Evidence Based Library and Information Practice

Natural Language Processing for Virtual Reference Analysis

Ansh Sharma, Kathryn Barrett et Kirsta Stapelfeldt

Volume 17, numéro 1, 2022

URI : https://id.erudit.org/iderudit/1088075ar
DOI : https://doi.org/10.18438/eblip30014

Aller au sommaire du numéro

Éditeur(s)

University of Alberta Library

ISSN

1715-720X (numérique)

Découvrir la revue

Citer cet article

Sharma, A., Barrett, K. & Stapelfeldt, K. (2022). Natural Language Processing for Virtual Reference Analysis. *Evidence Based Library and Information Practice*, *17*(1), 78–93. https://doi.org/10.18438/eblip30014

Résumé de l'article

Objective – Chat transcript analysis can illuminate user needs by identifying common question topics, but traditional hand coding methods for topic analysis are time-consuming and poorly suited to large datasets. The research team explored the viability of automatic and natural language processing (NLP) strategies to perform rapid topic analysis on a large dataset of transcripts from a consortial chat service.

Methods – The research team developed a toolchain for data processing and analysis, which incorporated targeted searching for query terms using regular expressions and natural language processing using the Python spaCy library for automatic topic analysis. Processed data was exported to Tableau for visualization. Results were compared to hand-coded data to test the accuracy of conclusions.

Results – The processed data provided insights about the volume of chats originating from each participating library, the proportion of chats answered by operator groups for each library, and the percentage of chats answered by different staff types. The data also captured the top referring URLs for the service, course codes and file extensions mentioned, and query hits. Natural language processing revealed that the most common topics were related to citation, subscription databases, and finding full-text articles, which aligns with common question types identified in hand-coded transcripts.

Conclusion – Compared to hand coding, automatic and NLP processing approaches have benefits in terms of the volume of data that can be analyzed and the time frame required for analysis, but they come with a trade-off in accuracy, such as false hits. Therefore, computational approaches should be used to supplement traditional hand coding methods. As NLP becomes more accurate, approaches such as these may widen avenues of insight into virtual reference and patron needs.

Evidence Based Library and Information Practice

Research Article

Natural Language Processing for Virtual Reference Analysis

Ansh Sharma
Emerging Professional in Application Development
University of Toronto Scarborough Library
Toronto, Ontario, Canada
Email: ansh.sharma@mail.utoronto.ca

Kathryn Barrett
Social Sciences Liaison Librarian
University of Toronto Scarborough Library
Toronto, Ontario, Canada
Email: kathryn.barrett@utoronto.ca

Kirsta Stapelfeldt
Head, Digital Scholarship Unit
University of Toronto Scarborough Library
Toronto, Ontario, Canada
Email: kirsta.stapelfeldt@utoronto.ca

Received: 29 July 2021 **Accepted**: 14 Dec. 2021

DOI: 10.18438/eblip30014

Abstract

Objective – Chat transcript analysis can illuminate user needs by identifying common question topics, but traditional hand coding methods for topic analysis are time-consuming and poorly suited to large datasets. The research team explored the viability of automatic and natural language processing (NLP) strategies to perform rapid topic analysis on a large dataset of transcripts from a consortial chat service.

Methods – The research team developed a toolchain for data processing and analysis, which incorporated targeted searching for query terms using regular expressions and natural language processing using the Python spaCy library for automatic topic analysis. Processed data was exported to Tableau for visualization. Results were compared to hand-coded data to test the accuracy of conclusions.

Results – The processed data provided insights about the volume of chats originating from each participating library, the proportion of chats answered by operator groups for each library, and the percentage of chats answered by different staff types. The data also captured the top referring URLs for the service, course codes and file extensions mentioned, and query hits. Natural language processing revealed that the most common topics were related to citation, subscription databases, and finding full-text articles, which aligns with common question types identified in hand-coded transcripts.

Conclusion – Compared to hand coding, automatic and NLP processing approaches have benefits in terms of the volume of data that can be analyzed and the time frame required for analysis, but they come with a trade-off in accuracy, such as false hits. Therefore, computational approaches should be used to supplement traditional hand coding methods. As NLP becomes more accurate, approaches such as these may widen avenues of insight into virtual reference and patron needs.

Introduction

Librarians strive to understand patron needs in order to provide efficient and accurate reference services. Reviewing questions from different reference channels offers an opportunity to identify common question types and to ensure that library staff have sufficient training and familiarity with resources to answer them. Since virtual reference software preserves a record of the interaction, librarians can analyze chat transcripts to understand the breadth and frequency of question topics (Chen & Wang, 2019). Researchers have traditionally used qualitative methods involving hand coding to determine the most common question types on chat services, but these methods are time consuming and not suited to analyzing large datasets. Consequently, some researchers have begun to explore computational analysis for large corpora (Paulus & Friend Wise, 2019a).

Automated topical analysis of chat transcripts has not received much attention in the library and information science literature (Ozeran & Martin, 2019). Consequently, our research team at the University of Toronto Scarborough (UTSC) Library sought to explore the viability of automatic and natural language processing (NLP) methodologies for chat transcript analysis. The team hoped to uncover how these methods might streamline or complement other forms of topical analysis. As the researchers were all members of the UTSC Library's Data and Digital Scholarship Committee, a secondary goal of the project was to better understand patron needs concerning data and digital scholarship topics. The project received support from the library's Emerging Professionals Program, which extends employment to students enrolled in UTSC's Computer Science Co-op program to expose them to digital scholarship and information science.

UTSC is the eastern campus of the University of Toronto system, offering over 250 programs to 13,843 students, rendering it similar in size to a mid-sized Canadian university. UTSC is served by one library,

which is part of the larger University of Toronto Libraries (UTL) system made up of over 40 libraries. UTL participates in the *Ask a Librarian* virtual reference service, which is offered by the Ontario Council of University Libraries (OCUL), a consortium representing the libraries of all 21 universities in Ontario, Canada. The service reaches approximately 400,000 full-time equivalent students across the province. It is managed by Scholars Portal (SP), the service arm of OCUL. *Ask a Librarian* connects students, faculty members, and researchers with real-time assistance through chat for 67 hours per week during the academic year. Staffing is managed through a collaborative model, in which participating libraries provide staffing hours relative to their student populations and service usage. On evenings and weekends, staffing is supplemented by part-time virtual reference operators hired by OCUL. *Ask a Librarian* is a busy service point; it handles over 25,000 chats per year, with roughly a third of chats originating from the University of Toronto, OCUL's largest member.

The high level of activity on the *Ask a Librarian* service offers a large dataset to test the viability of natural language processing. Since a team of UTL and SP researchers had recently hand-coded a sample of chats by question type as part of an assessment project, there was also an opportunity to compare the results of automatic and manual approaches to topic analysis on the same service. Overall, the project sought to answer the following research questions:

Can a toolchain using automatic and natural language processing effectively perform broad topic analysis of chat transcripts?
What types of insights can be gathered from the processed data, regarding patrons, chat operators, and users' questions?
To what extent are patrons asking questions related to data and digital scholarship topics?
How does natural language processing compare to traditional hand coding methods in terms of accuracy?

Literature Review

Since the emergence of virtual reference services in the late 1990s, chat reference has grown in popularity and become a core library service (Matteson et al., 2011; Schiller, 2016). Almost half of universities and colleges in North America offer chat services, with roughly a quarter providing them through a consortium (Yang & Dalal, 2015). Many users prefer virtual reference to face-to-face reference due to the personalized nature of the service, as well as its convenience and immediacy (Chow & Croxton, 2014; Connaway & Radford, 2011). Given the popularity of chat reference services, it is essential to continually evaluate and improve them to ensure service effectiveness and quality. Chat reference software typically offers detailed metadata about chat interactions, granular usage statistics, and complete transcripts of conversations, all of which can be used for assessment purposes. Chat transcripts are a particularly rich source of data, as they can illuminate the topics being discussed within chats, helping librarians to identify user needs and adjust reference services and staff training in response (Chen & Wang, 2019).

Chat transcripts can be analyzed using qualitative or quantitative approaches. Qualitative methods take an iterative approach to analysis; multiple readings of the data allow patterns to emerge, and the patterns are categorized to answer research questions (Paulus & Friend Wise, 2019b). For example, one common qualitative method, thematic analysis, allows researchers to understand the topic of online conversations. Researchers code categories and themes that emerge from the transcripts of chats using their own interpretations, rather than an a priori framework. In contrast, quantitative approaches require that a predetermined theoretical framework, such as a coding scheme, be chosen at the start of the project and used to find patterns in the data, transforming the chat into a numerical representation (Paulus & Friend

Wise, 2019a). Once a quantitative picture of the chat has been generated, it can be transformed into variables or used for calculations. For instance, content analysis requires researchers to manually assign numerical values to messages based on the degree to which characteristics are present, such as the frequency of certain words, using a coding scheme. While these methods permit an in-depth understanding of chats, they require significant researcher training, and they are time-intensive and inefficient for large chat datasets (Chen & Wang, 2019).

New computational methods are being explored to conduct content analysis in an automated fashion at scale using natural language processing techniques (Paulus & Friend Wise, 2019a). For example, topic modeling is a method of computational analysis in which a collection of documents, such as chat messages or transcripts, are represented as a set of topics. The topics are identified through statistical analysis by examining patterns of word co-occurrence within the corpus. There are several algorithms for topic modeling that can reveal the semantic structure of the collection, with Latent Dirichlet Allocation being the most common (Kherwa & Bansal, 2019). Topic modeling has been used across numerous domains, including healthcare (Wang et al., 2016), education (Afacan Adanir, 2019; Willis et al., 2017), and technology (Bulygin et al., 2018). Notably, it has been employed to understand communications with customers in fields such as banking (Hristova, 2021; Pronoza et al., 2018).

The body of scholarship on computational analysis in the library and information field is limited, as most researchers have approached chat transcript analysis through traditional hand coding methods (Ozeran & Martin, 2019). Several studies have used a semi-automatic approach for chat transcript analysis. Schiller (2016) explored the learning taking place on Wright State University's chat reference service using a two-step transcript coding process. First, transcripts with the most words (representing 1% of the data) were manually coded, resulting in a codebook with five main coding categories. Then the remaining transcripts were automatically coded using a text mining software that queried and extracted text strings matching each code category. Bungaro, Muzzupapa, and Tomatis (2017) used a semi-automatic approach for selection and categorization of chats at the University of Turin. The researchers designed a script to extract blocks of text beginning with keywords related to the natural sciences to understand the extent of questions addressed to their psychology library, then manually categorized chats according to the READ Scale and reviewed how the chat evolved throughout the reference interview.

A limited number of studies have used purely algorithmic analysis methods for library chats. Chen and Wang (2019) used Latent Dirichlet Allocation topic modeling to extract topics from chat transcripts from Carnegie Mellon's library, then applied VOSviewer, a network analysis tool, to generate a term map, and found that the most prominent chat topics discussed library resources. Kohler (2017) examined chats from Rockhurst University's Greenlease Library using natural language processing techniques. They ran three different topic extraction algorithms to explore common questions and determined that Non-Negative Matrix Factorization (NMF) was the best topic extractor, due to the even distribution of its topical clusters and the clarity of topic descriptors. Latent Dirichlet Allocation and Latent Semantic Analysis had topical distributions that were much more skewed. Ozeran and Martin (2019) also tested different algorithms for topic modeling using chat data from the University of Illinois at Urbana-Champaign, determining that Latent Dirichlet Allocation, Phrase-Latent Dirichlet Allocation, and Non-Negative Matrix Factorization had the most promise for large datasets. The researchers called on other libraries to employ similar approaches on their own chat reference services. This study answers the call by using natural language processing techniques to identify the frequency of topics within a busy consortial chat reference service in the province of Ontario.

Methods

In order to access chat transcripts from the *Ask a Librarian* chat service, we submitted a formal research data request which was approved by OCUL's Ask a Librarian Research Data Working Group. The Ask a Librarian coordinator then provided 132,263 transcripts in the comma-separated value (.csv) file format, covering the period of August 2016 to January 2021. Due to the possible presence of identifiable information within the dataset, transcripts had to be anonymized, and careful attention had to be paid to data custody throughout the project. All data was uploaded to a secure server operated by the UTSC Library's Digital Scholarship Unit (DSU) for processing using a Python toolchain developed by one of the authors, Ansh Sharma. Processed data was then exported to additional .csvs and used to generate a series of dashboards using the popular data visualization software Tableau. Note that private servers and fully licensed versions of Tableau were used for this part of the process as per the agreements with OCUL for use and protection of the data.

Processing

Each record in the chat transcript .csv files provided metadata (Table 1) as well as the full-text interaction between patron and chat operator.

Table 1
Source Chat Metadata Fields

Chat Identification Number (ID)	A unique identifier applied to the chat.
Guest Identification String	A unique string generated by the chat platform to identify the guest.
Protocol used to initiate the chat (web)	The network protocol type used to begin the chat.
Queue	The queue through which the chat was initiated (institutional affiliation).
Profile	The back-end information displayed to operators during the chat; corresponds to the queue field.
Date and Time	Date and time the chat was initiated.
Wait Duration	Time (in seconds) the patron waited for an operator to pick up the chat.
Duration of the chat	How long the chat lasted, in seconds.
Operator Data	A string of text containing the operator's name and institution.
IP address of the patron	Identifiable information about the patron's location.
Referrer Link	The specific URL of the webpage used to initiate the chat.
Chat Log	HTML content of the interaction between patron and operator (the chat itself).

A toolchain developed by Sharma processed the data and exported a .csv or spreadsheet of "hits" on relevant information, including additional metadata. The resulting metadata categories for the processed data are provided in Table 2. New metadata added as a result of processing has been indicated.

Table 2
Output Chat Metadata Fields

Chat Identification Number (ID)	A unique identifier applied to the chat.
Guest Identification String	A unique string generated by the chat platform to identify the guest.
Protocol used to initiate the chat (web)	The network protocol type used to begin the chat.
Queue	The queue through which the chat was initiated (institutional affiliation).
Profile	The back-end information displayed to operators during the chat; corresponds to the queue field.
Date and Time	Date and time the chat was initiated.
Wait Duration	Time (in seconds) the patron waited for an operator to pick up the chat.
Duration of the chat	How long the chat lasted, in seconds.
Referrer Link	The specific URL of the webpage used to initiate the chat.
Referrer Domain [a]	The web domain of the URL used to initiate the chat. Used to identify sources of traffic.
Operator Institution [a]	The institution of the operator.
UofT Operator Role [a]	If the operator is from UofT, their UofT-specific role.
UofT Operator Campus [a]	If the operator is from UofT, the campus they are affiliated with.
Hit Type [a]	The type of "hit," or identified text, observed.
Hit [a]	The specific text identified within the transcript.
Hit Context [a]	The context where the hit is identified.
Sent by [a]	The user who sent the text in which the hit was identified. May either be the patron or the operator of the chat.
Proper noun classification [a]	If the hit is a proper noun, the type of proper noun it is as identified by the spaCy library.

[a] New metadata added as part of processing.

This approach allowed the team to anonymize most of the data, as the name and username of each operator, IP address of each patron, and all complete chat transcripts were removed upon export, leaving only the metadata and hits found within the chat transcripts. Beyond anonymization, additional metadata was created as a result of the processing script. As chats could be launched from several parts of a website, obtaining the distribution of referral URLs while ignoring any subdirectories allowed the

team to better aggregate categories for understanding where patrons accessed the chat service. For example, the link https://onesearch.library.redacted.ca/ask was shortened to https://onesearch.library.redacted.ca/ and extracted as a separate column under the header "Referrer Domain." The institutional affiliations of all operators were also recorded in a separate column, based on a suffix in the operator's username that corresponded to their home institution. A list of operators relevant to the University of Toronto queue along with their roles (e.g., librarian, library technician, student worker) was precompiled. This key was used in processing so that any operators found in this list had their roles identified as part of the metadata. To enable differentiation between patron and operator, metadata about whether individual messages were sent by patron or operator was programmatically determined based on the location of the guest's identification string.

The chat logs themselves were separated and programmatically analyzed. A chat of relevance would become one or more rows in the output spreadsheet, which was structured around "hits" rather than "chats." Each row identified its hit type, and the unique chat ID was maintained, allowing for the team to preserve an understanding of the context of hits and the number of chats implicated in the results set. The specific term that was located was also recorded in metadata, as well as a small snippet of the text surrounding the term, to aid in contextual analysis of the hit and to assess the validity of the hit.

Cleaning

A dictionary of "canned" or automatic phrases used by the operator to respond to the patron was developed and such responses were identified and removed from the corpus. System and automatically generated messages were excluded from all searches.

Targeted Searching (Regular Expressions)

To answer research question 2a, regarding the extent to which chat patrons are asking about data and digital scholarship topics over chat, a list of data- and digital scholarship-related query terms (identified in the Appendix) was prepared. Regular expressions were used to identify chat logs with these search terms. To reduce the occurrence of false positives in detection, terms were only recorded if appropriate sentence structure was used and the term was not part of a larger word. For example, the term Git would register as a hit in the sentence "Git Version Control" and be recorded, but such a hit would not occur in the sentence "Legitimate." Hits from targeted searching were recorded with the identifying tag "Query." All queries except those for a master's in science degree were conducted in a case-insensitive format. Regular expressions also enabled the discovery of course codes, which follow a reliable pattern of concatenated letters and numbers. The script would check if any text had identical formatting to that of a course code from the University of Toronto. The regular expression was adjusted to match course codes from all three campuses: St. George, Mississauga, and Scarborough. Any hits found were recorded with the hit type "Course Code."

The mention of potential file extensions was of interest when analyzing transcripts. A regular expression was used to search through each message and identify patterns like those of file extensions. The regular expression searched for two to five lowercase letters preceded by a "." and succeeded by a space or the end of a sentence. Such extensions are also commonly seen in web domain names, so common domain names such as .com, .ca, and .org were not included as file extensions. Any hits found were recorded with the hit type "File Extension."

As part of the process, the query terms needed to be modified and qualified in the following ways:

- The terms "utsc.utoronto.ca" and "utoronto.ca" were intended to be recorded as parts of URLs, so the regular expression pattern was adjusted when searching for these terms to unconditionally include any mention of these terms.
- To group together matches of "Library," "Librarian," and any similar terms, the term "Librar" was used for querying.
- A specific string search was used to group together "reference", "references", and similar terms. A similar process was used for the terms "citation" and "protocol."
- The terms "Nvivo" and "Nvivohub" were grouped together.
- An alternative regular expression was used to query abbreviations of a master's degree.
- An alternative regular expression was used to query abbreviations of a master's in science degree.

Natural Language Processing

While refining the approach, it was determined that it might be possible to use the popular spaCy Python library to do automated analysis of the chat transcripts and surface common topics in the corpus; it is believed that this helps in determining the ratio of common topics in chat transcripts. Each transcript was analyzed by an English natural language processing pipeline (referred to as "en_core_web_sm" by spaCy) to analyze potential proper nouns, and each proper noun was recorded as a hit with hit type "Proper Noun." Additionally, the type of proper noun identified by spaCy was also recorded. Proper nouns were analyzed by the pipeline based on the full context of the chat transcript. The usage of an additional library (Gensim) to analyze topic keywords was considered; however, its functionality was very similar to that of spaCy in identifying proper nouns. All remaining data were exported to a Tableau Software workbook to be visualized in multiple interactive ways. The types of insight gleaned from the visualizations developed were broad in scope.

Results

Using the processed data, the team was able to derive simple counts like the volume of chats from the queues for each participating institution. The data also illustrated the sharp growth in chat activity at the beginning of the COVID-19 pandemic across all queues. The team was also able to determine the percentage of chats that were answered by librarians, technicians, and student workers in the University of Toronto system. Tableau provides affordances for data to be reviewed by time period. This approach permits a deeper understanding of the chat activity of different types of staff, which has implications for training and scheduling.

The team also determined the volume of chats responded to by operators at each of the participating institutions across the consortium, as well as the alignment of the institutional affiliations of users and operators across the service. Consequently, it was possible to understand the extent to which University of Toronto questions were responded to by local operators.

The top referring URLs for the service over time were also identified, with and without subdomains, as well as the most common course codes in chat logs over time. Course code data were illuminating, revealing which departments inspired the largest number of course-specific questions, and from which campus. For example, users identifying themselves as coming from a particular course came overwhelmingly from biology departments over the time period represented by the corpus being analyzed.

Top query hit terms were "APA" and "Reference." Top popular proper nouns from the chats included popular databases, such as "ProQuest," suggesting that many patrons are sourcing help with vendor-provided databases. Most of the terms in our targeted search pertaining to data and digital scholarship topics were not discovered in the corpus. Terms that did appear within conversations mainly related to specific software for which the library negotiates licensing or that is used in research, such as SPSS. In addition, the census was a topic of several chats, suggesting that sourcing government-published data is of interest to patrons using the chat service. Interest in the census may have been increased because the 2020 census took place during the period of data analysis.

The file extension data derived suggested that most files discussed in chat related to articles and webpages, as well as citation formats, a conclusion that is borne out in the dominance of citation formats as a topic in the NLP processing results.

Discussion

Can a Toolchain Using Automatic and NLP Effectively Perform Broad Topic Analysis of Chat Transcripts?

The toolchain developed for the project was able to process the chat transcripts, identify query hits from the list of terms, and perform topical analysis using automatic and natural language processing. The process took significantly less time than hand coding and was able to handle over 130,000 transcripts, confirming the benefits of computational analysis for large datasets previously identified in the literature (Chen & Wang, 2019; Paulus & Friend Wise, 2019a; Ozeran & Martin, 2019). However, the programmatic data processing needed human mediation to interpret the meaning of some terms and to fix inaccuracies. Automatic processing simplified the anonymization of the chat transcripts. The NLP process was very successful in identifying potential data for scrubbing, but human intervention was required to interpret some hits. For example, we needed to identify where hits referred to the proper noun “Eric” as a person (where the row for the hit should be deleted) or the “Education Resources Information Center” (often also abbreviated as "Eric"). Other "false hits" included misspellings and stray punctuation, which the script could not identify. For example, “.hi” was identified as a filetype, when context suggested this was more likely a greeting. For this tertiary stage of human cleaning, having the "snippet" of text around the hit was very useful. However, this data needed to be removed before the data was exposed in Tableau, as the words around a “hit term” often contained chat patron and operator identifying information. The NLP processing also introduced inaccuracies of other types, such as misidentifying a phrase as a proper noun and incorrectly labelling the type of proper noun.

While lacking the precision of hand coding, the team believes that programmatic processing of transcripts can be a beneficial supplement to hand coding, as automatic processing requires substantially less time commitment and can therefore be performed more often and across a larger corpus, permitting teams managing chat services to respond quickly to emerging topic trends. Moreover, this approach lends itself to adoption across digital reference formats, as it can be modified to address any structured data format. In this case, the team reviewed a second set of data emerging from the email ticketing system, which could also be exported in .csv format. The code for the analysis was generalized and is licensed as open source (Sharma, 2021).

What Types of Insights Can Be Gathered From the Processed Data?

Automatic and natural language processing provided insights about activity taking place on the chat service, such as the number of questions submitted by patrons at each participating library, the number of questions responded to by the operator groups of each library, and the proportions of chats picked up by different library staff types. These insights provide evidence regarding service usage and can inform decisions about staffing and training. The processed data also showed the extent to which questions were answered by operators from the users' home library.

This dimension of the findings is of significant value when assessing service quality, as awareness of mismatches between the institutional affiliations of the user and patron can lead to user dissatisfaction (Barrett & Pagotto, 2019). The processed data also captured the top referring URLs for the service, which can inform decisions about where to place chat buttons on each library's website to drive service usage.

These insights were a valuable complement to the statistical reports offered by *Ask a Librarian*'s chat platform, LibraryH3lp, which allows library administrators to view limited information about chat volume and operator activity. The processed data permitted the observation of additional chat metadata not ordinarily visible in the reports, such as library staff type and institutional match between the operator and patron. Additionally, the data visualizations in the Tableau dashboard can help staff easily compare various fields.

In terms of topical analysis, valuable insights were gathered from course codes, file extensions, and query hits. For instance, the team learned that the biology department generated the largest number of course-specific questions. In terms of file extensions, it was found that students asked about .pdfs, a likely indication that they were searching for full-text journal articles, and citations, suggesting that they were asking questions about citation managers. For the query hits and proper nouns derived from NLP, common topics mentioned were related to citation and subscription databases, indicating that students needed help with formatting references according to different citation styles, as well as searching in library resources.

To What Extent are Patrons Asking Questions Related to Data and Digital Scholarship Topics?

The processed data indicated that few patrons are asking data- and digital scholarship-related questions through the *Ask a Librarian* chat service. Only query terms related to the census and licensed software, such as SPSS, returned hits. This supports recent work by Mawhinney (2020), which suggests that reference interactions involving data and digital scholarship are not well suited to the virtual environment. Data and digital scholarship questions require longer reference interviews and often generate follow-up meetings. Getting assistance for these types of questions requires a substantive time commitment, but virtual reference users prioritize short interactions. Patrons with data or digital scholarship questions may therefore choose a different reference medium to get assistance, such as email or in-person service.

How Does NLP Compare to Traditional Hand Coding Methods in Terms of Accuracy?

In 2016, a research team from the University of Toronto Libraries and Scholars Portal hand-coded a sample of *Ask a Librarian* chats by topic using a coding scheme developed for the service (for coding key, see Maidenberg et al., 2012; for topic analysis, see Pagotto, 2020). Of the chats in the sample, 56% were research-based (see Figure 1); of these, questions about locating known items and finding sources on a

particular topic were most common. Other recurrent question types concerned policy, library accounts, e-resources, and citations. The results of the automatic chat analysis substantiated the hand coding. The most common question types related to research and citations, with frequent topics related to searching databases, accessing PDFs, and formatting citations according to different styles.

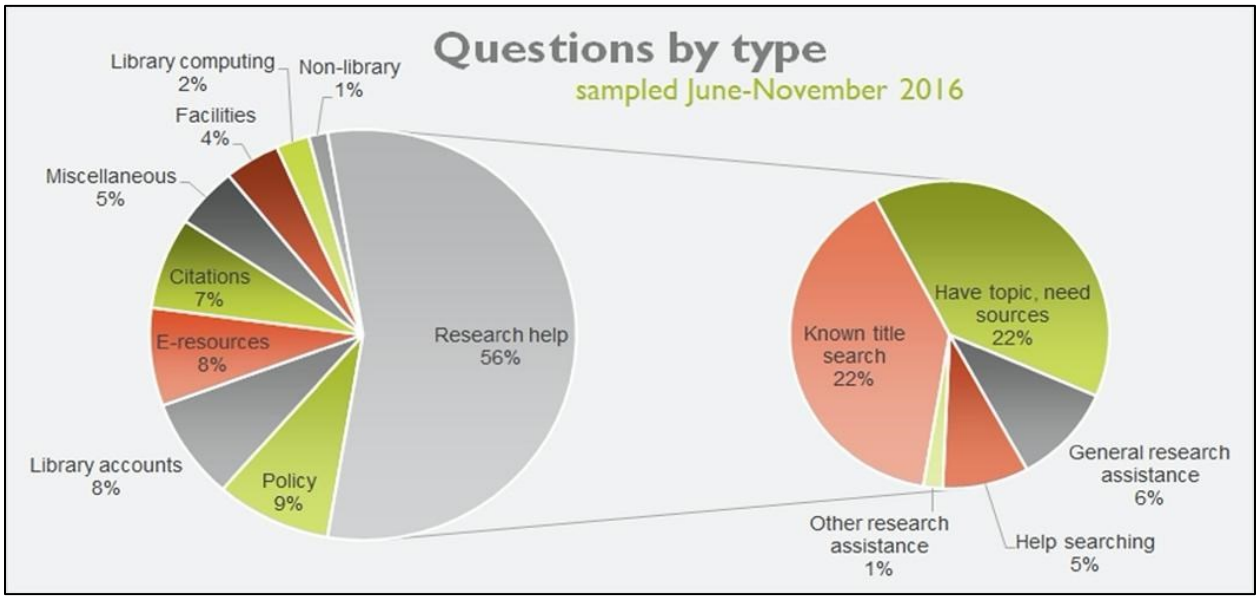

Figure 1
Ask a Librarian questions by type (hand-coded).

Inquiries about data and digital scholarship did not account for many questions in the hand-coded data. Questions related to data were coded under research help, in the general research assistance sub-category, which applied to any chat asking "how do I find this type of thing?". While this sub-category accounted for 6% of chats overall, the percentage of data-related questions on the service was lower in practice, since the category also encompassed questions about government documents, theses, and other publication types. Most digital scholarship questions would have fallen under the category of library computing, a category that accounted for only 2% of questions overall. Since few questions were found related to data and digital scholarship in the processed data, the automatic and natural language processing approach substantiated the hand-coded topic analysis.

Conclusion

Automatic and natural language processing toolchains have limitations when it comes to accuracy, and benefits when it comes to speed and the relative size of a corpus that can be analyzed. The team believes these benefits mean that automatic and natural language processing methods are well positioned to supplement hand coding, but results should be correlated with the results from studies utilizing hand coding techniques. It is possible that the precision of automatic and NLP approaches will continue to improve in accuracy. This study also provided a great opportunity for transdisciplinary learning and better understanding of emergent research tools.

Acknowledgements

The research team wishes to thank Scholars Portal, particularly Guinsly Mondésir, for providing access to chat metadata and transcripts to support this project. The team also wishes to acknowledge Judith Logan of the University of Toronto Libraries and Sabina Pagotto and Amy Greenberg of Scholars Portal for their work in hand coding chats by question type as part of the 2016 *Ask a Librarian* assessment project, which were used as a comparison point in this work. The team would also like to acknowledge David Kwasny of the University of Toronto Scarborough Library's Digital Scholarship Unit for his contributions to initial conversations about natural language processing in the library.

Author Contributions

Ansh Sharma: Methodology, Data curation, Software, Formal analysis, Visualization, Writing – original draft **Kathryn Barrett:** Conceptualization, Methodology, Data curation, Investigation, Formal analysis, Writing – original draft (lead), Writing – review & editing **Kirsta Stapelfeldt:** Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft

References

Afacan Adanir, G. (2019). Detecting topics of chat discussions in a computer supported collaborative learning (CSCL) environment. *Turkish Online Journal of Distance Education, 20*(1), 96-114. https://doi.org/10.17718/tojde.522398

Barrett, K., & Pagotto, S. (2019). Local users, consortial providers: Seeking points of dissatisfaction with a collaborative virtual reference service. *Evidence Based Library and Information Practice, 14*(4), 2-20. https://doi.org/10.18438/eblip29624

Bulygin, D., Musabirov, I., Suvorova, A., Konstantinova, K., & Okopnyi, P. (2018). *Between an arena and a sports bar: Online chats of esports spectators.* ArXiv. https://arxiv.org/abs/1801.02862

Bungaro, F., Muzzupapa, M. V., & Tomatis, M. S. (2017). *Extending the live chat reference service at the University of Turin – A case study*. 38th International Association of Scientific and Technological University Libraries Conference, Free University of Bozen-Bolzano, Italy. https://docs.lib.purdue.edu/iatul/2017/challenges/1

Chen, X., & Wang, H. (2019). Automated chat transcript analysis using topic modeling for library reference services. *Proceedings of the Association for Information Science and Technology, 56*(1), 368-371. https://doi.org/10.1002/pra2.31

Chow, A. S., & Croxton, R. A. (2014). A usability evaluation of academic virtual reference services. *College & Research Libraries, 75*(3), 309-361. https://doi.org/10.5860/crl13-408

Connaway, L. S., & Radford, M. L. (2011). *Seeking synchronicity: Revelations and recommendations for virtual reference.* OCLC Research. https://www.oclc.org/research/publications/2011/synchronicity.html

Hristova, G. (2021). Topic modeling of chat data: A case study in the banking domain. *AIP Conference Proceedings, 2333*(1), 150014. https://doi.org/10.1063/5.0044139

Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24), e2, 1-16. https://doi.org/10.4108/eai.13-7-2018.159623

Kohler, E. (2017, November 3). *What do your library chats say?: How to analyze webchat transcripts for sentiment and topic extraction.* 17th Annual Brick & Click: An Academic Library Conference, Maryville, MI, United States. https://eric.ed.gov/?id=ED578189

Maidenberg, K., Greenberg, A., Whyte-Appleby, J., Logan, J., & Spence, M. (2012). *Reference query coding key*. TSpace. http://hdl.handle.net/1807/94126

Matteson, M. L., Salamon, J., & Brewster, L. (2011). A systematic review of research on live chat service. *Reference & User Services Quarterly, 51*(2), 172-189. https://doi.org/10.5860/rusq.51n2.172

Mawhinney, T. (2020). User preferences related to virtual reference services in an academic library. *The Journal of Academic Librarianship, 46*(1), 102094. https://doi.org/10.1016/j.acalib.2019.102094

Ozeran, M., & Martin, P. (2019). "Good night, good day, good luck": Applying topic modeling to chat reference transcripts. *Information Technology and Libraries, 38*(2), 49-57. https://doi.org/10.6017/ital.v38i2.10921

Pagotto, S. (2020). *Online communication* [unpublished presentation]. Scholars Portal. https://spotdocs.scholarsportal.info/download/attachments/179013243/Online_Communication.pdf?version=2&modificationDate=1604339039000&api=v2

Paulus, T. M., Friend Wise, A., & Singleton, R. (2019a). How will the data be analyzed? Part one: Quantitative approaches including content analysis, statistical modeling, and computational methods. In T. M. Paulus & A. Friend Wise (Eds.), *Looking for insight, transformation, and learning in online talk* (pp. 127-159). Routledge.

Paulus, T. M., & Friend Wise, A. (2019b). How will the data be analyzed? Part two: Qualitative approaches including thematic, narrative, conversation, and discourse analysis. In T. M. Paulus & A. Friend Wise (Eds.), *Looking for insight, transformation, and learning in online talk* (pp. 160-196). Routledge.

Pronoza, E., Pronoza, A., & Yagunova, E. (2018, October 22-27). *Extraction of typical client requests from bank chat logs.* 17th Mexican International Conference on Artificial Intelligence, Guadalajara, Mexico. https://doi.org/10.1007/978-3-030-04497-8_13

Schiller, S. Z. (2016). CHAT for chat: Mediated learning in online chat virtual reference service. *Computers in Human Behavior, 65*(1), 651-665. https://doi.org/10.1016/j.chb.2016.06.053

Sharma, A. (2021). *digitalutsc/communication_analysis* (Version 1.03) [Source code]. https://doi.org/10.5281/zenodo.4918306

Wang, T., Huang, Z., & Gan, C. (2016). On mining latent topics from healthcare chat logs. *Journal of Biomedical Informatics, 61*, 247-259. https://doi.org/10.1016/j.jbi.2016.04.008

Willis, A., Evans, A., Kim, J. H., Bryant, K., Jagvaral, Y., & Glass M. (2017). Identifying domain reasoning to support computer monitoring in typed-chat problem solving dialogues. *Journal of Computing Sciences in Colleges, 33*(2), 11-19. https://dl.acm.org/doi/10.5555/3144645.3144647

Yang, S. Q., & Dalal, H. A. (2015). Delivering virtual reference services on the web: An investigation into the current practice by academic libraries. *The Journal of Academic Librarianship, 41*(1), 68-86. http://dx.doi.org/10.1016/j.acalib.2014.10.003

Appendix
List of Terms

Library/Librarian
Nvivo/Nvivohub
Reference
Citation
Protocol
Grad/Graduate
M.A/Master's
MSC
Data
Map
Mapping
Download
Collection
Software
Analyze
Google
Digital Scholarship
Digital Humanities
License
Open Source
Creative Commons
Wikipedia
Wikidata
Online
Digital
Visualize
Visualization
Git
Github
Metadata
[campus].[university].ca
[university].ca
ArcGIS
Python
Javascript
Undergrad
PhD
Covidence
SPSS
Qualitative
Quantitative
Tableau
RStudio
Microdata
Odesi

PowerBI
Census
Chass
Workshop
Bibliography
Works Cited
APA
MLA
IEEE
ASA
Chicago
Turabian
Vancouver
Zotero
RefWorks
EndNote
Stata
SAS
CSE
QGIS
Latex
Overleaf
ICMJE
OJS
PRISMA
JIRA
Prospero
Omeka
WordPress
Drupal
CMS
Website