten and a quarter	**billion**	rounds of small-arms ammunition , five times greater
George_Herbert_Walker_Bush_2.txt	for generations to come , the money to plant a	**billion**	trees a year. And tonight , and tonight let me say
Dwight_D._Eisenhower_4.txt	national life. Our economy , approaching the 400	**billion**	dollar mark , is at an unparalleled level of prosperity
Bill_Clinton_1.txt	plan , by contrast , will by 1997 cut	**billion**	in that year alone from the

TOTAL COUNT: 474

TOP TOPICS P(Z|W)

TOP DOCUMENTS P(D|W)

Harry_S._Truman_1.txt
Jimmy_Carter_4.txt
Lyndon_Baines_Johnson_5.txt
Dwight_D._Eisenhower_9.txt
Ronald_Reagan_1.txt
Lyndon_Baines_Johnson_4.txt
John_F._Kennedy_3.txt
Lyndon_Baines_Johnson_3.txt
Gerald_Ford_1.txt
Franklin_Delano_Roosevelt_6.txt
0
1

Figure 2: A view of the word "billion" as it occurs in the "programs federal" topic.

"billion" in the word cloud to see how that word has been used in various addresses (Figure 2). Or click on "Harry_S._Truman_5.txt" to see President Truman's message with words belonging to the "programs federal" topic highlighted (Figure 3).[2]

3 Metrics

Topic-centric document exploration is enhanced by means of *metrics*. Metrics are functions that give users additional insight into the nature of topics or documents. Topic metrics range from the simple, such as the number of word tokens and types labeled with the topic, to more complicated metrics like dispersion across documents, prevailing sentiment, or semantic coherence of the topic's words (Newman et al., 2010b).

Pairwise topic metrics describe relationships between topics.[3] These include Pearson correlation on both documents and words. Pairwise metrics are used to automatically display a list of similar topics and to weight the edges of the graph used to generate topic maps (discussed further in Section 5).

Similar to topic metrics, document metrics can also be computed. Beyond simple metrics like token count in the document, these include metrics such as the entropy of the topic distribution of the document (Misra et al., 2008). As with topics, we make use of pairwise document metrics such as topic correlation (Blei and Lafferty, 2009) to show similar documents.

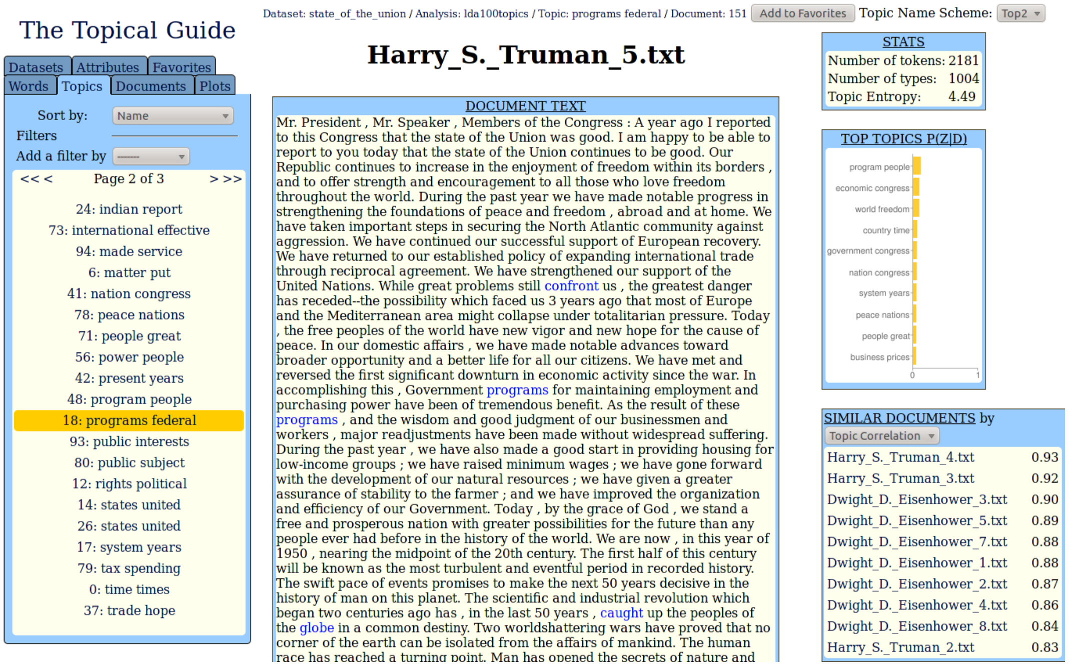
4 Metadata and Plots

Acknowledging that more is known about most documents than just their content, the Topical Guide provides *document attributes*, a mechanism for association of arbitrary metadata with documents. This accomodates metadata provided by a dataset curator, often including facts about document provenance and author identity. Document attributes also allow for metadata obtained as part of the output of a model. These attributes can be used to filter the list of documents, or as independent variables for plots.

The Topical Guide allows users to interactively

[2] In this case the token-level assignments seem to poorly reflect the dominant themes of the topic. While the MCMC sampling algorithm employed guarantees convergence at the distribution level, this guarantee does not extend to individual components of a sample, meaning that many topic assignments will fall outside of what humans would deem semantically coherent. Nevertheless, overall topic distributions capture an underlying reality recognized by human users (Chang et al., 2009).

[3] Pairwise metrics do not necessarily constitute "metrics" in the formal sense. We leave it to implementers to determine whether to satisfy the triangle inequality, etc.



Dataset: state_of_the_union / Analysis: lda100topics / Topic: programs federal / Document: 151

Harry_S._Truman_5.txt

DOCUMENT TEXT

Mr. President , Mr. Speaker , Members of the Congress : A year ago I reported to this Congress that the state of the Union was good. I am happy to be able to report to you today that the state of the Union continues to be good. Our Republic continues to increase in the enjoyment of freedom within its borders , and to offer strength and encouragement to all those who love freedom throughout the world. During the past year we have made notable progress in strengthening the foundations of peace and freedom , abroad and at home. We have taken important steps in securing the North Atlantic community against aggression. We have continued our successful support of European recovery. We have returned to our established policy of expanding international trade through reciprocal agreement. We have strengthened our support of the United Nations. While great problems still confront us , the greatest danger has receded--the possibility which faced us 3 years ago that most of Europe and the Mediterranean area might collapse under totalitarian pressure. Today , the free peoples of the world have new vigor and new hope for the cause of peace. In our domestic affairs , we have made notable advances toward broader opportunity and a better life for all our citizens. We have met and reversed the first significant downturn in economic activity since the war. In accomplishing this , Government programs for maintaining employment and purchasing power have been of tremendous benefit. As the result of these programs , and the wisdom and good judgment of our businessmen and workers , major readjustments have been made without widespread suffering. During the past year , we have also made a good start in providing housing for low-income groups ; we have raised minimum wages ; we have gone forward with the development of our natural resources ; we have given a greater assurance of stability to the farmer ; and we have improved the organization and efficiency of our Government. Today , by the grace of God , we stand a free and prosperous nation with greater possibilities for the future than any people ever had before in the history of the world. We are now , in this year of 1950 , nearing the midpoint of the 20th century. The first half of this century will be known as the most turbulent and eventful period in recorded history. The swift pace of events promises to make the next 50 years decisive in the history of man on this planet. The scientific and industrial revolution which began two centuries ago has , in the last 50 years , caught up the peoples of the globe in a common destiny. Two worldshattering wars have proved that no corner of the earth can be isolated from the affairs of mankind. The human race has reached a turning point. Man has opened the secrets of nature and

STATS

Number of tokens:	2181
Number of types:	1004
Topic Entropy:	4.49

Harry_S._Truman_4.txt	0.93
Harry_S._Truman_3.txt	0.92
Dwight_D._Eisenhower_3.txt	0.90
Dwight_D._Eisenhower_5.txt	0.89
Dwight_D._Eisenhower_7.txt	0.88
Dwight_D._Eisenhower_1.txt	0.88
Dwight_D._Eisenhower_2.txt	0.87
Dwight_D._Eisenhower_4.txt	0.86
Dwight_D._Eisenhower_8.txt	0.84
Harry_S._Truman_2.txt	0.83

Figure 3: A view of Harry Truman's fifth State of the Union address. Words belonging to the "programs federal" topic are highlighted in blue.

generate two types of plots. The first shows topic trends over the values of an attribute—for example, the year of an address, or the political party of the president delivering it. This kind of plot has been used to visualize topic models almost since their introduction (e.g. Griffiths and Steyvers, 2004). In our system they can be generated over any attribute for any topic or combination of topics in the corpus. Figure 3 shows trends for four chosen topics across the "year" attribute, giving a view of topic lifecycles across time as in Griffiths and Steyvers (2004), Wang and McCallum (2006), and others.

The second type of plot is a scatter plot of topic metrics, by which one topic metric is plotted against another. This is particularly suited to discovery of relationships between topic metrics. For example, we have found 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 yet implemented.

5 Topic Maps

With topic-to-topic relationships described by means of pairwise topic metrics, graph-based visualization of the topic space becomes straightforward. We construct a topic graph $G = (N, E)$ for a set of topics T as follows:

N is a set of $|T|$ nodes such that

$$\forall_{t \in T} weight(N_t) = \tau_1(t)$$

and

$$\forall_{t \in T} color(N_t) = \tau_2(t)$$

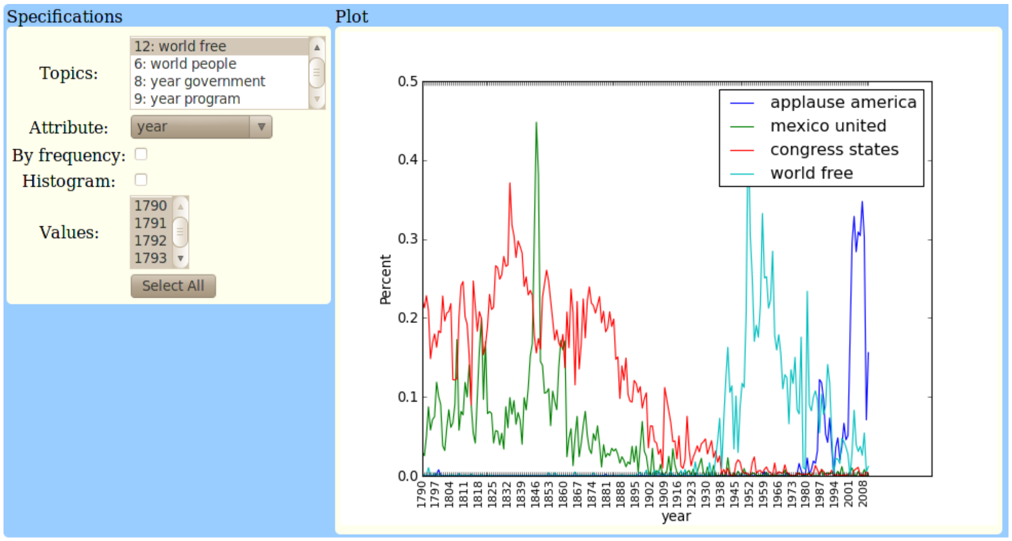
where τ_1 and τ_2 are topic metrics (potentially the same). E is constructed as a set of $|T|^2$ edges such that

$$\forall_{(t,u) \in T \times T, t \neq u} weight(E_{t,u}) = \pi(t, u)$$

where π is a pairwise topic metric.

We use the Gephi Toolkit[4] to generate the graphs and render them as images. We then integrate the visualization into the overall browsing experience to help users quickly assess how topics relate to each other. Figure 1 shows such a graph as part of a topic view. We employ a force-directed layout algorithm to arrange the nodes so that, generally speaking, nodes joined by edges of higher weight are closer together, and nodes joined by edges of lower weight

[4] http://www.gephi.org



Specifications

Topics: 12: world free, 6: world people, 8: year government, 9: year program

Attribute: year

By frequency: □

Histogram: □

Values: 1790, 1791, 1792, 1793

Select All

Figure 4: Selected topics over time as a percentage of overall tokens.

are further apart. (A similar approach focused on visualization of the document space is described in Newman et al. (2010a).) In the final rendering of the image, edges are omitted to reduce visual complexity. However, the distances between nodes are still determined by the interaction of the layout algorithm and the edge weights.

6 Topic Name Schemes

As LDA does not assign names to the topics it generates, automatic generation of topic names is of interest to researchers wishing to make topic models human-usable. Research in this area is ongoing (Mei et al., 2007; Lau et al., 2010). In order to facilitate investigations in this area, we equipped the Topical Guide with a fully pluggable topic naming system. Any number of topic name schemes can be used to produce names for all topics in a model. Within the user interface users can select a name scheme, which is then reflected throughout the interface. By default we use a concatenation of the two words with the highest probability for a given topic; alternative schemes can be imagined and easily implemented.

7 Data Import Backend

Automatically turning a raw document collection into an interactive browsing experience requires extensive preprocessing. Documents must be converted into a representation accepted by the topic model learner. Topics must then be inferred and model output indexed for quick retrieval by the web front-end. Additionally, metrics must be computed, topic names generated, and graphs rendered before they become accessible via the user interface. Unsurprisingly, the dependencies in the import process form a directed acyclic graph, represented in Figure 5. In order to make workflows simpler and more efficient, we implemented the import process using the `doit` task automation tool.[5] This allows full import of a dataset using a single invocation of the import script.

Enabling the Topical Guide to browse a new dataset can be as simple as choosing a name and description, converting the documents to plaintext, placing them in a directory, and providing a JSON file containing any document metadata. By default, the Topical Guide import system will automatically invoke MALLET's LDA implementation to train a topic model (McCallum, 2002); however, the import system can also make use of existing topic model output in the MALLET format. The system also allows all elements of the pipeline to be overridden, allowing for customization and clean integration with existing code.

[5] http://doit.sourceforge.net/

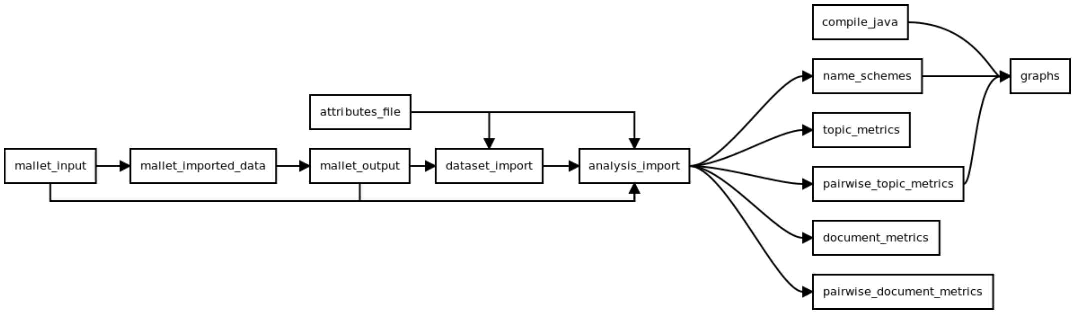


Figure 5: Dependencies within the data import pipeline. Arrows indicate that the source is prerequisite for the target. All stages in the import process shown here can be overridden, allowing clean, dataset-specific customizations.

8 Conclusion

We have presented the Topical Guide, an open-source web application for interactive, topic-centric exploration and visualization of topic model output. The Topical Guide demonstrates the ability of topic modeling to improve discoverability in and provide usable online access to otherwise obscure document collections. The free availability of such a tool also makes it easy to expose topic model output to third-parties for scrutiny, implicitly challenging the field to adopt a higher standard of transparency in future research.

References

David M. Blei and John D. Lafferty, 2009. *Topic Models*, chapter 4. Chapman & Hall/CRC.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

S. Brody and N. Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 804–812.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In *Coling 2010: Posters*, pages 605–613, Beijing, China, August.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 490–499, New York, NY, USA.

D. Mimno, H. M Wallach, J. Naradowsky, D. A Smith, and A. McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, page 880–889.

Hemant Misra, Olivier Cappé, and François Yvon. 2008. Using lda to detect semantically incoherent documents. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 41–48.

David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. 2010a. Invited paper: Visualizing search results and document collections using topic maps. *Web Semant.*, 8:169–175, July.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010b. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108.

Xuerui Wang and Andrew McCallum. 2006. 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*, KDD '06, pages 424–433, New York, NY, USA.

C. Wang, D. Blei, and D. Heckerman. 2008. Continuous time dynamic topic models. In *Proc. of UAI*.