Investigating topic modeling techniques for library chat reference data

ABSTRACT

Analysis of chat transcripts can provide librarians with rich insights into improving the quality of library resources, services, accessibility, and spaces. In practice it is burdensome for librarians to go beyond simple quantitative analysis (e.g., chat duration, message count, word frequencies) with existing tools, given that chat transcripts consist of unstructured text data. Currently, a great deal of time and effort are required to obtain rich insights. The lack of chat analysis tools hinders librarians' reaction to patrons' wants and needs in a timely manner, in an age when our patrons' information needs have been changing.

With a one-year planning grant, Dr. HyunSeung Koh at the University of Northern Iowa (UNI) Rod Library and Dr. Mark Fienup of the UNI Computer Science Department will analyze the UNI Rod Library's chat reference transcripts using multiple topic modeling techniques. Topic modeling, which reveals hidden patterns and topical trends in text data, is well-suited to the analysis of large volumes of chat transcripts. Use of topic modeling could enable librarians to move beyond simple analysis to a deeper and more nuanced understanding of patron needs by uncovering these hidden patterns and trends. This project will identify the most appropriate topic modeling technique(s) in the context of chat reference data, make program source codes (Python scripts) ready for other librarian users' immediate use, and initiate developing a user-friendly chat analysis and assessment tool. A collaborator, John Russell, who is a faculty librarian at the Pennsylvania State University with expertise in digital humanities and their related text analysis technologies/techniques, will serve as a consultant for this project. He will identify Advisory Board members, coordinate feedback and advice from them, and provide directional feedback on topic modeling techniques/approaches and their practicality in library settings.

Outcomes and findings of this project, which will be disseminated as open source codes and via diverse publication venues, will be beneficial immediately or in the longer term for a wide range of library staff including: (1) assessment librarians with technical expertise in running Python scripts, who will be able to analyze vast amounts of chat data using the Python scripts ready for use from this project; (2) chat-reference librarians who will be able to analyze chat data to improve the quality of chat reference services and monitor student employees' performances; and (3) decision makers, from administrative library staff to individual librarians in each unit, who need to make data-driven decisions that improve library resources, services, and spaces. In addition, outcomes and findings of this project will be used in our immediate follow-up project for developing a user-friendly analysis and assessment tool for chat reference data, with the goal of aiding librarians in navigating and analyzing their own vast amounts of chat transcripts in efficient and timely ways without needing a background in programming, as well as other potential follow-up comparative chat analysis studies across similar or different types of academic institutions (e.g., teaching vs. research institutions) using the Python scripts ready for use from this project.

NARRATIVE

Dr. HyunSeung Koh at the University of Northern Iowa (UNI) Rod Library and Dr. Mark Fienup of the UNI Computer Science Department request $100,000 for a one-year planning grant (July 1, 2019 - June 30, 2020) to explore UNI Rod Library's chat reference transcripts using multiple topic modeling techniques, identify the most appropriate topic modeling technique(s) in the context of chat reference data, make program scripts ready for other librarian users' immediate use, and initiate developing a prototype chat-analysis and assessment tool. Outcomes and findings of this project, which will be disseminated as open source codes and via diverse publication venues, will be used in our immediate follow-up project for developing a user-friendly analysis and assessment tool for chat reference data and potential follow-up comparative chat analysis studies across similar or different types of academic institutions (e.g., teaching vs. research institutions), based on the availability of additional funding.

Statement of National Need

With the rise of online education (Frieman, 2018), library chat services are an increasingly important tool for student learning (Desai & Graves, 2008; Oakleaf & VanScoy, 2011; Schiller, 2016). Library chat services have the potential to support student learning, especially for distant learners who will have a lack of opportunity to come and learn about library and research skills in person. In addition, unlike traditional in-person reference services that have declined drastically, library chat services have become an important communication channel that connects patrons to library resources, services, and spaces (Côté, Kochkina, & Mawhinney, 2016; LeMire, Rutledge, & Brunvand, 2016; Scales, Turner-Rahman, & Hao, 2015).

Analysis and findings of chat transcripts could provide librarians with rich insights into improving the quality of these resources, services, and spaces. Furthermore, these rich insights can be used in demonstrating and improving student success, which is increasingly in demand for higher education (Howard, 2019). In practice it is burdensome for librarians to go beyond simple quantitative analysis (e.g., chat duration, message count, word frequencies) with existing tools, given that chat transcripts consist of unstructured text data. Currently, a great deal of time and effort are required to obtain rich insights. The lack of chat analysis tools hinders librarians' reaction to patrons' wants and needs in a timely manner, in an age when our patrons' information needs have been changing. In particular, small and medium size academic libraries have seen a shortage of librarians and need to hire and train student employees, so librarians' capabilities in real-time analysis and assessment will become critical in helping them take appropriate actions to best meet user needs.

The ultimate aim of this project is to develop a user-friendly tool that aids librarians in navigating and analyzing their own vast amounts of chat transcripts in efficient and timely ways without needing a background in programming. This planning stage will identify appropriate topic modeling technique(s) in the context of chat reference data, make program scripts (Python scripts) ready for other librarian users' immediate use, and initiate developing a prototype of user-friendly chat-analysis and assessment tool.

Environmental Scan - Gaps and Opportunities

Our project will add new knowledge to existing bodies of knowledge by utilizing opportunities and filling in gaps that are identified from previous or existing chat transcript analysis methods/techniques in library and non-library settings, topic modeling techniques and tools, and library chat software (platforms) and their analytics features as below:

1) **Chat transcript analysis methods/techniques in library settings**
In order to identify gaps and opportunities in terms of research methods in analyzing chat reference data, we examined types of research methods in analyzing library chat transcripts, which are one major data source of library chat service research (Matteson, Salamon, & Brewster, 2011). One type of research method in analyzing chat transcripts is coding-based qualitative content analysis with or without predefined categories (e.g., Fuller & Dryden, 2015; Passonneau & Coffey, 2011). Another type of qualitative research method is conversation or language usage analysis (e.g., Dempsey, 2016; Waugh, 2013) but it is not a dominant type of research method, as compared to coding-based qualitative content analysis. The other type of research method is quantitative methods, most of which are simple descriptive count- or frequency-based analyses that are accompanied by qualitative coding-based content analyses (e.g., Brown, 2017; Maximiek, Rushton, & Brown, 2010). In some recent research, advanced quantitative research methods, such as cluster analysis and topic modeling techniques, have been used (e.g., Kohler, 2017; Stieve & Wallace, 2018; Tempelman-Kluit & Pearce, 2014), but they have not been fully explored yet. Our project will be built upon this line of recent research by exploring not only individual topic modeling techniques but also combinations of existing topic modeling techniques.

2) **Chat transcript analysis methods/tools/techniques in non-library settings**
We have also explored commonly-used methods/tools/techniques of chat transcript analysis in non-library settings, in order to learn from other disciplines and identify gaps that can be filled by our project, as below:

Disciplines	Platforms/Sources of Chat Transcript Data	Chat Transcript Analysis Methods/Tools/Techniques
Education	Chat rooms, text chat (e.g., Kassner & Cassada, 2017; Golonka, Tare, & Bonilla, 2017)	Qualitative content analysis
Health	Social media (e.g., Hamad et al., 2016; Richardson, 2016)	Qualitative & quantitative content analysis
Business	In-game chat features, chatbots (e.g., Ho et al., 2016; Park, Aiken, & Salvador, 2018)	A spell-checker, readability scores, the number of spelling and grammatical errors, Linguistic Inquiry and Word Count (LIWC) program, logistic regression analysis, Decision Tree, Support Vector Machine (SVM)
Criminology	Instant messengers, Internet Relay Chat (IRC) channels, internet-based chat logs, social media (e.g, Basher & Fung, 2014; Drouin et al., 2017; Kuang, Brantingham, & Bertozzi, 2017); Miah, Yearwood, & Kulkarni, 2015)	LIWC program, cluster analysis, Latent Dirichlet Allocation (LDA)

We found that a wide range of disciplines have used various research/analysis methods, ranging from qualitative manual coding methods to data mining and machine learning techniques, in analyzing chat data from diverse technology platforms. Topic modeling techniques are one of the chat analysis methods but again it seems that they have not been fully explored yet in chat analysis, even though they have been used in a wide range of contexts (Boyd-Graber, Hu, & Mimno, 2017). Our project will advance knowledge of topic modeling techniques in non-library settings, as well as library settings, by enabling its findings to be used in similar projects with similar contexts in other disciplines, such as information-seeking and educational chat contexts.

3) **Topic modeling techniques and their strengths and weaknesses**

Given the potential for topic modeling, as an appropriate technique for revealing hidden patterns and topical trends in large volumes of chat data, we examined strengths and weaknesses of the most common topic modeling techniques - Vector Space Model, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, and Latent Dirichlet Allocation - which we will use for this project (Aggarwal & Zhai, 2012; Alghamdi & Alfalqi, 2015; Anaya, 2011; Barde & Bainwad, 2017; Blei, 2012; Chen, Thomas, & Hassan, 2016; Günther & Quandt, 2016; Ignatow & Mihalcea, 2017; Jansen, 2018; Liu, et al. 2016; Mohr & Bogdanov, 2013; Salton, 1975; Xu et al., 2015; Zhai, 2008; Zhang & Ma, 2015).

	Acronym	Definitions	Strengths	Weaknesses
Vector Space Model	VSM	A document is represented as a vector of words in the vector space.	Easy to incorporate different models.	Cannot deal with polysemy (that is, a word with multiple meanings, such as lead, which can be a noun or a verb).
Latent Semantic Analysis	LSA	A document is represented as a vector of words in the reduced dimensional vector.	Can deal with polysemy to some extent.	Hard to obtain and to determine the number of topics.
Probabilistic Latent Semantic Analysis	pLSA	It is similar to LSA, but topics in a document are "probabilistic instead of the heuristic geometric distances." (Zhang & Ma, 2015, p74)	Can deal with polysemy issues.	Over-fitting problems.
Latent Dirichlet Allocation	LDA	It is an extension of pLSA and a document consists of different or distribution of topics.	Prevents over-fitting problems.	Need to manually remove stopwords. Does not show relationships among topics.

Individual techniques of topic modeling, "a type of statistical model for finding hidden topical patterns of words" (Barde & Bainwad, 2017, p. 745), have their own strengths and weaknesses. In order to examine which technique or combinations of techniques, from relatively old to recent, are most beneficial for analyzing library chat data, this project will explore these four techniques individually and combinations of these four techniques.

4) **Tools in topic modeling and their usage contexts**
In order to build upon existing topic modeling tools ready for use and to identify contributions that we can make from this project, we investigated four representative topic modeling tools and their common contexts or sources of chat data - MALLET, Gensim, BigARTM, and Stanford Topic Modeling Toolbox (Barde & Bainwad, 2017; Liu, et al. 2016) - as below:

Tools	Websites	Objectives	Contexts (Sources) of Chat Transcript Data
MALLET	http://mallet.cs.umass.edu/	Document classification, clustering, topic modeling, information extraction	Historical newspapers (e.g., Yang, Torget, & Mihalcea, 2011); Students' essays (e.g., Resnik, Garron, & Resnik, 2013); Electronic petitions (e.g., Hagen, 2018)
Gensim	https://radimrehurek.com/gensim/	Extraction of semantic structure	Social media (e.g., Guo et al., 2016);
BigARTM	http://bigartm.org/	Sparsing, smoothing, topics decorrelation	Social media (e.g., Farseev, 2016);
Stanford Topic Modeling Toolbox (*No longer updated or maintained*)	https://nlp.stanford.edu/software/tmt/tmt-0.4/	Analysis of large amount of textual data for social science researchers	Electronic health records (e.g., Kayi et al, 2017); Social media (e.g., Manna & Phongpanangam, 2018)

As summarized above, common topic modeling tools have been used in non-library contexts such as newspapers, social media, health records, and student essays. Utilizing additional tools/packages/modules/libraries (e.g., NumPy, SciPy, Gensim, NLTK) for Python, which is the programming language to be used for this project, our project will contribute to determining the strengths and limitations of the most common topic modeling techniques in a library context. In addition, we found that these command-line tools require programming expertise to some extent. Our quick and easy library-user-friendly tool, which we will initiate developing toward the end of this project, will be beneficial for all library staff and student employees, regardless of technical expertise.

5) **Library chat software (platforms) and their analytics features**
Last, in order to identify constraints of existing library chat platforms in terms of analytics, we examined features and capabilities of chat transcript and activity analysis in four widely-used chat software tools, including a chat platform in the PIs' institution, - LibraryH3lp, QuestPoint, LibChat, and LivePerson (Côté, Kochkina, & Mawhinney, 2016; Yang, & Dalal, 2015). Our findings are summarized below:

	Location of Documentation Consulted	File Type of Transcripts Downloaded	Features that Help Analyze Chat Transcript Contents	Filters that Help Quantify Chat Activities
LibraryH3lp	https://docs.libraryh3lp.com/	.txt	Keyword searches	e.g., Start/End Date, Include Not Answered, Minimum Chat Duration
Question Point	https://help.oclc.org/Discovery_and_Reference/QuestionPoint	.xml	Keyword searches	e.g., Time and Date, Session Time, Wait times
LibChat (*A platform in the PIs' institution*)	https://ask.springshare.com/libanswers/faq/1764	.csv	Keyword searches	e.g., Timestamp, Wait Time, Duration
LivePerson	https://www.liveperson.com/services/technical-support/engagement-history	.csv	Keyword/Phrases searches	e.g., Start Time, Engagement Length

As we demonstrated above, content-level in-depth analysis tools/features in chat software are limited to simple manual keyword searches. These findings show gaps to be filled by quick and easy content-level in-depth analysis features/tools. In addition, we found that something that would help clean text data, regardless of file type, and make them ready for immediate analysis will be beneficial. In short, outcomes of our project will contribute to developing a tool that helps quick and easy transcript content analysis, regardless of these three representative types of files.

To summarize, our project will contribute to manifesting the strengths and weaknesses of each topic modeling technique/approach and their combinations and will identify the most appropriate topic modeling techniques/approaches in the context of library chat reference data. This will help develop a quick and easy in-depth chat transcript analysis tool, including our immediately available Python scripts and prospective user-friendly tool, that will be initiated toward the end of this project, and will further advance knowledge of topic modeling techniques.

Project Design

Our project will emphasize both processes and outcomes equally. In other words, we will complete our project through an iterative process of reflecting upon outcomes from each step, adjusting or adapting in a subsequent step accordingly, and unfolding and reporting outcomes from each step.

1) **Defining success**

A key success measure of our project is to manifest not only results of techniques/approaches in terms of the accuracy of topic extraction but also processes of our exploration in terms of how to incorporate one technique/approach to another in achieving our goal of accuracy. In other words, we will report successful and unsuccessful techniques/approaches and solutions to challenges and attempts to address unsolved challenges throughout these processes. For example, we found from our preliminary exploratory study that tuning of allowable topic words versus stop words was an important step in improving the quality of chat-topics identified (see the Implementation section below for more details about strategies employed in improving the accuracy of topic extraction in our preliminary study). This will help other researchers expand our approaches and fill in gaps identified from our successes and challenges.

2) **Data management (confidentiality and privacy)**

In order to protect confidentiality and privacy of library users, the PIs will immediately delete voluntarily entered identifiable information (name and contact information) as soon as chat data are downloaded from a LibChat module of the library platform SpringShare. Chat data downloaded without identifiable information, Python scripts (.py), and documentations (.txt, .rtf, or .docx) will be saved in a shared project folder in the PIs' institutionally-secured Google Drive, which is accessible only to members designated by the PIs including a consultant and advisory board members, and on institutional computers, which will be accessible only to the PIs and a student assistant. Also, the PIs will remove any additional unexpected identifiable information encountered during analysis. At the stage of publication, results and findings of this analysis will be addressed in an aggregated manner. (For more details, please see the IRB document attached in Appendix A.)

3) **Collaborations**

Dr. HyunSeung Koh, Assessment Librarian & Assistant Professor of Library Services at University of Northern Iowa, has expertise in library assessment, reading and technology, Human-Computer Interaction, and research methods, and experience in programming and text analysis techniques. As PI, she will manage and oversee the performance and timeliness of all key tasks throughout the project. She will provide feedback to Dr. Fienup on processes and outcomes of the new approaches/techniques and will consult with other library staff in charge of chat reference services as needed. She will mediate all project team communications including the consultant and advisory board members, and will be responsible for disseminating results through a conference paper, final report, and presentation.

Dr. Mark Fienup is an Associate Professor in Department of Computer Science at University of Northern Iowa and received his master's and Ph.D. degrees in Computer Science from Iowa State University with relevant coursework in Artificial Intelligence. As co-PI, he will write Python scripts as scheduled, securely save all versions of scripts, document all details about successes, challenges, and solutions, communicate with a consultant and advisory board members about technical aspects of the project, and recruit and oversee a student assistant.

John Russell is Digital Humanities Librarian and Associate Director of the Center for Humanities and Information at the Pennsylvania State University. He was formerly part of the Digital Scholarship Center at the University of Oregon, where he taught digital scholarship methods, supported digital scholarship projects (including serving as a text encoding consultant on an ACLS-funded project), and worked with librarians to incorporate digital tools and methods into their practice of librarianship. He regularly teaches "An Introduction to Digital Humanities for Librarians" and other related courses for Library Juice Academy.

He is also currently a curator for humanities data as part of the Data Curation Network, a Sloan Foundation-funded project. As a consultant, he will help the PIs identify and recruit Advisory Board members, help manage the activities of the Advisory Board, and offer technical and procedural guidance during the grant. Contingent upon additional funding, he will be the PIs' collaborator on a future follow-up study that compares chat data between two different types of academic institutions using the same chat software, which will help us re-test and verify strengths and weakness of our topic modeling techniques.

4) **Recruitment**

We are planning to recruit two Advisory Board members who have expertise in topic modeling and two members who have expertise in library assessment. Two members with expertise in topic modeling will provide feedback on coding in terms of approaches and results and two members in library assessment will comment on the usefulness of approaches and results in practice. A consultant will find potential candidates by identifying scholars and/or librarians who have done research in areas relevant to this project via diverse channels (e.g., journal publications, conference presentations, social media, and professional networks) and contact and recruit them directly via email.

We will recruit an undergraduate student, who has an expertise in Python programming, as a student assistant in order to assist Dr. Fienup's coding. Dr. Fienup will contact a professor who teaches AI-related courses, get recommendations from him on students who are excellent in Python programming, send emails to potential candidates, conduct one-on-one interviews with them, and finalize and recruit a student.

5) **Implementation**

This Planning Grant project will expand and build upon a preliminary exploratory study in which we wrote multiple versions of scripts in the Python programming language using the IDLE editor. These scripts were used to extract topics from about 7,000 chat transactions occurring from April 10, 2015 to March 31, 2018 and downloaded from the PIs' institutional chat platform, called SpringShare, with a LibChat module. For topic extraction we utilized Latent Dirichlet Allocation (LDA), which is one commonly used technique, utilizing Python modules of numpy, scipy, gensim, and nltk.

The followings are examples of strategies that we employed at this preliminary stage:

- Eliminate non-Unicode characters (using Google Sheets Add-on called Power Tools)
- Word concordance with all chat texts (to identify word frequency & line #)
- Delete some basic stop words (e.g., why, who, what, when, how)
- Install NumPy, SciPy, Gensim, NLTK
- Import nltk and download stopwords and wordnet
- Eliminate first sentence (greeting) and last sentence (thank you) from each chat transcript
- Run a script with a first question (excluding a greeting sentence) from each chat transaction
- Add "contextualized" words (e.g., Rod Library) to stop words by increasing a threshold of word frequency to 1000
- Run a script that maps found chat topics back to a set of chat transcripts by line # in the .csv file

In short, we found that tuning of allowable topic words versus stop words was an important step in improving the quality of chat-topics identified. Also, we found that some topics are more accurate than others in representing topics of each chat. A topic of "Interlibrary Loan" is one example that stands out very accurately and is easily identifiable on which chats are associated with this topic.

In order to continue to find the most appropriate technique(s) for identifying accurate topics in the context of the PIs' library reference data, we will analyze chat data, collected from oldest to most recent chat data from April 10, 2015 to May 31, 2019, using the aforementioned four common topic modeling techniques - Vector Space Model (VSM), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) - and incorporation of those multiple techniques.

See our detailed schedules below:

- Planning & Initiating (Jul-19)
  - Identify Advisory Board members
  - Plan and schedule with Advisory Board members
  - Create a shared folder for schedules, progress, and feedback
  - Recruit a student assistant for coding
  - Collect and clean chat data and start coding

- Coding (Aug-19 - Nov-19)
  - Explore chat transcripts using individual techniques of VSM, LSA, pLSA, & LDA
  - Get feedback from consultant at the end of each month

- Mid-year progress reviews & coding (Dec-19)
  - Mid-year reviews on coding with all members including Advisory Board members
  - Explore chat transcripts using combinations of VSM, LSA, pLSA & LDA
  - Integrate data from all four common Library chat programs

- Coding (Jan-20 - Mar-20)
  - Explore chat transcripts using combinations of VSM, LSA, pLSA & LDA
  - Get feedback from consultant
  - Write and submit an abstract for International Conference on Qualitative and Quantitative Methods in Libraries (QQML)

- Final progress reviews & coding (Apr-20)
  - Final reviews on coding with all members including Advisory Board members
  - Identify, modify, and finalize most appropriate technique(s)
  - Get feedback from consultant
  - Write and submit a paper at QQML

- Wrapping up & final report (May-20 - Jun-20)
  - Prepare and present a paper at QQML
  - Start writing and finish the final report
  - Disseminate source codes via GitHub including usage documentation
  - Initiate developing a first prototype of a user-friendly chat analysis tool

6) **Dissemination**
Our partial preliminary findings associated with LDA were presented as a poster at the Library Assessment Conference in 2018, and received one of 9 Poster Awards (out of 82 Posters). We plan to present our findings related to this grant via other publication venues, such as International Conference on Performance Measurement in Libraries, Evidence Based Library and Information Practice Conference, Association for the Assessment of Learning in Higher Education Annual Conference, and QQML, whichever fits our schedule of completion best. Currently, we anticipate that QQML will be the most appropriate conference venue.

7) **Sustainability**
All of the code produced, including versions of Python scripts (.py), metadata (.txt) (e.g., library contextualized stop-words and non-stop words), and detailed instructions (.txt, .rtf or .docx) on how to install and utilize these scripts, will be shared via GitHub. Additionally, in the future with additional funding we will consider doing follow-up comparative analysis studies across similar or different types of academic institutions (e.g., teaching vs. research institutions) that might have similar or different types of patrons in terms of information needs, which will help make improvements to our tool or identify a most appropriate usage context for our tool.

8) **Diversity Plan**
Chat features in library websites are open to everybody who has internet access from anywhere in the world, not only to populations of home institutions but also to members of local, national, and international communities. As Powers and Costello (2019) argued that librarians can make virtual reference services more inclusive using diverse strategies such as outreach to multicultural groups on campus, we believe our current and potential patrons in our own institution and other potential academic and public libraries can be better served by librarians who are equipped with an easy-to-use chat transcript analysis tool.

National Impact

1) **Immediate impact in terms of library services with Python scripts as a command-line based tool**
Our Python scripts, which will be ready for immediate use from this project, will be beneficial for library staff who need to analyze vast amounts of chat data in a quick and easy way. For example, assessment and chat reference librarians in the PIs' institution will be able to analyze chat transcripts, from first implementation of chat reference services until now, identify how information needs of faculty and students in their institution have changed over time, and propose appropriate actions accordingly. Also, librarians in other academic institutions will be able to analyze their own chat transcripts using the Python scripts, which will be immediately released via GitHub, with a minimal amount of training about how to run Python scripts.

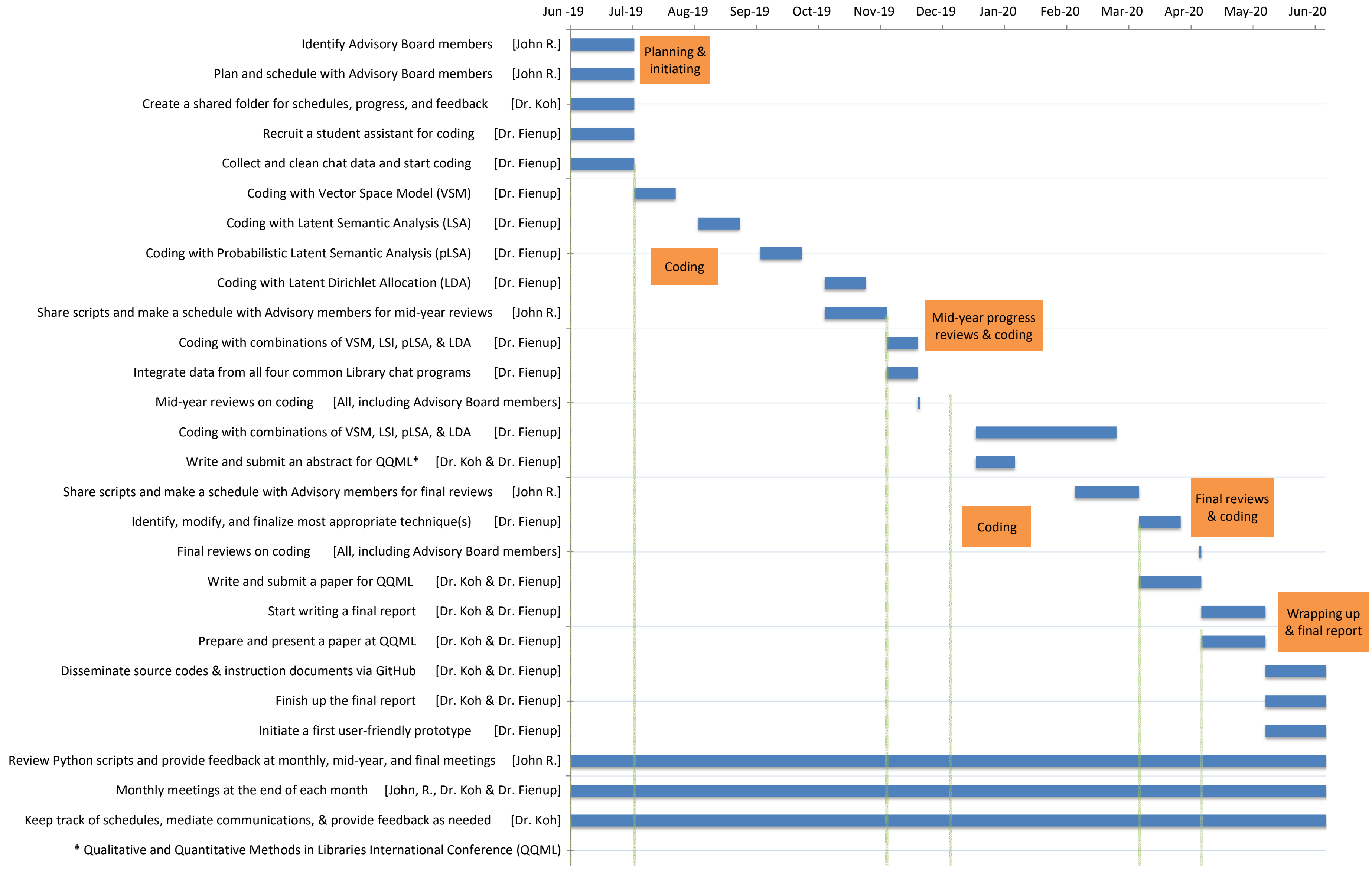
2) **Immediate, short- or long-term impact in terms of topic modeling techniques**
While we continue to improve our analysis technique by improving its accuracy of topic extraction, we are expecting to add new context-specific (i.e., library chat data) knowledge to existing bodies of knowledge about topic modeling techniques by sharing our source codes and process with online communities, such as GitHub and Stack Overflow, and our experiences and findings via library-specific and other disciplinary

emailing lists, conferences, and journals, such as the Library and Information Technology Association (LITA) emailing lists, International Conference on World Wide Web, International Conference on Learning Analytics & Knowledge, Code4Lib Journal, Communications of the Association for Computing Machinery (ACM), and Journal of Machine Learning Research. In particular, we are expecting that findings from our project can be utilized for other similar “educational” contexts in other disciplines by sharing our metadata (e.g., library contextualized stop-words and non-stop words) in our Python scripts.

3) **Short- or long-term impact in terms of library services with a user-friendly tool**
Dissemination and usage of our user-friendly tool, which will be initiated toward the end of this project, by other libraries will allow librarians without technical expertise to easily identify a landscape of topics in their chat data, map each topic back to a small set of chat data for deeper qualitative analysis, and take appropriate actions in a timely manner. For example, it will help chat reference librarians, who need to improve the quality of chat services and monitor student employees’ performance, to identify recurring issues quickly, seek solutions for them, and minimize future redundant efforts to address the same issues. Any decision makers, from administrative library staff to individual librarians in each unit, will be benefited by this tool in making data-driven decisions in finding ways to improve library resources, services, and spaces.

4) **Long-term impact in terms of library services with a user-friendly tool with visualization features**
Ultimate improvements to usability, which will be implemented as a follow-up long-term project in the future contingent upon additional funding, include application of visualization techniques as well as the ability to accept diverse types of chat data from other commonly used chat platforms. We believe that librarians at a broad range of institutions will take advantage of such a user-friendly high-level chat-analysis tool. In particular, given that chat features are accessible to anybody from anywhere and have high potential for attracting and reaching out a wide range of patrons, we hope that this tool enables librarians to identify “hidden” needs of local, national, and international community members and take appropriate action rapidly without the need for additional specialized personnel or resources, which will eventually help improve the well-being of these communities.

Investigating topic modeling techniques for library chat reference data: LG-34-19-0074



Identify Advisory Board members [John R.]

Plan and schedule with Advisory Board members [John R.]

Create a shared folder for schedules, progress, and feedback [Dr. Koh]

Recruit a student assistant for coding [Dr. Fienup]

Collect and clean chat data and start coding [Dr. Fienup]

Coding with Vector Space Model (VSM) [Dr. Fienup]

Coding with Latent Semantic Analysis (LSA) [Dr. Fienup]

Coding with Probabilistic Latent Semantic Analysis (pLSA) [Dr. Fienup]

Coding with Latent Dirichlet Allocation (LDA) [Dr. Fienup]

Share scripts and make a schedule with Advisory members for mid-year reviews [John R.]

Coding with combinations of VSM, LSI, pLSA, & LDA [Dr. Fienup]

Integrate data from all four common Library chat programs [Dr. Fienup]

Mid-year reviews on coding [All, including Advisory Board members]

Coding with combinations of VSM, LSI, pLSA, & LDA [Dr. Fienup]

Write and submit an abstract for QQML* [Dr. Koh & Dr. Fienup]

Share scripts and make a schedule with Advisory members for final reviews [John R.]

Identify, modify, and finalize most appropriate technique(s) [Dr. Fienup]

Final reviews on coding [All, including Advisory Board members]

Write and submit a paper for QQML [Dr. Koh & Dr. Fienup]

Start writing a final report [Dr. Koh & Dr. Fienup]

Prepare and present a paper at QQML [Dr. Koh & Dr. Fienup]

Disseminate source codes & instruction documents via GitHub [Dr. Koh & Dr. Fienup]

Finish up the final report [Dr. Koh & Dr. Fienup]

Initiate a first user-friendly prototype [Dr. Fienup]

Review Python scripts and provide feedback at monthly, mid-year, and final meetings [John R.]

Monthly meetings at the end of each month [John, R., Dr. Koh & Dr. Fienup]

Keep track of schedules, mediate communications, & provide feedback as needed [Dr. Koh]

Planning & initiating

Coding

Mid-year progress reviews & coding

Coding

Final reviews & coding

Wrapping up & final report

* Qualitative and Quantitative Methods in Libraries International Conference (QQML)

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

All applications must include a Digital Product Form.

□ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The digital products (i.e., program scripts) will be subject to the GNU General Public License. The GNU General Public License was chosen because it is a copyleft license so derivative works can only be distributed under the same license.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

Our organization does not reserve any ownership rights to the new digital products. The products will be made widely available on GitHub as open-source codes subject to the GNU GPL.

A. 3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

The Python program scripts being produced will analyze chat data of the user, but nothing we produce will involve products that involve privacy concerns.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

We will write multiple iterative versions of scripts using the Python program language with the source code (.py files) being text files, which will be used for initiating developing a chat-analysis tool to be used by libraries. Using these Python scripts, we will analyze approximately 7,000 or more chat transcripts stored in a Microsoft Excel spreadsheet, which will be collected from a chat platform at the University of Northern Iowa Rod Library. Dates of chat transactions will range from April 10, 2015 to May 31, 2019. These library chat data are for internal usage with the University IRB's approval.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

We will use University of Northern Iowa computers, servers, and free development tools (mainly Python 3.7).

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

The digital file formats will be text files of some format: source Python programs (.py) and comma-separated values (.csv) files.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

We will employ the standard Waterfall model of software development, which means that we will move forward step by step incrementally toward improvements in terms of accuracy of topic extraction. Deliverables (i.e., program scripts) in each milestone will be saved as different versions. Since we are evaluating several topic-modeling methods, their results will be compared for relative accuracy.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

During the award period's development phase, a shared project folder in the PIs' institutionally-secured Google Drive will be utilized to store digital assets. This shared project folder is accessible only to members designated by the PIs including a consultant and advisory board members. Towards the end of the grant, Python scripts (.py), metadata files (.txt) (e.g., a list of library contextualize stop-words and non-stop words), and detailed instructions for usage will be shared via GitHub.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

The metadata produced in our exploration of topic modeling techniques (on chat-transcripts) - Vector Space Model, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, and Latent Dirichlet Allocation - will need to associate words or sets of words to real numeric values. Thus, text comma-separated value (.csv) files are all that are required.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

During the award period's development phase, a shared project folder in the PIs' institutionally-secured Google Drive will be utilized to preserve and maintain metadata. This shared project folder is accessible only to members designated by the PIs including a consultant and advisory board members. Towards the end of the grant, Python scripts, metadata files, and detailed instructions for usage will be shared via GitHub.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

We will share our Python scripts, including metadata, and process with online communities, such as GitHub, and our experiences and findings via diverse publication venues, such as conferences, journals, and online listservs, whichever fits our schedule of completion. Examples include International Conference on Performance Measurement in Libraries, International Conference on Qualitative and Quantitative Methods in Libraries, Evidence Based Library and Information Practice Conference, Code4Lib Journal, Communications of the ACM, and the Library and Information Technology Association (LITA) email lists.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The Python scripts, metadata files, and detailed instructions for usage will be shared via GitHub.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

The University of Northern Iowa has many examples of publicly available digital content, some of which are computer science related projects listed below:
UNI-FOLD: http://bioinfo.cs.uni.edu/ LOSST (Laboratory of Security & Storage Technology): http://www.cs.uni.edu/~diesburg/losst/index.html UNI LLAMA (Learning Lab for Applied Manipulative Applications): http://www.cs.uni.edu/~diesburg/llama/index.html Reai-Time Embedded Systems Laboratory: http://www.cs.uni.edu/~mccormic/RealTime/

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Not applicable for this project

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

Not applicable for this project

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

Not applicable for this project

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

Not applicable for this project

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

Not applicable for this project

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

Not applicable for this project

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

Not applicable for this project

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

Not applicable for this project

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

Not applicable for this project

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

Not applicable for this project

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Not applicable for this project

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

Not applicable for this project

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

Not applicable for this project

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Not applicable for this project

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Not applicable for this project

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

Not applicable for this project

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Not applicable for this project

A.8 Identify where you will deposit the dataset(s):

Name of repository:

Not applicable for this project

URL:

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

Not applicable for this project