Final Performance Report

Grant # HD5162712

Topic Modeling for Humanities Research

Project Director: Jennifer Guiliano

University of Maryland

May 1, 2013

Final Performance Report

Topic Modeling for Humanities Research, a one-day workshop directed by Assistant Director of MITH Dr. Jennifer Guiliano, received a Level 1 Digital Humanities start up from the National Endowment for the Humanities on April 19, 2011. The workshop facilitated 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. The workshop was 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.

Of particular note in this final review is the completion of all project goals including the workshop itself on November 3, 2012 at the Maryland Institute for Technology in the Humanities (MITH) on the University of Maryland College Park campus.

**Project Activities**

As outlined below as of May 1, 2013, the Topic Modeling Workshop has met all goals within the allotted time. These goals with the outlined accomplishments are listed below in table format for ease of reading.

|  |  |
| --- | --- |
| Initial Goals | Actual Accomplishments |
| Establishment of Topic Modeling Website | Completed May 8, 2012.  |
| Finalized List of Speakers | Initial list completed as part of grant submission. Finalized list with all invitations sent to potential speakers June 15, 2012. |
| Call for Participation for Event | The initial call for participation was announced on July 10, 2012 to major listservs, public boards, and professional organizations as well as social media outlets. Applications consisted of a two-page statement of interest and a c.v. |
| Notification of selected attendees | Selected attendees were notified on August 27th and 28th, 2012. Thirty-six submissions were received and accepted.  |
| Completion of pre-workshop materials | Pre-workshop materials including a topic modeling Zotero group (45 members), a 215 item bibliography of resources, and establishment of a #dhtopic twitter tag were completed September 15, 2012. |
| Workshop | The workshop was slated to be held November 3rd or 4th, 2012. It was held as scheduled on November 3, 2012. A full agenda is attached in Appendix A. |
| Post-workshop materials | MITH elected to create a post-workshop sandbox environment for workshop attendees to use for their topic modeling explorations. That library and toolkit was activated on December 18, 2012. It can be reviewed at: https://github.com/umd-mith/topic-modeling  |
| Post-workshop survey | The post workshop survey was completed November 10, 2012. That data has been tabulated and is shared here in the survey appendix. |

**Accomplishments**

We are quite proud to report a number of significant accomplishments related to this project. First, as promised, we were able to gather a cross-section of scholars using topic modeling approaches and those developing topic modeling software and research methodologies. Of particular note were our speakers which included: Matthew Jockers of the Department of English and Center for Digital Research at the University of Nebraska (author of *Macroanlyses*), Robert Nelson of the Department of American Studies and Digital Scholarship Lab at the University of Richard (lead investigator of *Mining the Dispatch*), and Jo Guldi of the Department of History at Brown University (co-developer of *Paper Machines*). Speaking to the computational side of the equation, we were delighted to have Jordan Boyd-Graber of the School of Information Studies and Institute for Advanced Computer Studies of the University of Maryland, David Mimno of the Department of Computer Science at Princeton University (co-developer of the MALLET topic modeling software) and David Blei of the Department of Computer Science at Princeton University (co-founder of the field of Computational Approaches to Topic Modeling) join us.

We were also able to achieve our goal of a deeper understanding of the role of literary and historical corpora and their role as data within topic modeling by encouraging a frank discussion of the challenges and opportunities presented by individual datasets. In conversations, participants noted the need for the digitization of texts "on the margin", a shared repository of "clean" data that can be easily leveraged for experimental purposes, partnerships with bibliographers and scholars that focus on the history of the book to address the creation of new corpora, and the need to incorporate born-digital materials into corpora creating as they are generally easier to prepare for topic modeling tools. This was coupled with our desire to provide a greater familiarity within the interpretation and vocabulary of LDA topic modeling. We are pleased to report that we were able to create a number of educational materials and a sandbox environment for people to learn about and experiment with topic modeling that allowed humanists who were just beginning to explore topic modeling approaches an opportunity to educate themselves and experiment without having to fully-commit to data collection, cleaning, and tool installations.

**Audience**

When coupled with the attendees (itemized in appendix B: attendee list), the project was able to gather roughly 95% of humanists working in topic modeling. Our gender split was 19 of 52 attendees or roughly 37%. We also captured the current status of the attendee as well as geographic locale.

Figure 1: Attendees by Type

Figure 2: Attendees by State

We provided a number of ways for the results of this project to be extended to the general public as well as those unable to attend:

1. through our social media feed on twitter (hashtag: #dhtopic). Our social media feed included 1585 unique tweets, 243 links to topic modeling related materials and 423 retweets stretching from October 26th through April 12, 2013, almost five months after the close of the workshop
2. via our publicly available twitter list of all attendees and presenters (available at: https://twitter.com/UMD\_MITH/dhtopic)
3. via our workshop website (available at: http://www.mith.umd.edu/topicmodeling/)
4. via a Zotero group "Digital Humanities Topic Modeling". The group currently has 49 members, is open to the general public, and contains 216 unique bibliographic items for those interested in topic modeling (available at: https://www.zotero.org/groups/digital\_humanities\_topic\_modeling)
5. via our Vimeo channel which provides recordings of the presentations. Direct links are as follows:
* Jo Guildi and Chris Johnson-Roberson: http://vimeo.com/53078693
* Matt Jockers and Rob Nelson: http://vimeo.com/52959139
* David Mimno: http://vimeo.com/53080123

**Continuation of the Project and Long-term Impact**

The qualitative feedback we received from the Office of Digital Humanities with regards to submissions received for its grant competitions following the workshop highlights the short-term impact of this workshop. Brett Bobley reports that they have seen a significant increase in the number of applications from scholars working in topic modeling. This increase in requests for support, we believe, highlights the effectiveness of the event that our attendees were motivated by their attendance to seek additional support for the next stage of their research. It suggests that the workshop will have a long-term influence on forthcoming research efforts.

Lastly, a significant portion of our discussion explored fundamental research questions for researchers in developing the models and methods to apply to topic modeling and humanities corpora. Below is a items articulated as part of this discussion as potential areas for future discussion, research, and resource allocation:

* partnerships with XSEDE or other high performance computing centers that can rapidly load and develop models for use with differing workflows.
* the publication of workflows used by scholars to generate their results
* the publication of documentation of models to allow for critique and evaluation.
* the documentation of provenance of data, models, and results
* a need for a humanities-centric model that is non-linguistically/scientific driven
* the identification and documentation of assumptions that impact topic models
* additional training opportunities in statistics
* additional training opportunities in pre-modeling workflows including data harvesting and preparation for the differing tools and models
* the publication of script libraries and documentation of "simple" actions that humanists can use to jump-start their ability to adopt MALLET
* the creation of a digital space for humanists with interesting data and research questions to solicit involvement from computer scientists
* the creation of a topic modeling handbook that will address responsive actions a humanist could take when their results do not fit their needs
* the identification and creation of a community of topic modeling scholars interested in partnerships with librarians who can assist with topic modeling projects using library resources
* the creation of standard workflow pipelines and methodologies for humanist use
* the facilitation of sustainable spaces for humanist-coders to undertake their work.

**Appendix A: Agenda**

9:00-9:30 am Registration and Refreshments

9:30 am-9:45 am Welcome and Setting of Workshop Goals: Jennifer Guiliano, Maryland Institute for Technology in the Humanities

Humanities Applications

9:45-10:15 am Presentation: Matthew Jockers, Department of English and Center for Digital Research in the Humanities, University of Nebraska

Title: Thematic Change and Authorial Innovation in the 19th Century Novel

Abstract: This presentation explores my use of latent Dirichlet allocation (topic modeling) to extract 500 themes from a corpus of 3,500 19th-century novels. By linking the derivative topical data to bibliographic metadata about the texts (their publication dates but also information about the gender and nationality of the authors) macroscale trends across time, nationality, and gender can be observed and employed to contextualize anecdotal readings of the individual texts in the corpus. The approach allows us to see and measure thematic innovation across the century and also according to author nationality and author gender. In the end, this larger context of thematic change allows us to better understand and assess the contributions of specific writers such as Jane Austen, Hermann Melville, and Maria Edgeworth.

10:15-10:45 am Presentation: Robert Nelson, Department of American Studies and Digital Scholarship Lab, University of Richmond

Title: Analyzing Nationalism and Other Slippery “Isms”

Abstract: Franco Moretti’s concept of “distant reading” has been all but ubiquitous in the rationales historians, literary critics, and other digital humanists have provided for different text-mining techniques. Practitioners of text mining have seized upon both Moretti’s methodological challenge and his memorable phrase, claiming that text mining offers the tantalizing prospect of addressing hugely ambitious interpretive questions by using computation to analyze massive amounts of digitized text. This presentation will take a somewhat different tack. Using topic modeling to chart the relative prominence of nationalistic rhetoric in two newspapers during the course of the American Civil War, this presentation will use those charts to infer the particular instrumental political purposes of that rhetoric for different individuals and groups at different moments in time. In doing so, it will suggest that topic modeling enables us to develop subtler—one might even say closer—readings of something as abstract and amorphous as an ideological discourse like nationalism.

10:45-11:30 am Discussion (all): Facilitator, Katrina Fenlon, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

11:30-11:45 am Break

Models and Extensions

11:45-12:15 pm  Presentation: Jordan Boyd-Graber, School of Information Studies and Institute for Advanced Computer Studies, University of Maryland

Title: Incorporating Human Knowledge and Insights into Probabilistic Models of Text

Abstract: Topic models aren’t perfect; errors hamper adoption, degrade performance in downstream computational tasks, and prevent users from making sense of large datasets.  I begin by first explaining that the answer is not simply to build fancier statistical models, as the usefulness of topic models does not always correlate with likelihood, the traditional objective function of statistical models.  Given this understanding, what is possible to improve topic models?

As an attempt to answer this question, I will present models incorporate human knowledge into topic models via ontologies, through direct interaction with users, or by mimicking social processes.  After describing the statistical formalisms that allow these knowledge repositories to be seamlessly integrated and the associated computational challenges, I demonstrate how these models can contribute to real-word natural language processing tasks such as classifying documents, predicting sentiment, topic segmentation, and detecting who is influential in a conversation.

12:15- 12:45 pm Discussion (all): Facilitator, Scott Weingart, School of Library and Information Science, Indiana University

12:45-2:00 pm Catered Lunch: 6137 McKeldin Library

Implementations and Tools

2:00- 2:30 pm Presentation: Jo Guldi, Department of History, Brown University and Christopher Johnson-Roberson, Ethnomusicology, Brown University
Title: Paper Machines: A Tool for Analyzing Large-Scale Digital Corpora
Abstract: Paper Machines,developed in the summer of 2012 through a Harvard grant and a grant from Google Summer of code, is designed to help scholars parse through large sets of information, capitalizing on current work in computer science, drawing upon current work in topic modeling and visualization to generate iterative, time-dependent visualizations of how what a hand-curated body of texts talks about and how it changes over time. At a conceptual level, we believe that this tool will be a powerful resource in examining large quantities of information and allow knowledge seekers to consider a broader, richer, often ignored corpus of text. In doing so, we hope to enlist the power of digital humanities to tame the pile of paper, and redistribute the power that “official” paper took away.

2:30- 3:00 pm Presentation: David Mimno, Department of Computer Science, Princeton University

Title: The details: how we train big topic models on lots of text
Abstract: It’s possible to treat a topic model as a black box: text goes in, topics come out. But to really understand what a model is telling you about your corpus, you need to understand what the algorithm is actually doing. In this talk I’ll cover some of the different choices we can make in training models and what their implications are for efficiency, scalability, and topic quality. As a running example, I’ll use the Mallet topic modeling package. Issues covered will include Gibbs sampling, variational inference, hyperparameters, and automated quality diagnostics. No particular mathematical background is required.

3:00- 3:30 pm Discussion (all): Facilitator, Travis Brown, Maryland Institute for Technology in the Humanities

3:30-3:45 pm Break

3:45-5 pm Topic Modeling in the Humanities Roundtable Discussion (all): Facilitator, David Blei, Department of Computer Science, Princeton University

**Appendix B: Attendee Roster**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Last Name** | **First Name** | **University** | **Academic title** |
| Ms. | Alami | Sarita | Emory University | PhD Candidate in History / Mellon Graduate Fellow at Emory's Digital Scholarship Commons |
| Ms. | Auvil | Loretta Sue | University of Illinois at Urbana Champaign | Senior Project Coordinator |
| Mr. | Bailey | Jefferson Jacob | Metropolitan New York Library Council | Strategic Initiatives Manager |
| Dr. | Black | Chad Thomas | University of Tennessee | Associate Professor, History |
| Dr. | Blei | David | Princeton University | Associate Professor, Computer Science |
|  | Bobley | Brett | National Endowment for the Humanities | Director, Office of Digital Humanities |
| Dr. | Boyd-Graber | Jordan | University of Maryland | Assistant Professor |
| Mr. | Brown | Travis | University of Maryland | Assistant Director, MITH |
| Mr. | Burton | Matt | University of Michigan | Doctoral Candidate |
| Dr. | Cecire | Natalia Aki | Yale University | ACLS New Faculty Fellow |
| Ms. | Chang | Shian | University of Maryland | IT Architect |
| Dr. | Cordell | Ryan | Northeastern University | Assistant Professor |
| Dr. | Croxall | Brian | Emory University | Digital Humanities Strategist and Lecturer of English |
| Ms. | Fenlon | Katrina | University of Illinois at Urbana Champaign | Doctoral Candidate |
| Dr. | Ferster | Bill | University of Virginia | Research professor |
| Dr.  | Fraistat | Neil | University of Maryland | Director, MITH |
| Dr. | Garfinkel | Susan | Library of Congress | research specialist |
| Mr. | Gil | Alex | Columbia University Libraries | Digital Scholarship Coordinator |
| Ms. | Glasgow | Kimberley | university of maryland college park |  |
| Dr. | Goodwin | Jonathan | University of Louisiana at Lafayette | Assistant Professor of English |
| Dr. | Guiliano | Jennifer | University of Maryland | Assistant Director, MITH |
| Dr. | Guldi | Jo | Brown University | Assistant Professor |
| Mr. | Hollis | Luke Robert | N/A |  |
| Dr. | Jockers | Matt | University of Nebraska | Assistant Professor |
| Mr. | Johnson-Roberson | Christopher | Brown | Doctoral Candidate |
| Prof. | Kill | Melanie | U of Maryland | Assistant Professor |
| Mr. | Kirkpatrick III | Robert H. | Yale University | Academic Technology Specialist |
| Prof. | Klein | Lauren | Georgia Institute of Technology | Assistant Professor |
| Prof. | Kleinman | Scott | California State University, Northridge | Professor |
| Mr. | Knox | John Thomas | University of South Carolina | Doctoral Candidate |
| Mr. | Kochis | Matthew | University of Tulsa | Doctoral Candidate |
| Dr. | Koeser | Rebecca Sutton | Emory University |  |
| Mr. | Leonard | Peter | UCLA |  |
| Mr. | Matienzo | Mark A. | Yale University Library |  |
| Dr. | Mimno | David | Princeton University | Postdoctoral Researcher |
| Dr. | Murphy | James Stephen | Unaffiliated |  |
| Dr. | Murray-John | Patrick | George Mason University | Research Assistant Professor |
| Dr. | Nelson | Rob | University of Richmond | Director, Digital Scholarship Lab |
| Mr. | Nguyen | Thang Dai | U of Maryland |  |
| Mr. | Padilla | Thomas George | University of Illinois at Urbana Champaign | Center for Informatics Research in Science and Scholarship Graduate Research Assistant |
| Dr. | Remy | Eric | Gettysburg College | Director, Instructional Technology |
| Ms. | Rhody | Lisa Marie | University of Maryland | Doctoral Candidate |
| Dr. | Rice | Claiborne | University of Louisiana at Lafayette | Assistant Professor |
| Ms. | Schleser | Jamie | American University | PhD Fellow |
|  | Serventi | Jennifer | National Endowment for the Humanities | Senior Program Officer |
| Mr. | Summers | Edward Hugh | Library of Congress | Information Technology Specialist |
| Ms. | Visconti | Amanda | University of Maryland | Webmaster and Ph.D. Student |
| Ms. | Walden | Justine Angelique | Yale University | Ph.D. Candidate |
| Mr. | Weingart | Scott | Indiana University | Doctoral Candidate |
|  | Williams | Kristin | George Washington University |  |
| Mr. | Wolff | Mark | Hartwick College | Associate Professor |