White Paper

Grant # HD5162712

Topic Modeling for Humanities Research

Project Director: Jennifer Guiliano
University of Maryland

July 20, 2013
White Paper

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.

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 Macroanalyses), 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 creation 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.

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


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.

We’d also like to highlight products produced by workshop attendees and the larger digital humanities community responding to the event:

- Some reflections by Thomas Padilla
- A wrap-up by Sarita Alami
- Some questions by Trevor Owens (Available from: http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/)
- A collection of Tweets by Scott Kleinman
- Detailed notes by Brian Croxall (Available from: https://docs.google.com/document/d/1T12WHhCvORnOXr0dk7VXHSCRc2uUAW4Be9dLlg
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 list of 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.

Several of these issues and possibilities were the focus of further investigation and development work at MITH after the workshop, and are discussed in more detail in the following section.

**Technical Challenges and Opportunities**

**More flexible tools for working with topic modeling results**

In the final sessions of the workshop, we asked participants to identify obstacles that they had faced (or anticipated facing) in the application of topic modeling techniques in their own research. One of the most significant technological gaps that this exercise highlighted was the lack of intermediate tools for working with the output of topic modeling implementations. In the case of MALLET, for example, which is currently the most widely used implementation of LDA in the humanities, on the one hand there is the command-line application itself, which can be a challenge to configure effectively, and which produces output in formats that are likely to be opaque to many humanities scholars; on the other hand, there is also a growing ecosystem of high-level, visualization-oriented applications (including Paper Machines and MITH’s own Woodchipper) that build on the MALLET library to provide new ways for users to explore their corpora and identify relationships between documents.
In many cases the humanities scholar may be interested in questions that neither category of application allows them to answer easily. Suppose for example that he or she has run MALLET and has noticed the following topic in the command-line output that scrolls by during the training process:

eloquence style great manner public subject taste writing genius ...

The scholar may then want to know the relative importance of each of these words in the topic, or to identify which documents in his or her corpus are most strongly associated with it. It is possible to have MALLET produce output that will get him or her closer to an answer to at least the latter question (using the --output-doc-topics flag), but the format is difficult to work with without additional processing:

#doc name topic proportion ...
0 fc-000001 18 0.5907526455111526 2 0.4029643419420837 8 9.3640340201027E-4 ...
1 fc-000002 1 0.43632157542165595 7 0.2821595426868692 2 0.16166496467260072 ...

Here we can easily identify the topics that a given document is most strongly associated with, and we have all the data we need to identify the documents that are most strongly associated with a given topic, but we cannot easily make that kind of query without parsing the file. For scholars who do not have expertise in a scripting language (or access to someone who does), this can be an immediate dead-end.

MITH had encountered this problem in several projects before the workshop, and in early 2012 the MITH development team had begun building a set of utilities designed to make working with the output of MALLET easier for non-programmers. Specialized versions of these utilities have played a key role in projects such as Amanda Visconti’s analysis of citation networks in Digital Humanities Quarterly, Winnemore Fellow Lisa Rhody’s Review, Revise, Requery, and Faculty Fellow Peter Mallios’s Foreign Literatures in America. When the feedback from workshop participants made it clear that other scholars and projects would benefit from these kinds of tools, the MITH development team generalized and documented the utilities, and published them as a library on GitHub in December 2012.

One of the most important pieces of functionality provided by this library is the ability to take the binary “model file” generated by MALLET and create an Excel spreadsheet capturing much of the information contained in the model, in a format that is easy to manipulate. Given the default configuration, for example, the spreadsheet will include a worksheet with each row providing the smoothed probability distribution over topics for each document. This view allows the user to sort documents by their association with a particular topic with a few simple user interface actions. Other worksheets provide the word distributions for each topic and document similarities based on the symmetrized Kullback-Leibler divergence between topic distributions. The library is also designed to be easily extensible, so that for example adding a worksheet that provides the Hellinger distance between documents (instead of K-L divergence) would be a matter of writing a dozen lines of code.

Managing and documenting experiments

Another technical challenge identified by workshop participants involved the management and
documentation of experiments. Many different kinds of design decisions are required to run MALLET on a corpus, for example, including how to divide the text into documents, whether to lemmatize words or perform other kinds of preprocessing, what words to remove before analysis (i.e., stopwords), how many topics to generate, and whether to use hyperparameter estimation. Because the methods used for inference in LDA are randomized algorithms, it can also be necessary to document the random seeds that are used by the topic modeling engine, if being able to reproduce experiments exactly is important. Even in a workflow where all of the relevant information about inputs is being captured and documented, it can be useful to store particular representations of the output (for example the binary model file that MALLET can be configured to generate) in order to avoid having to perform expensive or time-consuming processes multiple times.

MITH has drafted best-practice guidelines for experiment documentation in the context of several specific projects, but the general formulation of the problem—how do digital humanists manage, record, and share the quantitative and algorithmic aspects of their work?—is largely an open question, and its answer will depend on both the gradual emergence of scholarly conventions and the development of new platforms and tools.

**Graphical user interfaces**

There was active debate at the workshop about the role and importance of graphical user interfaces in the context of text analysis in the humanities. A number of participants argued for the necessity of “gateway GUIs” that would make text analysis more accessible to scholars who are not comfortable with the command line. Other participants argued that the ability to install, configure, and run command-line tools is currently essential for this kind of work, and that GUI applications are less flexible and harder to compose into complex workflows (in addition, of course, to being more expensive to develop).

MITH’s approach to these issues has been pragmatic: the libraries and utilities we have developed for topic modeling and other kinds of text analysis have tended to be command-line applications, and we have expected our close collaborators (such as interns, graduate students, and fellows) to familiarize themselves with command-line navigation and tools. These decisions have been made on the basis of project resources and scope, however, and we see a value in the development of graphical tools that would bring these kinds of methods to a wider audience.

**Virtual machine sandboxes**

In preparation for the workshop, MITH created an Amazon Web Services EC2 (Elastic Cloud Computing) virtual machine image with a number of data sets, text analysis tools, and programming language compilers and interpreters pre-installed and configured. We started an instance of this machine on the day of the workshop, distributed accounts to interested participants, and maintained the instance for two weeks after the workshop. Although we are not currently running the sandbox instance, we are maintaining the machine image, and have used it (often with modifications) in subsequent collaborations. This approach to designing, maintaining, and reusing EC2 machine images has proven useful in a wide range of MITH projects, and we believe that the possibility of promoting the use of particular digital
methods (such as topic modeling) through the documentation and distribution of specialized virtual machine images deserves further investigation.

**Topic modeling digitization artifacts to support corpus curation**

Most successful applications of topic modeling in the humanities to date have focused on exploration and discovery: they have enabled scholars to navigate their corpora in new ways, and to find patterns that they may not have been able to recognize otherwise. One thread of conversation at the workshop, however, involved the possibility of using topic modeling to aid in the curation of large text collections. The [HathiTrust Digital Library](https://www.hathitrust.org), for example, contains over ten million digitized volumes, most of which have been transcribed using OCR (optical character recognition) engines with no human correction, and therefore contain many misspellings and other digitization artifacts. Because of the way these digitization artifacts tend to occur, they are often identified as “topics” by LDA, even though they do not necessarily have any semantic coherence. For example, running LDA on a HathiTrust collection that contains books published in the eighteenth-century is likely to result in at least one tall non-terminal s topic (e.g., “fo fome fuch fame thofe thefe”). Topics representing other kinds of OCR errors (e.g. “tlie tliat tliese”, where the “th” is recognized as “tli”) or containing primarily halves of hyphenated words are also common. While many researchers using topic modeling have noticed these effects, they have received little (or no) attention in the topic modeling literature.

These methods could be used to prioritize pages for human correction, to find instances of maps, musical notation, and other non-textual materials (which are often characterized by a particular OCR error signature), or to identify passages of foreign languages within a primarily English text. Careful configuration of stopword lists and tokenization could lead to the identification of other applications. MITH is currently working on two pilot projects designed to explore some of these possibilities.