"Good Night, Good Day, Good Luck": Applying Topic Modeling to Chat Reference Transcripts

Megan Ozeran
and Piper Martin

ABSTRACT

This article presents the results of a pilot project that tested the application of algorithmic topic modeling to chat reference conversations. The outcomes for this project included determining if this method could be used to identify the most common chat topics in a semester and whether these topics could inform library services beyond chat reference training. After reviewing the literature, four topic modeling algorithms were successfully implemented using Python code: (1) LDA, (2) phrase-LDA, (3) DMM, and (4) NMF. Analysis of the top ten topics from each algorithm indicated that LDA, phrase-LDA, and NMF show the most promise for future analysis on larger sets of data (from three or more semesters) and for examining different facets of the data (fall versus spring semester, different time of day, just the patron side of the conversation).

INTRODUCTION

The library at the University of Illinois at Urbana-Champaign has included chat reference services since the spring of 2001.[1] Today, this service is extensively used by library patrons, resulting in thousands of conversations each semester. While in-person reference edges out chat for the largest number of interactions at the main library information desk over the most recent four years, chat questions have a higher number of more complex questions that incorporate teaching or strategizing.[2] Since the initial implementation of chat, the library has continually assessed and improved chat reference by evaluating the software, measuring the effectiveness and value of the service, and providing staff training.[3]

For several years, librarians at the University of Illinois have used chat transcripts for training graduate assistants and new employees and chat statistics for determining staffing. Unlike other forms of reference interactions, chat offers a textual record of the conversation, so librarians have used this unique opportunity in a couple different ways. In a training exercise, students read through actual transcripts and are guided in recognizing both well-developed and less-than-ideal interactions. They are then asked to think about ways those chat conversations could have been improved and to share strategies for doing so. Graduate assistant supervisors also use chat transcripts to evaluate the performance of individual graduate assistants, checking for appropriate levels of helpfulness and for adherence to the library's chat policies. Finally, part of the library's assessment strategy looks at chat interaction numbers, such as chats per hour, the duration of each conversation, and the level of complexity of each conversation to help make decisions about optimal chat staffing levels. However, prior to the project described here, the library had not yet

Megan Ozeran (mozeran@illinois.edu) is Data Analytics & Visualization Librarian, University of Illinois Library. **Piper Martin** (pm13@illinois.edu) is Reference Services & Instruction Librarian, University of Illinois Library.

analyzed the chat reference conversations on a large scale to understand the range and consistency of topics being discussed.

While these uses of chat data have been successful, such a large body of information from patrons about the library and its collections and services seemed underutilized. In an environment of growing data-informed decision-making, both within the broader library community and at the University of Illinois in particular, it was now an opportune time to implement this kind of large-scale topic analysis. If common themes emerged from the chat interactions beyond simply showing the most frequently asked questions, these themes could inform the library's reference services beyond just training for chat reference. For example, patterns in the number of citation questions could indicate the best times to offer a citation management tool workshop; multiple inquiries about a new resource or tool might prompt planning a new workshop; and repeated confusion regarding a service or policy may signal a need to bolster the online guides or FAQ. Since the number of chat transcripts was so large, automating analysis through a programming language such as Python seemed the best course of action. This article presents the results of a pilot project that tested the application of algorithmic topic modeling to chat conversations. The outcomes for this project included (1) determining if this method could be used to identify the most common chat reference topics in a semester; and (2) whether this information indicated if it could be used to inform reference services beyond just training for chat, such as improving FAQs, workshops, the library website, or other instruction.

LITERATURE REVIEW

Chat reference services are well established in academic libraries, and there are abundant examples in the literature exploring these services. However, there is a lack of research on ways to employ automated methods to analyze chat reference. Numerous articles approach chat analysis via traditional qualitative methods, where research teams hand-code chat themes, topics, or question categories.[4] Schiller employed a tool called QDA Miner to partially automate the otherwise human-driven coding process, using the software to automatically generate clusters of manually created codes.[5]

Only one paper appeared to explicitly address the issue primarily by using algorithmic analysis methods. In addition to conducting sentiment analysis, Kohler applied three topic modeling algorithms to chat reference conversations at Rockhurst University.[6] Kohler identified the algorithm of non-negative matrix factorization (NMF) as the "winning topic extractor" based on how evenly it distributed the topic clusters across all the chat conversations.[7] The other algorithms Kohler tested, latent Dirichlet allocation (LDA) and latent semantic analysis (LSA), had much more skewed distributions of topics. The most common topic identified by LDA appeared in so many of the chat conversations that it was essentially meaningless as a category. LDA is one the most well-established topic modeling algorithms, but as Kohler found, it does not work very well with short texts like chat conversations.

To supplement the lack of library research in this area, non-library research that has applied topic modeling to short texts was also reviewed. Interestingly, although the NMF algorithm worked well for Kohler's analysis of library chat conversations, there was little mention of NMF in the non-library literature. On the other hand, it was not surprising that LDA was one of the most commonly discussed algorithms, either as an example of what doesn't work or as a basis upon which a modified algorithm was created to perform better for short texts.[8] Another common algorithm

was biterm topic modeling (BTM). Proposed by Cheng et al., BTM takes pairs of words (biterms), rather than individual words, as the unit on which to base topics.[9] By creating biterms, the researchers increased the number of items to sort into topics, thus mitigating a common problem with analyzing short texts. A final commonly used algorithm was the Dirichlet mixture model (DMM).[10] A key feature of DMM for analyzing short texts is that it assumes each text (in this project, each chat conversation) is associated with only one topic. While longer texts like articles or books likely encompass many topics, it is plausible that a chat conversation could be summarized in one topic.

METHODOLOGY

At the time of this project (spring 2018), the library was using locally developed chat software called IWonder. The chat widget is embedded on the library homepage, on the "Ask A Librarian" page, in LibGuides, and within the library's interface for its licensed EBSCO databases. The chat service was available 87 hours per week at the time the data was collected. During the day, chat service is provided by a mix of librarians, library staff, and graduate assistants, most of whom are scheduled at the main library's information desk. Subject-specific libraries, including the engineering library, the agricultural and life sciences library, and the social sciences, health, and education library, also contribute many hours on chat reference from their respective locations. The evening and weekend shifts are all covered by graduate assistants from the University of Illinois School of Information Sciences.

The authors decided that one semester of chat transcripts would be the most appropriate corpus with which to work for this pilot project because it would encompass a substantive and meaningful (but also manageable) number of conversations. In preparation, Institutional Review Board approval was received, and a graduate student completing a degree in information management from the School of Information Sciences was selected to assist with this project through the school's practicum program. This practicum student is an experienced programmer, and his presence on the team allowed the project to proceed more quickly than if the authors had pursued the project without his expertise.

To begin the project, all chat conversations from the spring 2017 semester were obtained by querying the local server using MySQL Workbench, limiting the query to chat logs between the dates 1/17/2017 and 5/12/2017 (inclusive). Because each line of a chat conversation was saved as a separate line in the database, this meant retrieving approximately 90,000 lines of data. The actual text of the chat conversations was unstructured (by its nature), but the text was saved with related metadata. For instance, each chat conversation was automatically given a unique identifier, so the individual lines could be grouped into conversations and put in order by their timestamp. The 90,000 lines represented almost 6,000 individual conversations.

The chat logs were cleaned using a combination of OpenRefine (primarily for ASCII character cleanup) and Python code to remove personally identifiable information (PII) and to make the data easier to analyze.[11] By default, the chat software did not collect any information about patrons, but sometimes patrons volunteered PII because they thought it was needed to answer their questions. Therefore, part of the cleaning process involved removing as much of this patron PII as possible, replacing it with the word "REMOVED" to denote the change. In addition, library staff usernames were scrubbed by replacing each username with a generic "staff###", where "###" was a unique (incremented) number assigned to each original username. This maintained

the ability to track a single staff member across multiple conversations, if desired, without identifying the actual person. Another important part of the data cleaning was to remove URLs, because these would be unnecessary in identifying topics, and they significantly increased the number of unique "words" that the analysis algorithms identified. The URLs were nearly always saved within an <a> HTML tag, so most URLs were easily identified for removal. The data cleaning process has been described here in a linear fashion for ease of understanding, but over the course of the project it was actually an iterative process, as more cleaning issues were discovered during analysis.

Based on the analyses performed in the related literature, the practicum student wrote code to test five topic modeling algorithms: (1) latent Dirichlet allocation (LDA), (2) phrase-LDA (LDA applied to phrases instead of words), (3) biterm topic modeling (BTM), (4) Dirichlet mixture modeling (DMM), and (5) non-negative matrix factorization (NMF). Ultimately, the processing power and time required to implement BTM meant that this algorithm could not be implemented for this project. However, for the other four models, LDA, phrase-LDA, DMM, and NMF, were all successfully implemented. All code related to this project, including the cleaning and analysis, are available on GitHub (https://github.com/mozeran/uiuc-chat-log-analysis).

RESULTS

Outputs of the LDA, phrase-LDA, DMM, and NMF modeling algorithms are shown in tables 1 through 4. After removing common stop words, the remaining words were put into lowercase and stemmed before topic modeling algorithms were applied. The objective of the stemming process was to convert singular and plural versions of a word to a hybrid form so that they are treated as the same word. Thus, many words ending in "y" are shown ending in "i". For instance, "library" and "libraries" would both be converted to "librari" and thus be treated as the same word. The phrase "easi search" refers to "Easy Search," the all-in-one search box on the library homepage. The word "ugl" refers to the undergraduate library (UGL). The word "remov" showed up in the topic lists surprisingly frequently, probably because patron PII was replaced with the word "REMOVED." Since explicitly denoting the removal of PII is unlikely to be of import, it makes sense in the future to simply remove the PII without replacement.

Table 1: LDA	
(top 10 words in each topic)	
Topic 1	music map laptop remov find ok one also may score
Topic 2	look search find help databas thank use articl research would
Topic 3	book librari thank help check look remov reserv would els
Topic 4	help use student find articl librari hi look tri question
Topic 5	request librari account item thank ok get help loan number
Topic 6	thank chat good know one night go okay think hi
Topic 7	thank look librari remov help would contact inform find like
Topic 8	search articl databas click thank journal help page ok find
Topic 9	articl thank journal access look help remov full link find
Topic 10	access tri link thank use work get campu remov let

Table 2: Phrase-LDA

(top 10 phrases in each topic)	
Topic 1	interlibrari loan, lose chat, chat servic, lower level, chat open, writer workshop, spring break, studi room, call ugl, add chat
Topic 2	good night, great day, good day, good luck, drop menu, sound good, nice day, ye great, remov thank welcom, make sens
Topic 3	anyth els, tri find, abl find, find anyth, feel free, ll tri, social scienc, tri access, ll back, abl access
Topic 4	easi search, academ search, find articl, search box, tri search, databas subject, search bar, search term, databas search, search databas
Topic 5	graduat student, grad student, peer review, undergrad student, illinoi undergrad, scholarli sourc, univers illinoi, undergradu student, primari sourc, googl scholar
Topic 6	main librari, librari catalog, librari account, librari homepag, call number, librari websit, netid password, main stack, creat account, borrow id
Topic 7	page remov, click link, open new tab, link remov, send link, remov click, left side, remov link, page click, error messag
Topic 8	give one moment, contact inform, moment pleas, faculti staff, give minut, pleas contact, email address, staff member, faculti member, unit state
Topic 9	full text, journal articl, access articl, find articl, databas journal, light blue, articl titl, titl articl, journal databas, found articl
Topic 10	request book, request item, check book, doubl check, print copi, cours reserv, copi avail, physic copi, book avail, copi past

Table 3: DMM
(top 10 words in each topic)

Topic 1	work open chat way onlin say specif avail day sourc
Topic 2	check titl research much onlin avail day text sourc say
Topic 3	pleas sourc day onlin titl found right hello may take
Topic 4	chat also copi pleas think onlin undergrad sourc work way
Topic 5	pleas sorri found item chat way right open work time
Topic 6	found also right much think could research undergrad sorri way
Topic 7	contact hello account sorri could ask titl moment may think
Topic 8	copi onlin sorri ask think say right also much sourc
Topic 9	much research way may right think open take hello result
Topic 10	abl avail also titl catalog pleas say campu onlin take

Table 4: NMF
(top 10 words in each topic)

Topic 1	request take titl today moment way item may place say
Topic 2	specif start type journal topic research tab way subject result
Topic 3	ugl today ask wonder call may contact peopl someon talk
Topic 4	sourc univers scholarli research servic resourc tell illinoi guid librarian
Topic 5	account log set vpn us password id say campu problem
Topic 6	main locat undergradu call tab review two circul ugl number
Topic 7	reserv class time undergradu cours websit show im titl onlin

Topic 8	text full troubl problem still pdf websit onlin send moment
Topic 9	chat night hey yeah oh well time tonight take yep
Topic 10	unfortun uiuc onlin wonder version graduat print seem way grad

DISCUSSION

Interpreting the results of a topic model can be a bit of a guessing game. None of these algorithms look at the semantic meaning of words, so the resulting topics are not based on semantics. Each algorithm simply employs a different method of mathematically determining the likelihood that words are related to each other. When this likelihood is high enough (as defined by the algorithm), the words are listed within the same topic. Identifying topics mathematically is much quicker than a person hand-coding conversations. However, automatic classification also means that the resulting topics could make absolutely no sense to people, who understand the semantic meaning of the words within a topic.

This lack of coherent meaning is most present in the results of the DMM model (table 3). For instance, the words that comprise Topic 1 are the following: "work open chat way online say specify available day source." It is difficult to imagine what overarching concept links all, or even most, of these words. Only a few words appear to have any significance at all: "open" could refer to open access, or to the library's open hours; "online" may refer to finding resources online, or the fact that a student is taking online classes; and "source" is likely some reference to a research resource. These words barely relate to each other semantically, and the remaining seven words don't provide much clarification. Thus, it appears that DMM is not a particularly good topic modeling algorithm for library chat reference.

The results seen from the LDA model (table 1) appear slightly more comprehensible. In Topic 2, for instance, the words are as follows: "look search find help database thank use article research would." While not all the words relate to each other, a common theme could emerge from the words *look, search, find, database, article,* and *research*. It's possible that this Topic 2 identified chat conversations where a patron needed help finding research articles. Even Topic 6, at first glance a silly list of words, makes some sense: "thank chat good know one night go okay think hi." Greetings and sign-offs probably comprised a good number of the total words in the corpus, so it is understandable that a "greetings" topic could be mathematically identified. Overall, LDA appears to have potential in topic modeling chat reference, but it probably needs to be further tweaked.

When applying the LDA model to phrases (table 2), the coherence increases within the phrases, but the topics are not always as coherent. Topic 1 includes the following phrases: "interlibrary loan, lose chat, chat service, lower level, chat open, writer workshop, spring break, study room, call UGL, add chat." Each phrase, individually, makes perfect sense for the context of this library; as a collection, however, the phrases don't comprise one coherent topic. Four of the phrases explicitly mention chat services (an interesting meta-topic), while the rest appear completely unrelated. On the other hand, Topic 10 does show more semantic relation between the phrases: "request book, request item, check book, double check, print copy, course reserve, copy available, physical copy, book available, copy past." It seems pretty clear that this topic refers to books—whether on reserve, being requested, or checking if they are even available. With the wide difference in topic coherence, the phrase-LDA algorithm is not perfect for topic modeling chat reference, but further exploration is warranted.

The final algorithm, NFM (table 4), is also imperfect. It is possible to distill each topic into an actual semantic concept, but there is almost always at least one word that makes it a little less clear. Topic 5 probably provides the best coherence: "account log set VPN use password ID say campus problem." It seems clear this topic refers to identity verification, likely for off-campus use of library resources. The other topics given by the algorithm have more confusing elements, such as in Topic 1 where the relatively meaningless words *may, way,* and *say* all appear. It's interesting that Kohler found NMF to work very well, while the results above are not nearly as coherent as those identified in her implementation.[12] This is a perfect example of how the tuning of many different parameters can affect the ultimate results of each topic modeling algorithm. This is why the authors think it is worth continuing to explore how to improve the implementation of LDA, phrase-LDA, and NMF algorithms for chat conversations, as well as share the original code for others to test and revise. It will take many different projects at many different libraries before an optimum topic model implementation is found for chat reference.

NEXT STEPS

For the most part, the more coherent results from the LDA and NMF topic modeling algorithms support anecdotal understanding of the primary themes in chat conversations. Currently, two members of the Research & Information Services unit, the department responsible for scheduling the chat reference service at the main library, are examining the model outputs to determine whether any of the results are strong enough at this stage to suggest changes to services or resources. They will also share the results with the chat coordinators at other libraries on campus in case the results indicate changes for them. Additionally, results will be shared with the library's Web Working Group, since repeated questions about the same services or locations may suggest the need to display them in a more prominent place on the library website or provide a more discoverable online path to them. Since this was a pilot project that used a fairly small data set, it is anticipated that years of transcripts—along with improved topic model implementation—will reveal even more significant and robust themes.

With the encouraging results of this pilot project, there is much to continue to explore.[13] One future question is whether there are differences between fall and spring semesters. If some topics arise more frequently in one semester than the other, perhaps the library needs to offer more workshops during that semester. Alternatively, perhaps support materials should be created (such as handouts or online guides) that emphasize the related services and place them more prominently, while withdrawing or de-emphasizing them in the other semester. Another area for further analysis is how the topics that emerge in the late-night chat interactions compare to other times of day. This will help the library design more relevant training materials for the graduate assistants who staff those shifts, or potentially change who is staffing the shifts. Also of interest is comparing the text written by the chat operators versus the chat users, as this would further spotlight the terminology that patrons use. If patrons are using significantly different terms from staff, then modifying the language of the library's website may reduce confusion.

There are also improvements to make to the data cleaning process, such as better identifying when to remove stop words and when to remove punctuation. These steps weren't perfectly aligned, which is why; for example, the "ll" that appears in Topic 3 of the phrase-LDA results (table 2) is most likely a mutation of the contractions like "I'll," "we'll," and "you'll." Generating "ll" as a word from multiple different contractions not only created a meaningless word, but since "ll"

occurred more frequently than any unique contraction, it was potentially treated as more important by the topic modeling algorithms.

CONCLUSION

This project has demonstrated that topic modeling is one possible way to employ automated methods to analyze chat reference, with mixed success. The library will continue to improve chat reference analysis based on this project experience. The authors hope that other libraries will use the lessons from this project and the code in GitHub as a starting point to employ similar analysis for their own chat reference. In fact, a related project at the University of Northern Iowa Library is evidence of growing interest in topic modeling of chat reference transcripts.[14] Considering how frequently patrons use chat reference, is it important for libraries to explore and embrace whatever methods will allow them to assess and improve such services.

ACKNOWLEDGEMENTS

NOTES

[1] Jo Kibbee, David Ward, and Wei Ma, "Virtual Service, Real Data: Results of a Pilot Study," *Reference Services Review* 30, no. 1 (Mar. 1, 2002): 25–36, https://doi.org/10.1108/00907320210416519.

[2] The library uses the READ scale (Reference Effort Assessment Data scale), which allows reference transactions to be translated into a numerical scale that takes into account the effort, skills, knowledge, teaching moment, techniques and tools used by the staff in the transaction. See readscale.org for more information.

[3] David Ward and M. Kathleen Kern, "Combining IM and Vendor-Based Chat: A Report from the Frontlines of an Integrated Service," *Portal: Libraries and the Academy* 6, no. 4 (Oct. 2006): 417–29, https://doi.org/10.1353/pla.2006.0058; JoAnn Jacoby et al., "The Value of Chat Reference Services: A Pilot Study," *Portal: Libraries and the Academy* 16, no. 1 (Jan. 2016): 109–29, https://doi.org/10.1353/pla.2016.0013; David Ward, "Using Virtual Reference Transcripts for Staff Training," *Reference Services Review* 31, no. 1 (2003): 46–56, https://doi.org/10.1108/00907320310460915.

[4] Robin Brown, "Lifting the Veil: Analyzing Collaborative Virtual Reference Transcripts to Demonstrate Value and Make Recommendations for Practice," *Reference & User Services Quarterly* 57, no. 1 (Fall 2017): 42–47, https://doi.org/10.5860/rusq.57.1.6441; Maryvon Côté, Svetlana Kochkina, and Tara Mawhinney, "Do You Want to Chat? Reevaluating Organization of Virtual Reference Service at an Academic Library," *Reference & User Services Quarterly* 56, no. 1 (Fall 2016): 36–46, https://doi.org/10.5860/rusq.56n1.36; Donna Goda and Corinne Bisshop, "Frequency and Content of Chat Questions by Time of Semester at the University of Central Florida: Implications for Training, Staffing and Marketing," *Public Services Quarterly* 4, no. 4 (Dec. 2008): 291–316, https://doi.org/10.1080/15228950802285593;

Kelsey Keyes and Ellie Dworak, "Staffing Chat Reference with Undergraduate Student Assistants at an Academic Library: A Standards-Based Assessment," *The Journal of Academic Librarianship* 43, no. 6 (2017): 469–78, https://doi.org/10.1016/j.acalib.2017.09.001; Michael Mungin, "Stats Don't Tell the Whole Story: Using Qualitative Data Analysis of Chat Reference Transcripts to Assess and Improve Services," *Journal of Library & Information Services in Distance Learning* 11, no. 1–2 (Jan. 2017): 25–36, https://doi.org/10.1080/1533290X.2016.1223965.

[5] Shu Z. Schiller, "CHAT for Chat: Mediated Learning in Online Chat Virtual Reference Service," *Computers in Human Behavior* 65 (Dec. 2016): 651–65, https://doi.org/10.1016/j.chb.2016.06.053.

[6] Ellie Kohler, "What Do Your Library Chats Say?: How to Analyze Webchat Transcripts for Sentiment and Topic Extraction," in *Brick & Click Libraries Conference Proceedings* (Brick & Click, Maryville, MO: Northwest Missouri State University, 2017), 138–48, https://www.nwmissouri.edu/library/brickandclick/presentations/eproceedings.pdf.

[7] Kohler, 141.

[8] For example: Guan-Bin Chen and Hung-Yu Kao, "Re-Organized Topic Modeling for Micro-Blogging Data," in *Proceedings of the ASE BigData & SocialInformatics 2015*, ASE BD&SI '15 (New York, NY: ACM, 2015), 35:1–35:8, https://doi.org/10.1145/2818869.2818875.

[9] X. Cheng et al., "BTM: Topic Modeling over Short Texts," *IEEE Transactions on Knowledge and Data Engineering* 26, no. 12 (Dec.2014): 2,928–41, https://doi.org/10.1109/TKDE.2014.2313872.

[10] For example: Chenliang Li et al., "Topic Modeling for Short Texts with Auxiliary Word Embeddings," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM Press, 2016), 165–74, https://doi.org/10.1145/2911451.2911499.

[11] We used Python packages gensim, langid, nltk, numpy, pandas, re, sklearn, and stop_words for data cleaning and analysis.

[12] Kohler, "What Do Your Library Chats Say?"

[13] The library implemented new chat reference software after this project was completed, so analysis of chat conversations that took place after the spring 2018 semester will require a re-working of the data collection and cleaning processes.

[14] HyunSeung Koh and Mark Fienup, "Library Chat Analysis: A Navigation Tool," (Poster, Dec. 5, 2018), https://libraryassessment.org/wp-content/uploads/2018/11/58-KohFienup-LibraryChatAnalysis.pdf.