**Citation:**

Brown, Travis and Jennifer Guiliano. "Topic Modeling for Humanities Workshop" National Endowment for the Humanities, Grant Submission, University of Maryland, College Park, MD, 2011.

**Licensing:**



**Collaborating Sites:**

University of Maryland
Maryland Institute for Technology in the Humanities

**Team members:**

Maryland Institute for Technology in the Humanities
Travis Brown
Jennifer Guiliano
Kirsten Keister
Amanda Visconti

**Acknowledgments**

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the collaborating institutions or the National Endowment for the Humanities.

## Topic Modeling For Humanities Research: Level I

**Enhancing the humanities through innovation:** Topic modeling is a statistical technique that attempts to infer the structure of a text corpus on the basis of minimal critical assumptions. One widely used topic model is Latent Dirichlet Allocation, which employs the following hypothetical story about how documents are created: we assume that each document is made up of a random mixture of categories, which we'll call *topics*, and that each of these topics is defined by its preference for some words over others. Given this story, we would create a new document by first picking a mixture of topics and then a set of words, by repeatedly choosing at random first one of the document's topics and then a word based on the preferences of that topic. This obviously isn't how documents are actually created, but these simple assumptions allow the topic model to work in reverse, learning topics and their word preferences by assuming that this story explains the distribution of words in a given collection of documents. A sub-area within the larger field of natural language processing, topic modeling in the digital humanities is frequently framed within a "distant reading" paradigm, drawing upon the 2005 work of Franco Moretti in *Graphs, Maps, Trees*. Moretti argues that the intersections of history, geography, and evolutionary theory create the potential for quantitative data modeling. Categorized as network theory, Moretti's approach to topic modeling utilizes aggregated data to explore macro level trends, themes, exchanges, and patterns. Yet, humanities scholars often need to focus simultaneously on the macro (distant/many texts) and the micro (close/individual texts). As a result, visualization between the "distant" aspect of the text's high-level attributes — its "topics" — and the "close" aspects the text — its individual words — are crucial. Latent Dirichlet Allocation (LDA) topic modeling provides a way to explore "distant" and "close" simultaneously. "LDA is a three-level hierarchical Bayesian model in which each item of a collection is modeled as a finite mixture over an underlying set of topics," wrote David Blei, Andrew Ng, and Michael Jordan in their 2003 article defining this field. "Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities." LDA offers an "unsupervised" topic modeling approach, in which no knowledge of the content of the text is really needed — the algorithm simply cranks away at whatever text corpus it is working on, and discovers topics from it — and a "supervised" approach where scholars "train" the algorithm by making use of domain knowledge. For example in a supervised LDA approach to Civil War newspapers, related pieces of knowledge coming from contemporaneous sources external to a corpus are used as additional data source. Casualty rate data for each week and the Consumer Price Index for each month allow the algorithm to potentially discover more "meaningful" topics if it has a way to make use of feedback regarding how well the topics discovered by it are associated with one of these parameters of interest. Thus, the algorithm can be biased into discovering topics that pertain more directly to the Civil War and its effects. Even this "supervised" version of LDA is not supervised in the sense generally used in machine learning, in which it indicates that the learner has been trained on labeled data and is attempting to assign these same labels to new data. As an unsupervised machine learning technique, LDA topic modeling does not require any form of expensive human annotation, which is often unavailable for specific literary or historical domains and corpora, and it has the additional benefit of handling transcription errors more robustly than many other natural language processing methods. Topic modeling aspires to discover global properties and qualities of the text, while at the same time connecting those global, macro-level qualities to micro-level detail, and is therefore likely to appeal to humanities scholars in a way that purely distant approaches do not. It is an approach that not only answers pre-existing research questions but also generates new questions. Despite—or perhaps because of—the relatively widespread use of topic modeling for text analysis in the digital humanities, it is common to find examples of misapplication and misinterpretation of the technique and its output. There are a number of reasons for this: existing software packages generally have a significant learning curve, most humanists do not have a clear understanding of the underlying statistical methods and models, and there is still limited documentation of best practices for the

application of the methods to humanities research questions. As a result, the most promising work in topic modeling is being done not by humanists exploring literary or historical corpora but instead by scholars working in natural language processing and information retrieval. These scholars, even as they have generated promising new avenues of research, have recognized topic modeling as "something of a fad" and suggested that more attention should be paid to the wider context of latent variable modeling approaches. This one-day workshop will facilitate a unique opportunity for cross-fertilization, information exchange, and collaboration between and among humanities scholars and researchers in natural language processing on the subject of topic modeling applications and methods. Recent work in natural language processing has particular relevance for research questions in the humanities, including a range of extensions of the basic LDA model that incorporate time and geography. Our intent is to begin to repair the divide between humanities using topic modeling approaches/software and those developing and utilizing them in computer science and natural language processing. Our primary goals will be: 1) greater familiarity with the interpretation and vocabulary of LDA topic modeling (and other latent variable modeling methods) for scholars in the humanities; 2) a deeper understanding of literary and historical corpora and their role as data within topic modeling; and 3) increased involvement in articulating fundamental research questions for researchers developing the models and methods (as well as the software implementations). **Environmental scan:** Existing work in topic modeling and the digital humanities follows two major LDA approaches: synchronic, where the unit of analysis is not time bound, and diachronic, where the unit of analysis includes a measurement of time. Examples of synchronic work include Jeff Drouin's exploration of Proust and Brown's own work on Byron's narrative poem *Don Juan* and Jane Austen's *Emma* while examples of diachronic work include David Newman and Sharon Block's work on the *Pennsylvania Gazette*, Cameron Blevins' work on *The Diary of Martha Ballard*, and Robert Nelson's "Mining the *Dispatch*". While all effective in their conclusions, each speaks to its own content analysis more than they speak to innovations in pedagogy, approach, and methodology within LDA. Previous events that have focused on topic modeling have been dominated by workshops organized by computer scientists and information retrieval specialists working in natural language processing. These workshops tend to provide low-level technical explorations of particular machine learning approaches, which are obviously not tailored to the training or expertise of general humanities audiences. When the concerns of humanists do intersect with these events, it is often through presentations by computer scientists using humanities' derived corpora. An example of these phenomena is the 2010 workshop in Natural Language Processing Tools for the Digital Humanities presented at Stanford University during the annual Digital Humanities Conference. Taught by Christopher Manning, a computational scientist, the workshop was a "survey what you can do with digital texts, starting from word counts and working up through deeper forms of analysis including collocations, named entities, parts of speech, constituency and dependency parses, detecting relations, events, and semantic roles, co-inference resolution, and clustering and classification for various purposes, including theme, genre and sentiment analysis. It will provide a high-level not-too-technical presentation of what these tools do and how, and provide concrete information on what kinds of tools are available, how they are used, what options are available, examples of their use, and some idea of their reliability, limitations, and whether they can be customized." Significantly, this effort to "empower participants in envisioning how these tools might be employed in humanities research" did not close the feedback loop to computational science to imagine how natural language processing tools, including topic modeling software, can be improved to deal with humanities research questions. When humanists are interacting with topic modeling approaches, it is often as an uncritical consumer rather than as an engaged critical applied theorist. **History and duration of the project:** This workshop has received no previous support. Preliminary research on topic modeling, LDA, and the humanities has been undertaken by Primary Investigator Travis Brown, Research and Development Software Development Lead

at the Maryland Institute for Technology in the Humanities, for a one-year period prior to this application. Brown has been engaged with national dialogues about topic modeling undertaken by computer scientists and information retrieval specialists and has also participated in humanists' discussions of topic modeling via his roles at the Walt Whitman Archive and MITH. He is, as a result, uniquely positioned to facilitate the cross-fertilization process. University of Michigan Graduate Student Sayan Bhattacharyya, and UMD Graduate Student Clay Templeton, who have served as Interns in topic modeling at MITH via an Institute for Museum and Library Services Internship grant in Summer 2011, will aid him. Through their work on Woodchipper, a visualization tool for humanities usage that allows the user to search and select text from participating collections and display relationships among texts, Bhattacharyya and Templeton have aided Brown in identifying thematic areas where cross-fertilization of knowledge about topic modeling needs to occur between humanists, computer scientists, and information retrieval specialists. **Work plan:** The workshop will be organized into three primary areas: 1) an overview of how topic modeling is currently being used in the humanities; 2) an inventory of extensions of the LDA model that have particular relevance for humanities research questions; and 3) a discussion of software implementations, toolkits, and interfaces. Each area will be covered in a two-hour long session with two or three individual speakers giving 30-minute presentations. The initial overview will explore examples of topic modeling approaches currently being used in text analysis projects in the humanities. Potential speakers include, but are not limited to: Matt Jockers, Sharon Block, and Robert Nelson. The overview of extensions will cover a range of variants of the widely-used Latent Dirichlet Analysis topic model that are able to take into account time, geography, and other information about the documents being analyzed and their context, and may include speakers such as Jordan Boyd-Graber, Doug Oard, and Jason Baldridge. The final implementation session will focus on the development and explication of tools such as the Machine Learning for Language Toolkit (MALLET). Potential speakers include, but are not limited to, David Mimno and Taesun Moon. Each area session will culminate in a 30-minute exercise to identify areas of overlapping interest for further development. The workshop will close with an additional 45-minute session that will focus on extrapolating from the individual sessions to a larger understanding of how topic modeling approaches can advance humanities scholarship. **Staff:** The proposed workshop on topic modeling is fortunate to benefit from a variety of substantial relationships at the University of Maryland and MITH. Core project staff will include: Travis Brown, lead Research and Development Software Developer at MITH, who will lead the project and supervise all project activities; University of Michigan Graduate Student Sayan Bhattacharyya and UMD Graduate Student Clay Templeton will work with Mr. Brown to develop an appropriate cyber-environment to gather all associated publications, software, and presentation materials for workshop events; Dr. Jennifer Guiliano, Assistant Director of MITH, will provide logistical support for all workshop related activities including handling all local arrangements and handle fiscal reporting activities; Emma Millon, Community Lead at MITH, will be responsible for all community outreach including distribution of the workshop solicitation, documenting workshop activities via social media, and aiding Brown in completing the white paper. **Final product and dissemination:** We will document publicly the workshop and all associated presentations thereby encouraging other researchers to join our community, benefit from our investment of resources, and extend the discussions related to topic modeling. Using twitter, blogs, and video feeds, we will provide synchronous and asynchronous methods of workshop involvement. By utilizing the workshop website as an opportunity to create a public presence around topic modeling and the humanities, we hope to extend our impact by providing a space for scholars to engage pre- and post-workshop. To aid this, we will release a reflective white paper at the end of the grant documenting the various sub-areas within topic modeling in the digital humanities and attempt to extrapolate potential understanding of how topic modeling efforts can extend itself into humanities scholarship via specific recommendations for further development.