Topic Modeling for the Social Sciences

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning and Daniel A. McFarland*
Computer Science Department and School of Education*
Stanford University
Stanford, CA 94305
{dramage,emrosen,jcchuang,manning,dmcfarla}@stanford.edu

Abstract

As textual datasets grow in size and scope, social scientists need better tools to help make sense of that data. Despite the natural applicability of topic modeling to many such problems, word counts and tag clouds are often used as the primary means of gleaning information from textual data. We characterize two barriers to adoption encountered during a collaboration between the Stanford NLP group and social scientists in the school of education: accessibility and trust. Accessibility refers to the technical barriers that make text processing and topic modeling difficult. Trust comes when practitioners can explore and validate a model being used to discover or support a hypothesis. We introduce recent work aimed at solving these challenges including the Stanford Topic Modeling Toolbox software.

1 Introduction

Topic models hold great promise as a means of gleaning actionable insight from the text datasets now available to social scientists, business analysts, and others. The underlying goal of such investigators is a better understanding of some phenomena in the world through the text people have written. In the Mimir project at Stanford, computer scientists in the natural language processing group have worked closely with social scientists in the school of education. During this interaction, we discovered two main barriers to adoption of topic models in the social sciences. The first is accessibility of the models—text processing is messy, with most existing tools assuming a reasonable familiarity with scripting, command line software invocation, and data pre-processing. While many social scientists are technically capable, fewer are proficient at all these prerequisites. In Section 2, we describe this issue in more detail, and introduce the Stanford Topic Modeling Toolbox as a step toward more accessible topic modeling for the social sciences.

The more central issue, perhaps, is trust. Ultimately, the intended usage of topic models is to tell a compelling story about textual data in order to support or inspire hypotheses. For example, a social scientist might wish to understand relationships between teens and teachers in online social networks. Armed with a corpus of text from a social networking site, these investigators may seek to uncover distinctions in teens' posts when they are or are not viewable by teachers. Topics can act as natural means to characterize these differences. But how can an investigator trust a system describing text that—by nature of the problem size—he or she has never read? This is a fundamental concern in topic modeling for text, which we consider in Section 3, arguing both for improved models to overcome existing shortcomings and better support for interactive exploration.

2 Accessible topic modeling through better software

One barrier to the adoption of richer text modeling techniques in the social sciences is a technical one—many existing toolkits presume a working understanding of basic text processing techniques for converting documents into sets of appropriately transformed words. Furthermore, many toolkits eschew the existing computing environments with which the many computer-literate social scientists are most familar, such as spreadsheets and statistical programming environments like R. Topic modeling researchers need to provide better interfaces to existing computing environments if they wish to expand their impact within the social scientists.

We developed the Stanford Topic Modeling Toolbox[1] as a step toward answering these challenges. The toolbox is intended for use by social scientists and other non-engineers who have modest scripting abilities, and it assumes no background in text processing. It contains implementations of collapsed Gibbs samplers for Latent Dirichlet Allocation [2] and for Labeled LDA [4]. The toolbox reads input text and associated metadata from comma-separated value files (CSV), and can generate rich outputs as CSV, ready to be analyzed or plotted back in a spreadsheet environment. The software is written in Scala, a language with script-like syntax that targets the Java virtual machine. The toolbox includes high-level primitives for describing text manipulation pipelines, enabling a few lines of readable code to load text from a column of a CSV file, tokenize, case-fold, filter rare and common terms, before removing documents with too few words remaining. Users can interact with toolkit through short Scala scripts, and can adapt the provided examples for common topic modeling related tasks. The software is used by several social scientists at Stanford and is able to produce the kinds of outputs used in later figures in this paper.

3 Trusting topic model output

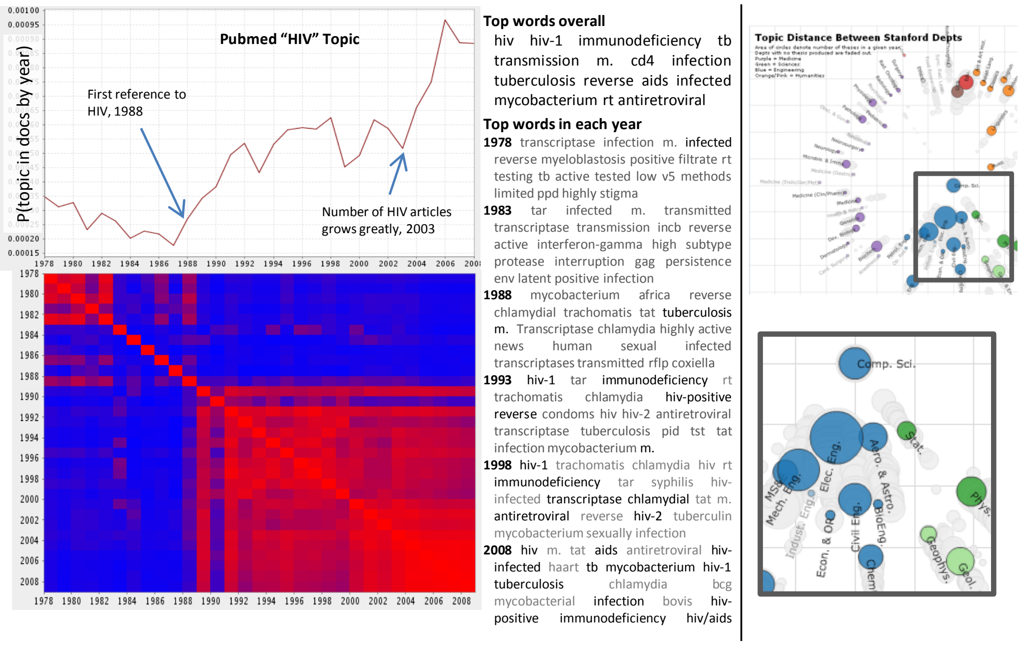
Trust is a recurring theme when using topic models as support or inspiration for social science hypotheses. How can we be sure of the trends discovered in the data? If they don't align with our intuitions, how do we know when it is the model or our intuitions that must change? Concretely, we outline some specific challenges in interpreting and ultimately trusting topic models. These challenges are known to varying extents within the topic modeling community and motivate our work. Where appropriate, we describe steps we have taken to address them and provide recommendations for other researchers interested in doing the same.

Characterizing topics is hard. Because topics are treated as hidden variables that represent latent dimensions in the data, LDA and most related models learn topics that have no inherent canonical descriptions. Commonly, the top-k most frequent words are used to describe the topic. At its worst, this characterization can mislead the investigator because each topic is a distribution over the full vocabulary. In practice, some documents will use a topic's low frequency terms disproportionately, so the top-k terms may be a poor representation of the topic in context. Topic modeling tools need to provide better information about *how* topics are used—both in individual documents and in aggregate—so that practitioners can make informed choices about how much to trust each topic's description. Solutions to this challenge may take the form of richer aggregate descriptors (such as in Figure 1 whose accompanying text is later in this section), better visualizations that support exploring topics' usage in individual documents, or new models that are easier to inspect.

Naming topics is hard. One specific type of characterization challenge is how to choose a simple, descriptive name for a given topic. Commonly, this is an ad-hoc process done by the practitioner after inspecting the topic's most common words. A recent alternative approach is to incorporate some supervision into the model so that the learned topics are designed to match labels determined in advance. Labeled LDA [4] (and the similar network-entity discovery model [3]) provide one promising avenue toward a solution when some document-level label information is available. Like, LDA, Labeled LDA treats each document as a mixture of topics, and each word is drawn from one of those topics. But unlike LDA, Labeled LDA does not treat the topic space as entirely latent—the set of topics for use on each document is taken as an observed variable, fixed in advance.[2] For example, web pages from *del.icio.us* that share a particular tag, news stories under a heading, or even documents that share a reference to the same entity (like [3]) will be tied together through shared use of some topics. The result is that the topics in these models have direct interpretability through the name of the label associated with each topic. But because not all dimensions of interest are necessarily labeled, the challenge of accurately naming latent dimensions remains.

[1] http://nlp.stanford.edu/software/tmt/

[2] Because the label space is observed, the authors of the Labeled LDA paper have noted that the model might better be referred to as *Blatant Dirichlet Allocation*.



Top words overall

hiv hiv-1 immunodeficiency tb transmission m. cd4 infection tuberculosis reverse aids infected mycobacterium rt antiretroviral

Top words in each year

1978 transcriptase infection m. **infected** reverse myeloblastosis positive filtrate rt testing tb active tested low v5 methods limited ppd highly stigma

1983 tar infected m. transmitted transcriptase transmission incb reverse active interferon-gamma high subtype protease interruption gag persistence env latent positive infection

1988 mycobacterium africa reverse chlamydial trachomatis tat **tuberculosis** **m.** Transcriptase chlamydia highly active news human sexual infected transcriptases transmitted rflp coxiella

1993 **hiv-1** tar **immunodeficiency** rt trachomatis chlamydia **hiv-positive** **reverse** condoms hiv hiv-2 antiretroviral transcriptase tuberculosis pid tst tat infection mycobacterium **m.**

1998 **hiv-1** trachomatis chlamydia hiv rt **immunodeficiency** tar syphilis hiv-infected **transcriptase chlamydial** tat m. **antiretroviral** reverse **hiv-2** tuberculin mycobacterium sexually infection

2008 **hiv** m. tat **aids** antiretroviral **hiv-infected** haart **tb mycobacterium hiv-1 tuberculosis** chlamydia bcg mycobacterial **infection** bovis **hiv-positive immunodeficiency hiv/aids**

Figure 1: *Left of vertical bar*: A sliced analysis of an "HIV" topic learned in a corpus of biomedical papers. The top words in the topic are shown in the top right, with the top words as used by the topic in selected years shown below. Words in common with the overall top words are bolded. The probability of the topic is plotted over time in the top left, with some substantial usage before the introduction of HIV into the medical literature in 1988. The bottom left shows the year-to-year cosine similarity of topic usage distributions. The block structure demonstrates that the topic converged to being mostly about HIV starting in 1988, with additional tightening in 2000 and again in 2003. *Right of vertical bar*: Viewing the output of Labeled LDA for department-department similarity based on Stanford dissertations: here percentage of words from each department that computer science dissertations borrow. The visualization supports drilling down to the level of individual dissertations.

Topics mean different things in different contexts. Although many topic models define their topics' term distributions as global variables, topics are not used uniformly across groups within a corpus. For example, consider Figure 1, which tracks the usage of a topic discovered using LDA on the open access subset of Pubmed, a database of biomedical publications. The figure demonstrates how the "HIV" topic (our name) evolves over time. Unless used with care, the graph of topic usage is inherently misleading because it assigns some substantial mass to the topic before the introduction of HIV as a concept in the medical literature (1988). Indeed, the words used in the early years of the graph pick up up more general references to transmitted infections, rather than specifically immunodeficiency diseases like HIV. Outputs like the figure can shed light on *how* a topic is used at varying levels of aggregation.

These figures were created by applying the same process usually done on the whole corpus to only the subset of documents present in each year. In particular, for a document group g's usage of topic k word j, we computed: $\beta_{k,j}^{(g)} \propto \sum_{d \in g} \#(z = k \wedge w = j)$ from the current point assignment (in Gibbs sampling) or counts (in variational estimates) associated with each z. It is worth noting that this example demonstrates how dynamic topic *modeling assumptions* [1] are not needed in order to get dynamic topic *usage* over time. In contrast, a recent trend in the literature has been toward topic models that are more aware of these document groupings, be it by incorporating time [1], multiple corpora [5], or labels [4]. However, even these richer models face the same underlying challenge: secondary aggregations of documents make use of shared topics in slightly different ways. We could as well have sliced a dynamic topic model's output by geographic region or by the author's educational background. The resulting usages may well differ substantially from the model's assumed generative distributions.

Topic models must find what we know is there. Ultimately, a topic model's trustworthiness must be determined by informed human judgments. In particular, the model must find the broad trends and facts known to be true by the practitioner of the domain. Without such support in finding the known, topic models have limited value in discovering the unknown—i.e. quantifying known trends or discovering unexpected ones. We believe a solution to this challenge lies in systems that enable better interactive exploration and sharing of topic model outputs as a means of validating that the model.

To explore these relationships, we built a tool for interactive visualization of the output of models from the topic modeling toolbox, as shown in the right half of Figure 1. The tool can be used to quickly validate a large set of a model's predicted relationships. Underlying the visualization, Labeled LDA is used to determine word distributions for each department at Stanford based on its dissertations (every dissertation is labeled with the department of all committee members). Next, each dissertation is allowed to be a mixture of all the per-department distributions (via regular LDA inference), in order to score every dissertation (and then department) according to how many words it likes to borrow from each other department. The browser enabled us to quickly test the model's behavior against our own intuitions: Do these departments belong closer together in this year than that one? Does this student's dissertation belong with these others? We used our own intuitions like a *training set* by updating the model until it answered our questions correctly. By sharing the visualization with others, we used their intuitions like a *test set*. In this way, we were able to leverage the visualization into a validation mechanism for our model, with final results still in preparation. We believe that trust-building interactions like these have an important role in the future adoption of topic modeling in the social sciences.

4 Conclusion

Both techniques and software need to mature in order for topic models to gain widespread adoption in the social sciences. This work has outlined several outstanding challenges to that end, and in some instances has described initial steps taken by our group and others to meet those challenges. The social sciences' demand for methodological rigor must also be satisfied: free parameters' choices must have warrants, topics and their usage should be characterizable, and results should be easily communicated visually. While complete solutions to the challenges outlined here remain open, one overarching theme has been made clear through our cross-disciplinary initiative: the present demand for techniques like topic modeling in the social sciences is strong and growing.

Acknowledgments

This research was supported in part by NSF grant NSF-0835614 CDI-Type II: What drives the dynamic creation of science? and in part by Stanford University via the Institute for Research in the Social Sciences' Initiatives Assessment project. We wish to thank our anonymous reviewers and the members of the Stanford Mimir Project team for their insights and engagement.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International conference on Machine learning*, pages 113–120, 2006.

[2] D. M. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[3] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: Augmenting social networks with text. In *Conference on Knowledge Discovery and Data Mining*, 2009.

[4] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore, August 2009. Association for Computational Linguistics.

[5] C. Wang, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In *The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 583–590, 2009.