Investigating topic modeling techniques for library chat reference data

Dr. HyunSeung Koh & Dr. Mark Fienup - IMLS, NLG-Libraries-FY19, Planning Grants, National Digital Infrastructures and Initiatives, Build Capacity

Summary:
Dr. HyunSeung Koh at the University of Northern Iowa (UNI) Rod Library and Dr. Mark Fienup of the UNI Computer Science Department request $99,819 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 in the context of chat reference data, and develop 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.

Statement of National Need:
In recent years, chat features on academic libraries' websites have become an important communication channel that connects patrons to library resources, services, and spaces. Analysis and findings of chat transcripts could provide librarians with rich insights into improving the quality of these resources, services, 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 are unstructured text data that require time and efforts in obtaining rich insights. This lack of chat analysis tools hinders librarians' reaction to patrons' wants and needs in a timely manner in the age where our patrons' information needs have been changing unexpectedly in some cases. 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 right on time. The ultimate aim of this project is to develop a 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 the most appropriate topic modeling technique for developing such a tool and initiate a first prototype of the tool.

Project Design:
In our preliminary research, we initially wrote multiple versions of scripts in the Python programming language that extracted topics from about 7000 chat transactions collected from April 10, 2015 to March 31, 2018 using our LibChat module in LibAnswers from Springshare. For topic extraction we utilized Latent Dirichlet Allocation (LDA), one natural language processing technique for topic extraction utilizing Python modules of numpy, scipy, gensim, and nltk. 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 for identifying most accurate topics in the context of our library reference data, we plan to analyze our chat data using other topic modeling techniques, such as Vector Space Model (VSM), Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and incorporation of those multiple techniques. We will compare results across different techniques, report their differences in terms of outcomes, and initiate a first prototype of the tool. Our partial preliminary findings associated with LDA will be presented as a poster accepted to the Library Assessment Conference in 2018. We also plan to present our findings via other publication venues such as International Conference on Performance Measurement in Libraries, International Conference on Qualitative and Quantitative Methods in Libraries, and the Association for the Assessment of Learning in Higher Education (AALHE) Annual Conference.

National Impact:

1) **Immediate impact in terms of library services**: Our findings and outcomes will be immediately beneficial for librarians and student employees in our institution. Our tool will allow librarians 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. We expect that other institutions which have similar challenges will take advantage of the outcomes of our project.

2) **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. We plan to share our source codes with online and offline communities, such as SourceForge, and our experiences and findings via diverse publication venues.

3) **Long-term impact in terms of usability and library services**: Ultimate improvements to usability include application of visualization techniques for chat-analysis results, and the ability to accept chat-data from other commonly used chat programs (e.g., LibraryH3Ip). We believe that librarians at a broad range of institutions will take advantage of a user-friendly high-level chat-analysis tool developed by this study. Chat features have high potential for attracting a wide range of patrons. Virtually, chat features are accessible to anybody from anywhere. In particular, we hope that this tool enables librarians to uncover "hidden" needs of local, national, and international community members and take actions right on time without the need for additional specialized personnel or resources.

Team:

Dr. Koh has expertise in library assessment, reading and technology, Human-Computer Interaction, and research methods. Her knowledge of text analysis and programming have been developed through projects, coursework, and workshops such as a project where she analyzed readers' electronically or manually written notations in-between text lines and on page margins using Rhetorical Structure Theory. Dr. Koh received a grant for her dissertation from the University Graduate School at Indiana University Bloomington and was an invited participant at the Doctoral Seminar for Research and Career Development of the American Society for Information Science and Technology (ASIS&T). Dr. Fienup received his master's and Ph.D. degrees in Computer Science from Iowa State University with relevant coursework in Artificial Intelligence.

Budget:

Total anticipated project cost: $99,819

The preliminary budget includes partial academic year and summer salary support for two PIs in the amount of $47,846, fringe benefits of $14,299, student assistant salaries of $3,840, conference travel costs of $5,900, publication costs of $2,000, and Facilities & Administrative costs at UNI's federally negotiated rate of 35.1% in the amount of $25,934.