NATIONAL ENDOWMENT FOR THE Humanities

OFFICE OF DIGITAL HUMANITIES

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the Office of Digital Humanities program application guidelines at http://www.neh.gov/grants/odh/digital-humanities-start-grants for instructions. Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: Topic Modeling for Humanities Research

Institution: University of Maryland, College Park

Project Director: Jennifer Guiliano

Grant Program: Digital Humanities Start-Up Grants, Level 1

NEH Application Cover Sheet
Digital Humanities Start-Up Grants

PROJECT DIRECTOR

Dr. Jennifer Elizabeth Guiliano
Assistant Director
B0131 McKeldin Library
University of Maryland
College Park, MD 20742-7011
UNITED STATES

E-mail: guiliano@umd.edu
Phone(W): 301-405-9528
Phone(H):
Fax: 301-314-7111

Field of Expertise: Interdisciplinary

INSTITUTION

University of Maryland
College Park, MD UNITED STATES

APPLICATION INFORMATION

Title: *Topic Modeling for Humanities Research*

Grant Period: From 5/2012 to 4/2013
Field of Project: Interdisciplinary
Description of Project: Topic Modeling for Humanities Research, a 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. 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.

BUDGET

Outright Request	$24,808.00	**Cost Sharing**	
Matching Request	$0.00	**Total Budget**	$24,808.00
Total NEH	$24,808.00		

GRANT ADMINISTRATOR

Ms. Stephanie Swartz
Contract Administrator
Office of Research Administration and Advancement

3112 Lee Building
University of Maryland

E-mail: sswartz1@umd.edu
Phone(W): 301-405-8280
Fax: 301-314-9569

Topic Modeling For Humanities Research

Table of Contents

List of Participants

Bhattacharyya, Sayan	University of Michigan
Brown, Travis	Maryland Institute for Technology in the Humanities
Guiliano, Jennifer	Maryland Institute for Technology in the Humanities
Millon, Emma	Maryland Institute for Technology in the Humanities
Templeton, Clay	University of Maryland

Abstract

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. 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. 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. The final implementation session will focus on the development and explication of tools such as the Machine Learning for Language Toolkit (MALLET). 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.

Statement of Innovation

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. This workshop will address these issues by providing an opportunity for humanists and scholars working in natural language processing jointly to identify potential areas of research and development within applications, extensions, and implementation of topic modeling.

Statement of Humanities Significance

Our workshop will provide humanities scholars with a deeper understanding of the vocabulary of LDA topic modeling (and other latent variable modeling methods) and best practices for interpreting the output of such analysis, and will articulate fundamental literary and historical questions for researchers outside of the humanities who are developing the models and methods (as well as the software implementations).

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's approach to topic modeling utilizes aggregated data to explore macro level trends, themes, exchanges, and patterns in literature. Yet, humanities scholars often need to focus simultaneously on the macro (distant/many texts) and the micro (close/individual texts).

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. Latent Dirichlet Allocation (LDA) topic modeling provides a way to explore "distant" and "close" simultaneously. 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 sources. Casualty rate data for each week of the war 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.

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.

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.

The proposed 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 scholars using topic modeling approaches/software and those developing and utilizing them in computer science and natural language processing. Our primary goals for the workshop 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 focused on topic modeling have been dominated 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 this phenomenon 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 of 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." 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 uncritical consumers rather than as engaged critical applied theorists.

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 Senior Personnel Travis Brown, lead Research and Development Software Developer, 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. Brown will be aided by 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. 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: Ex. B6 . 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, Ex. B6 . 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 develop and oversee the intellectual agenda of the workshop; University of Michigan Graduate Student Sayan Bhattacharyya and UMD Graduate Student Clay Templeton who 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 manage the project and provide logistical support for all workshop related activities including handling all local arrangements and coordinate 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 Guiliano and 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 be further extended into humanities scholarship. These extrapolations will form a core set of recommended areas of further inquiry.

NEH: Topic Modeling in Humanities Research, A Level One Digital Humanities Start Up Proposal		Applicant Institution: University of Maryland Project Director: Jennifer Guiliano Project Grant Period: 05/01/2012-04/30/2013		
	Computational Details/Notes		Year 1	Project Total
1. Salaries & Wages				$5,375
Project Director Jennifer Guiliano	FY12 salary: Ex. B6 x 3% COLA x 2% effort		Ex. B6	Ex. B6
Senior Personnel Travis Brown	FY12 salary: x 3% COLA x 2% effort			
Community Lead/Public Relations	FY12 salary: x 3% COLA x 2% effort			
Web Designer Amanda Visconti	Hourly Rate: Ex. B6 x 3% COLA x 24 hours			
Graduate Student Clay Templeton	Hourly Rate: Ex. B6 no COLA) x 96 hours			
2. Fringe Benefits				$1,359
Project Director Jennifer Guiliano	30% of funded portion		Ex. B6	Ex. B6
Senior Personnel Travis Brown	30% of funded portion			
Community Lead/Public Relations	30% of funded portion			
Web Designer Amanda Visconti	30% of funded portion			
Graduate Student Clay Templeton	8% of funded portion			
3. Consultant Fees				Ex. B6
U Michigan Graduate Student Sayan Bhattacharyya	Hourly Rate: Ex. B6 per hour (96 hours total)		Ex. B6	Ex. B6
4. Travel				$10,026
Speakers: Domestic Travel to Workshop	9 flights total (9 people for 1 workshop event) at $375 per flight; 9 local transportation fees at $55 per individual		$3,870	$3,870
Speakers: Domestic Accommodations	18 nights (9 people, 2 days each) at $175 per night		$3,150	$3,150
Speakers: Domestic Per Diem (dinner only)	18 days (9 people, 2 days each) at $24 per day		432	432
Bhattacharyya: Domestic Travel, Accommodations, Per Diem	(1 flight at $350; 4 nights at $175; 3 days at $42 per day, 2 dinners at $24 per day)		$1,224	$1,224
Guiliano: Local Travel NEH Directors Mtg			$0	$0
Local hosting	Lunch ($15 per person for 50 people); Coffee/Refreshments ($12 per person for 50 people)		1350	$1,350
5. Supplies & Materials			0	$0
6. Services			0	$0
7. Other Costs			0	$0
8. Total Direct Costs	**Per Year**			$17,912
9. Total Indirect Costs	**Per Year**			$6,896
	Indirect Cost Calculation: a. Rate: 38.5% of direct cost per year b. Federal Agency: DHHS c. Date of Agreement: 06/07/2011			

10. Total Project Costs (Direct and Indirect costs)				$24,807
11 Project Funding				
a. Requested from NEH	Outright:			$24,807
	Matching Funds:			$0
	Total Requested from NEH:			$24,807
b. Cost Sharing	Applicant's Contributions:			$0
	Third Party Contributions:			$0
	Project Income:			$0
	Other Federal Agencies:			$0
	Total Cost Share:			$0
12. Total Project Funding				$24,807

UNIVERSITY OF MARYLAND, COLLEGE PARK
BUDGET JUSTIFICATION
Title: Topic Modeling in Humanities Research
Principal Investigator: Jennifer Guiliano
Period: 05/01/2012-04/30/2013

Project Dollars

1a. Senior Personnel Ex. B6

The Principal Investigator, Dr. Jennifer Guiliano, at 2% of her annual salary Ex. B6, will manage the project and provide logistical support for all workshop related activities. Senior Personnel Mr. Travis Brown, at 2% of his annual salary, will develop and oversee the intellectual agenda of the workshop. Senior Personnel Ms. Emma Millon, at 2% of her annual salary, will be responsible for all pre-workshop publicity and outreach in consultation with Dr Guiliano. Web designer Amanda Visconti, at Ex. B6 per hour, will provide 24 hours of graphic and publicity design. Salaries are based on UMD FY2012 rates and are incremented at a rate of 3.0% each year.

1b. Other Personnel Ex. B6

Graduate Student Clay Templeton, at Ex. B6 per hour for 96 hours, will work with Mr. Brown to develop workshop materials and manage outreach through twitter and other social media apparatus.

2. Fringe Benefits **$1,359**

Fringe benefits on personnel and primary investigator salaries are budgeted at an average rate of 30% and charged to the grant as actual costs.

3. Consultant Fees Ex. B6

Funding is requested for University of Michigan Graduate Student Sayan Bhattacharyya, at Ex. B6 per hour for 96 hours, to work with Mr. Brown and Mr. Templeton on pre- and post-workshop tasks including identifying materials, resources, and scholars working in topic modeling. Mr. Bhattacharyya will also be charged with providing documentation on any topic modeling software that will be used during the workshop.

4. Travel **$10,026**

Domestic travel support is requested for nine invited speakers to present at the workshop at a cost of $375 per flight and $55 each for ground transportation. Domestic accommodation support is requested for nine invited speakers to be housed for 2 days each at a cost of $175 per night. Domestic per diem is also requested for nine invited speakers at $24 per day for dinner only, since breakfast & lunch will be provided (see below).

Domestic travel, accommodation, and per diem is requested for Consultant Bhattacharyya as follows: 1 flight at $350, 4 nights accommodation at $175, 3 days of per diem at $42 per day, and 2 dinners at $24 each.

Funds are requested in the amount of $15 per person to provide 50 lunches for the workshop attendees, and $12 per person to provide breakfast & coffee breaks to 50 attendees.

5. Not applicable **$0**
6. Not applicable **$0**
7. Not applicable **$0**
8. Total Direct Costs **$17,912**
9. Total Facilities and Administrative Costs **$6,896**

Facilities and Administrative costs are assessed at a rate of 38.5% of Modified Total Direct Costs (MTDC).

10. Total Project Costs **$24,807**

Biographies

Sayan Bhattacharyya is a master's student in the University of Michigan's School of Information and holds a PhD in Comparative Literature and a master's degree in Computer Science (both earned at Michigan).

Travis Brown holds an M.A. in English from the University of Texas at Austin and is a Ph.D. student in English with a focus on natural language processing and automation for textual scholarship. While at the University of Texas he worked as an editor for the Walt Whitman Archive and was the lead developer of eComma, a web application for collaborative textual annotation. He also participated in a range of projects in UT's Computational Linguistics Lab, where he developed tools for dependency parsing, semantic role labeling, and toponym resolution.

Jennifer Guiliano is Assistant Director at MITH and a Center Affiliate of the National Center for Supercomputing Applications. Jennifer received a Masters of Arts in History from Miami University (2002), and a Masters of Arts (2004) in American History from the University of Illinois before completing her Ph.D. in History at the University of Illinois (2010). She has previously served as Associate Director of the Center for Digital Humanities at the University of South Carolina where she was also a Research Assistant Professor of History and as Post Doctoral Research Fellow and Program Manager at the Institute for Computing in the Humanities, Arts, and Social Sciences at the University of Illinois .

Emma Millon is Community Lead at MITH. She serves as the communications representative and outreach coordinator for MITH projects. Prior to arriving at MITH, Emma worked for three years as text-encoder and project manager for the Accademia di San Luca research database at the Center for Advanced Study in the Visual Arts, National Gallery of Art. She holds a Bachelor of Arts in English Literature and Italian from Harvard College.

Clay Templeton is currently a third year doctoral candidate in the University of Maryland's iSchool.

Data Management Plan

Project Title: Topic Modeling for the Humanities **Institution:** University of Maryland
Project Director: Jennifer Guiliano **Budget:** $24,807
Beginning: 05/01/2012 **Ending:** 04/30/2013 **Duration:** 12 months
This data management plan was created on September 12, 2011, for submission to the Office of Digital Humanities (ODH), National Endowment for the Humanities as required by ODH Guidelines in the interest of securing funding for this project. This is the first version of the data management plan associated with this data.

Types of Data: The data produced by this project will consist of presentations to be given by participants in the proposed workshop. These data will comprise slide presentations, audio recordings, video recordings, and text files documenting an interdisciplinary workshop on the application of topic modelling approaches to humanities data. Social media will be a component of participation in the workshops and postings on social networks such as Twitter will be collected. The researchers will also create a web site to publicize the event and provide access to materials (see “Access & Sharing” below).

Data Standards and Capture: Data for this project will be captured as part of the process of running the proposed workshop. Invited speakers will be asked to upload slides, documents and other supporting materials to computers provided by Maryland (as the host institution). Maryland will also perform audio and video recording of the workshop events. Posts from social media services will be collected using application programming interfaces (APIs) provided by these services and open-source tools which comprise PHP code and MySQL database software.For presentations given at the workshop, widely-adopted formats, such as Microsoft Powerpoint, will be used. Recent versions of these formats are at relatively open but, to the degree they are not, these formats are widely-supported across multiple technical environments and are likely to remain readable. Some conversion may be performed to normalize presentation files to PDF format as necessary for long-term accessibility. Audio and video of workshop events will be captured as uncompressed audio or video but will be converted and maintained in the form of widely-adopted open formats such as MP4 that are suitable for storing efficiently and sharing over the network. Social media content will be collected using published application programming interfaces maintained by these services. Plain-text, open-formats such as XML or JSON (JavaScript Object Notation) are the most common formats for such services.The project will rely on consistent naming conventions created in accordance with local policies at Maryland. Furthermore, as part of deposit with digital collections hosted by the University of Maryland Libraries, all items will be given unique identifiers according to the handle system, an open protocol published by the Internet Engineering Task Force. This system will support persistent citation of data created by this project.

Metadata: Descriptive metadata will be created as part of the process of depositing items with the Maryland Libraries. Metadata elements will be manually supplied using the Dublin Core Metadata scheme by project personnel collaborating with library staff to ensure consistent application of content conventions and controlled values. Embedded technical metadata will be maintained with all audio and video recordings. Dublin Core is the chosen metadata standard for this project because it can be created using existing workflows for depositing research products into Maryland's institutional repository and also because Dublin Core metadata can be easily reused as embedded data in web pages related to the workshop. This alignment with wider web practices will help enhance the discoverability of materials produced by this project.

Legal Policy: Participants in the workshop will retain their copyright and other intellectual property rights in material submitted for the workshop but will agree to license materials prepared for the conference under permissive open source licenses (such as Creative Commons) according to Maryland's standard policy or a similar arrangement agreed between participants and the workshop organizer. Contributions to the final white paper will be considered under a similar agreement. Participants will be asked to sign waivers giving their consent for audio and video of presentations to be shared. There are no other legal policy issues associated with data created or captured for this project.

Data Storage, Security, and Backup: Once collected from invited speakers or from digital audio and video recording devices, data will be physically stored on a password-protected server maintained by the Information Technology Division of the Maryland University Libraries. Servers are housed in a secure machine room with redundant power. No data will reside on any other portable or external media. Data is backed up incrementally through a service provided by the Maryland Office of Information Technology, which has a proven record of and commitment to secure data archiving for the university. In addition to backups hosted on university servers, data will be copied to tape and stored at a geographically-distant site through a service provided by Iron Mountain information services. The specific volume of storage for this project is not anticipated to exceed 1 TB.

Access, Sharing & Re-use: All data from this project will be made available for download from either the dedicated site hosted by MITH or from the Digital Collections of the University of Maryland within 6 months of the end of the workshop. There will be no additional permissions required to download or reuse data except for those specified above. Data from this interdisciplinary workshop will be useful to researchers in the fields of machine learning and related disciplines in information science and computer science. Also, data may be used by humanities researchers who are interested in applying computational methods to detect patterns in large collections of cultural heritage material.

Long-Term Preservation: Within three years from the end of the grant period, data will be permanently archived with the University Libraries at Maryland. No data will remain on servers controlled by the Maryland Institute for Technology in the Humanities (MITH). Data will remain publicly available through the libraries' digital collections. The University of Maryland has an interest in investing in the management of data created by researchers affiliated with Maryland and to providing digital preservation services, such as file validation, integrity checks, and, if needed, format conversion.

September 12, 2011

Dear Travis,

I am delighted to write this letter to offer my commitment should our workshop proposal on topic modeling be accepted by the National Endowment for the Humanities. I will eagerly fulfill all responsibilities as outlined in the proposal and will be an active participant in the workshop. I encourage the NEH to consider our proposal and am confident in its potential impact on the humanities.

Sincerely,

Sayan Bhattacharyya

Sayan Bhattacharyya

Graduate Student (M.S.I. candidate)
School of Information
University of Michigan, Ann Arbor

September 12, 2011

Dear Travis,

I am delighted to write this letter to offer my commitment should our workshop proposal on topic modeling be accepted by the National Endowment for the Humanities. I will eagerly fulfill all responsibilities as outlined in the proposal and will be an active participant in the workshop. I encourage the NEH to consider our proposal and am confident in its potential impact on the humanities.

Sincerely,

Sept 12 2011

Clay Templeton

September 13, 2011

Neil Fraistat
Professor of English & Director
Maryland Institute for Technology in the Humanities (MITH)
B0131 McKeldin Library
University of Maryland
College Park, MD 20742-7011

Dear Neil,

I am writing to express the commitment of the Maryland University Libraries to providing staff time, computing resources, and technical infrastructure to support data curation and preservation services for grant projects undertaken by the Maryland Institute for Technology in the Humanities (MITH).

The Libraries have a strategic interest in developing and maintaining services and infrastructure to support the sustainable management of research data created by scholars affiliated with MITH and the University of Maryland. Our current agenda includes provision of secure data storage and backups provided in concert with the Office of Information Technology, basic bit preservation and file format migration as well as development of service models to provide consultations with researchers related to data management planning and best practices.

Your activities in the area of digital research will complement the Libraries' commitments to the development of a robust research agenda in data management, digital preservation, and digital collections more broadly. Please do not hesitate to involve us in your internationally-recognized efforts in digital humanities.

Sincerely,

Babak Hamidzadeh
Associate Dean of Information Technology
University of Maryland Libraries

September 19, 2011

Dear Travis,

I am delighted to write this Letter of Commitment for the 2011 NEH Digital Humanities Start-Up Level I Grant for Topic Modeling for Humanities Research. 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.

As an Assistant Professor at the University of Maryland iSchool and University of Maryland Institute for Advanced Computer Studies, I support efforts to bring together humanists and specialists in topic modeling in a workshop where there is potential to compare approaches and to generate best practices for scholars in the field.

My own research on exploring social science corpora with topic models and developing cross-cultural models is very relevant to the proposed event. I would be willing to commit to aid this workshop by identifying and recruiting participants, suggesting potential areas of interest, reviewing the final white paper.

I am confident in you and the team of personnel you have put together for this grant and look forward to working with you. I urge the NEH to grant this application and would be delighted to discuss this further if needed.

Sincerely,

Jordan Boyd-Graber
Assistant Professor, iSchool and Institute for Advanced Computer Studies
Affiliate, Cloud Computing Center, Computer Science,
Computational Linguistics and Information Science, and Language Science

Hornbake 2118C
University of Maryland School of Information Studies
Hornbake Bldg, South
College Park, MD 20742
jbg@umiacs.umd.edu

Appendices

References:

D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research* 3 (2003): 993–1022.

Cameron Blevins, "Topic Modeling Martha Ballard's Diary," *Historying*, http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Sharon Block and David Newman, "Common-place: Tales from the Vault", http://common-place.org/vol-06/no-02/tales/.

Travis Robert Brown, "CorporaCamp:: About Woodchipper," *CorporaCamp*, http://mith.umd.edu/corporacamp/tool.php.

Jeff Drouin, "Ecclesiastical Proust Archive," archive, *Ecclesiastical Proust Archive*, http://www.proustarchive.org/.

Matthew L. Jockers, "Matthew L. Jockers," *Matthew L. Jockers*, http://www.stanford.edu/~mjockers/cgi-bin/drupal/.

Christopher Manning, "Tutorial 6: Natural Language Processing Tools for the Digital Humanities | Digital Humanities 2011: June 19 – 22", https://dh2011.stanford.edu/?page_id=525.

David Mimno, "David Mimno," Personal, *David Mimno*, n.d., http://www.cs.umass.edu/~mimno/.

Franco Moretti, *Graphs, maps, trees: abstract models for a literary history* (London; New York: Verso, 2005).

Nelson, Robert, "Mining the Dispatch," *Mining the Dispatch*, http://dsl.richmond.edu/dispatch/pages/intro.

Dragomir Radev, "Dragomir Radev | UMSI," *Dragomir Radev*, http://si.umich.edu/people/dragomir-radev.

Chung-chieh Shan, "Chung-chieh Shan," *Chung-chieh Shan*, http://www.cs.rutgers.edu/~ccshan/.

Andrew J. Torget, "Home page for Torget," Dr. Andrew J. Torget, http://www.hist.unt.edu/faculty/Torget/Torget.htm.

University of Massachussets Amherst, "MALLET homepage," *MAchine Learning for LanguagE Toolkit*, n.d., http://mallet.cs.umass.edu/.

Itemized Schedule:

8:45- 9:00 am	Registration and Refreshments
9:00 am-9:15 am	Welcome and Setting of Workshop Goals: Travis Brown
9:15-9:45 am	Methods/Applications Speaker #1
9:45-10:15 am	Methods/ Applications Speaker #2
10:15-10:45 am	Methods/Applications Discussion
10:45 am- 11:00 am	Break
11:00-11:30 am	Extensions of Modeling Speaker #1
11:30- noon	Extensions of Modeling Speaker #2
12:00- 12:30 pm	Extensions of Modeling Discussion
12:30-1:30 pm	Catered Lunch/Break
1:30- 2:00 pm	Implementation Speaker #1
2:00- 2:30 pm	Implementation Speaker#2
2:30- 3:00 pm	Implementation Discussion
3-3:15 pm	Break
3:15-4 pm	Topic Modeling in the Humanities Roundtable Discussion (all)