Protecting Privacy in the Archives

Preliminary Explorations of Topic Modeling for Born-Digital Collections

Tim Hutchinson
University Archives & Special Collections
University of Saskatchewan Library
Saskatoon, Canada
tim.hutchinson@usask.ca

***Abstract*—Natural language processing (NLP) is an area of increased interest for digital archivists, although most research to date has focused on digitized rather than born-digital collections. This study in progress explores whether NLP techniques can be used effectively to surface documents requiring restrictions due to their personal information content. This phase of the research focuses on using topic modeling to find records relating to human resources. Early results show some promise, but suggest that topic modeling on its own will not be sufficient; other techniques to be explored include sentiment analysis and named entity extraction.**

Keywords-topic modeling, natural language processing, NLP, personal information, digital archives

I. INTRODUCTION

It has become increasingly obvious, especially as the extent of material continues to expand, that traditional appraisal and description techniques will not be suitable for born-digital collections. Indeed, even for collections dominated by paper-based and other analog materials, a growing body of literature as well as institutional practice and educational offerings have been influenced by Greene and Meissner's "More Product, Less Process" (MPLP) framework [1]. However, a concern often raised about MPLP is the lack of control over material containing personal information and requiring access restrictions [2]. At least in the Canadian legislative context, a more risk-based approach to identification of sensitive content is difficult [3]. A project that was initiated at Born-Digital Archiving & eXchange (BDAX) 2016 aims to establish a framework "to standardize and define levels of born-digital processing" [4]. The draft framework identifies the "intensive" level as one which, among other characteristics, "has specific access restrictions."

Natural language processing (NLP) offers an intriguing prospect to facilitate such elements of archival processing that remain time-intensive and require review by archival staff, often at the document level. As described further below, published studies relating to natural language processing for archives, have so far focused largely on digitized collections. The current study is focusing on born-digital collections, and especially aims to explore NLP techniques applied to identifying sensitive content and other files requiring access restrictions. The first phase of the project, reported here, experiments with topic modeling to identify documents related to human resources.

For a gentle introduction to topic modeling, see [5].

II. BACKGROUND: PROJECTS EXPLORING NATURAL LANGUAGE PROCESSING FOR ARCHIVES

Several projects in the last few years have been focusing on the potential of NLP for archival collections.

A. BitCurator

Phase three of the BitCurator project is dedicated to natural language processing. The BitCurator NLP project

> will develop software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections. The software will use existing natural language processing software libraries to identify and report on those items likely to be relevant to ongoing preservation, information organization, and access activities. These may include entities (e.g. persons, places, and organizations), potential relationships among entities (for example, by describing those entities that appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents" [6]. For topic modeling, BitCurator is using GraphLab and Gensim, with pyLDAVis for visualization [7].

At time of writing, the BitCurator software is still in development, but it's possible to install a working instance; all code is shared on GitHub [8]. Focusing on the topic modeling functionality, at this point it is possible to train a topic model and dynamically adjust the relevance setting to refine the topics. The source data can either be a set of files or a disc image. Some configuration (e.g. number of topics) is possible through the command line. The detailed results, such as identifying documents assigned to a particular topic, are not yet able to be accessed through the graphic interface. Cleanup of punctuation and application of a custom stop word list is also not yet working.

The track record and institutional support for the BitCurator project, and its maintenance strategy – with the BitCurator Consortium administered by the Educopia Institute – suggests that BitCurator NLP offers very good long-term prospects for natural language processing tools for digital archives. The BitCurator project previously developed tools for processing, analyzing and accessing disc images.

B. *ArchExtract*

ArchExtract was a demonstration project undertaken by UC Berkeley's Bancroft Library during 2014/2015. The software is web-based, built with Ruby on Rails and using the Mallet topic modeling library [9]. While the software is not being maintained, it is still possible to install a (mostly) functional version. There are some understandable limitations, and some customization was needed, but for testing and experimentation it has so far proved the most useful for the current project, avoiding the need to develop custom software or master the APIs of underlying libraries.

The original ArchExtract project focused on the Bancroft Library's John Muir collection (a digitized collection of correspondence dating 1856-1914).

> In testing the application, we found the automated text analysis tools in ArchExtract were successful in identifying major topics, as well as names, dates, and places found in the text, and their frequency, thereby giving archivists an understanding of the scope and content of a collection as part of the arrangement and description process. We called this process "dynamic arrangement and description," as materials can be re-arranged using different text processing settings so that archivists can look critically at the collection without changing the physical or virtual arrangement. … The topic models, in particular, surfaced documents that may have been related to a topic but did not contain a specific keyword or entity. [10]

C. *Other archival initiatives*

Fondz is an experimental tool developed by Ed Summers in 2013/2014 (and not currently under active development) "for auto-generating an 'archival description' for a set of born digital content found in a bag or series of bags" [11]. It bundles tools including FIDO for file format identification, exiftool for extracting image metadata, and Mallet for topic modeling.

Thomas Padilla, at the time a graduate student at the University of Illinois at Urbana-Champaign, also reported in fairly general terms in 2013 on the topic modeling of the electronic records of Carl Woese, a microbiologist at the University of Illinois [12]. This project is notable for its exploration of born-digital, contemporary archives (through a disc image of Woese's hard drive), rather than older, digitized material. Padilla used the Topic Modeling Tool, a graphical (Java-based) interface for Mallet [13].[1]

A large-scale experiment into topic modeling for digitized archival material, using archives of the European Commission and mapping topics to heading of the EUROVOC thesaurus, was reported at the 2016 Archival Computational Science workshop [15].

D. *Topic modeling for digital libraries*

Other research into topic modeling for digital libraries has also been undertaken; that is, focusing on full text of digitized and/or transcribed material. TOME (Interactive TOpic Model and MEtadata Visualization) is "a tool designed to support the exploratory thematic analysis of digitized archival collections." The test corpus for this Georgia Institute of Technology project completed in 2015 was 19th century abolitionist newspapers [16]. Similarly, historical texts are often included in the work reported by digital humanists, even if this is not in the context of complete archival collection or the function of archival arrangement and description. See, for example, research on topic modeling for a federated digital library [17]. Other early examples include "Mining the Dispatch" [18] and "Topic modeling Martha Ballard's diary" [19] Additional examples are included in the TOME white paper [16].

The Collections as Data project (principal investigator Thomas Padilla) is a broad initiative "that aims to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research and teaching in areas including but not limited to Digital Humanities, Public History, Digital History, data driven Journalism, Digital Social Science, and Digital Art History" [20]. Through periodic calls for "facets", it has the potential to be a hub for data-driven work happening in a number of institutions.

III. PRELIMINARY TOPIC MODELING RESULTS

For the purpose of this preliminary research, we used the ArchExtract software. While it is no longer under development, and there are some limitations and workarounds required, it was the best option that we located for experimentation with topic modeling that would not require a significant learning curve. For an archivist new to this area, it provided a good starting point. ArchExtract's default settings have also lent themselves to more meaningful topic groupings than has been achieved so far with BitCurator, at the current stage of development. ArchExtract also provides document-level information about

[1] The source code remains available, but at time of writing, the compiled Java application was not accessible. The tool is linked from the Milligan and Graham 2013 review of Mallet [14] as a Java GUI.

the extracted topics, which is not currently readily available through BitCurator.[2]

The text corpus used for this preliminary investigation is drawn from the records of a University of Saskatchewan Associate Vice-President for Information and Communications Technology, accessioned in 2000. While the full collection has an estimated 2500 documents, for the purpose of this research we have focused on documents that could be programmatically identified (using DROID) as Microsoft Word. That leaves 1675 documents in the full collection, but for this preliminary testing, we are reporting on a corpus of just over 1000 of those documents.[3] For text extraction using ArchExtract's built-in tools, LibreOffice was first used to convert these Word documents to PDF.

Using ArchExtract (which, as described earlier, provides a frontend for Mallet), topic modeling was run with 10, 15, 20, 40, and 100 topics (see tables I to V). The pre-processing configuration used was *Tagged—Nouns-Verbs-Adjectives, Stemmed, Removed Stop Words, Removed Rare Words*. Each configuration generates groupings of keywords suggestive to some extent of personnel/human resources. This corresponds to the type of content in this collection most likely to require restrictions.

TABLE I. 10 TOPICS: HR-RELATED TOPICS

Topic ID	Keywords
A8	people group management good work technical staff team systems position time change skill personnel support employees manager part job training
A5	areas area working level making basis addition past order expect issue made effort important time day feel make special general

TABLE II. 15 TOPICS: HR-RELATED TOPICS

Topic ID	Keywords
B4	management work support department computing time part team campus system project year resources personnel manager administration services development human systems
B12	staff technical change groups people effective assist ways group training related skills experience required times maintain communication successful person members

TABLE III. 20 TOPICS: HR-RELATED TOPICS

Topic ID	Keywords
C17	staff management work group team department good manager areas skills position client services year performance time part effort recommendation special
C11	management budget projects financial project external including impact institutional activities reporting personnel funds contract cost benefits resources implementation operating progress
C7	matter feel people notes general {first name A} unit mail talk {first name B} leave pressure sense asked result personally assignment happy things show

TABLE IV. 40 TOPICS: HR-RELATED TOPICS

Topic ID	Keywords
D6	management project resources systems process system support human functional make plan information processes role strategic procedures scope costs requirements departments
D8	manager services staff department work {first name C} management technical client {last name C} time computing September july {first name D} position personnel consulting {first name B} including
D11	issues support objectives work working developing important group met problems ideas responsibility summary efforts questions concerns role objective results people
D23	performance part year recommendation areas team time special good salary makes progress skills staff develop effectiveness job employees excellent leadership
D25	issue situation problem made case question concern long matter short money quickly regard put things past point view felt good

Unfortunately, document-level analysis so far suggests less success in terms of comprehensively identifying documents relating to human resources (HR). See Table VI for details; highlights are as follows:

- The strongest topics are A8, B4 and E25 (from the 10, 15 and 100 topic groupings, respectively), with HR-related documents representing 25% or more of the top-40 documents and 50% of the top-10 documents.
- Of the 40 highest ranked documents in each topic, the number of documents relating to HR range from 6 to 12 (median 8.5, mean 8.78).
- Of the 10 highest ranked documents in each topic, the number relating to HR ranges from 1 to 5 (median and mean 3). The best results are from the 10-topic and 40-topic groupings (five and four relevant documents in the top ten, respectively).
- The top-ranked HR document in a given topic ranges from 1 to 10, with all but two between 1 and 3.
- The Mallet scores for the HR-related documents range from 0.105 to 0.637 (median 0.383, mean 0.364)

[2] ArchExtract links the top documents under each topic; for more comprehensive analysis, the Mallet data was exported from the backend database.

[3] ArchExtract requires a flat directory (no subdirectories), so the missing documents at this point are a result of problems in automating this flattening. The data set will be refined to address this as well as to add other identifiable files such as PDF, RTF and PowerPoint.

TABLE V. 100 TOPICS: HR-RELATED TOPICS

Topic ID	Keywords
E7	employees responsibility actions group respect decisions continuing provide attention general member manner set personal employee person clients term members supervisor
E24	mail power manner harassment history conversation requests advantage departmental treatment knew behaviour differential drive starting acceptable thought game excessive morale
E25	performance part good effectiveness salary recommendation areas group year excellent appraisal employees role team complete makes assistant carry special significant
E27	position management experience support client manager change director personnel department team work staff positions duties job years successful aspa[4] level
E55	employment equity group responsible assist managers communication term issues tasks selection collaboration efficient information people application processes communicate effective organization
E88	completed leave employee aspa improvement increased dcs units cupe ability responsibilities member project communicate market forward added directly similar supervisory

TABLE VI. HR-RELATED DOCUMENTS

# Topics	Topic ID	Top 40 docs	Top 10 docs	Highest rank	Highest score
10	A5	7	2	2	0.483
10	A8	11	5	3	0.525
15	B4	10	5	3	0.383
15	B12	11	2	1	0.484
20	C7	6	3	3	0.558
20	C11	12	3	3	0.508
20	C17	8	2	3	0.361
40	D6	11	3	1	0.398
40	D8	6	2	2	0.178
40	D11	8	1	10	0.105
40	D23	9	4	2	0.382
40	D25	7	4	2	0.135
100	E7	8	3	1	0.637
100	E24	8	2	3	0.346
100	E25	11	5	1	0.395
100	E27	7	2	5	0.126
100	E55	9	3	3	0.314
100	E88	9	3	1	0.233

Because of the structure of the collection being used for this initial investigation, it is also possible (without document-by-document review of the full set) to pull out a number of documents known to be HR-related, and see how they score against the extracted topics. There are 79 documents in the personnel subfolders of the Department of Computing Services. All but four of them were categorized in at least one of the HR-related topics (500 documents are assigned to each topic). Thirty-four documents are found in the top-40 rankings; and 41 in the top-60 rankings. That is, focusing on the highest ranked documents, about half of the known HR documents are not immediately surfaced through these topic models. Further analysis is needed, but it appears that the topics relating to the functional areas covered by these positions (e.g. financial administration, systems, etc.) – and discussed in the documents – edge out the human resources nature of the documents. Improved techniques in training the topics may help address this.

The maximum score for each document (across topics), ranges between 0.623 to 0.000777. The top 44 documents score above 0.1 (top seven above 0.3); the next 30 range between 0.02 and 0.09; with the final document representing the extreme minimum.

If we limit the corpus to the documents in the personnel subfolders for the Department of Computing Services, generating a single topic leads to:

> department services {last name H} computing date time subject {first name H} work staff management university support performance campus group president development areas building

where the individual name in this case is that of the author of the memos. Training topics on a larger set of known HR documents might be a viable approach.

[4] Abbreviations in E27 and E88: ASPA and CUPE are University of Saskatchewan bargaining units; DCS was Department of Computing Services.

IV. OBSERVATIONS AND FURTHER RESEARCH

Based on the results so far, topic modeling seems to be more successful for high-level identification of topics than drilling down to the document level. Topic modeling can help identify documents that need to be reviewed, but there is a potential for them to be buried among other documents. Training topic models focused on management, planning, human resources, etc. might at least narrow down the documents needed to be individually reviewed.

An important caveat, however, is that the training of these topic models could certainly be more refined. There may be other combinations of pre-processing configuration that should be tested. It is also important to note that it has not yet been possible to apply a custom stop word list. For example, for the current corpus we should likely remove words found on departmental letterhead, although named entity extraction might help with that too.

Other directions for further testing and research include:

- Analysis of overlap between topics
- Analysis of document scores and rankings.
- Train topic models based on documents known to be relevant, e.g. a group of HR documents. Efron, Organisciak, and Fenlon have developed a method supporting the claim that "it may be beneficial to induce topic models using less, higher-quality data" [17]
- Correlate results from topic modeling with sentiment analysis of the same documents. The use of sentiment analysis has also been proposed by Baron and Borden [21].
- Correlate results from topic modeling with named entity extraction. For example, multiple occurrences of the same individual name might suggest a document is about that individual.

REFERENCES

[1] Mark A. Greene and Dennis Meissner, "More Product, Less Process: revamping traditional archival processing," American Archivist, vol. 68 (Fall/Winter 2005), pp. 208-263.

[2] Stephanie H. Crowe and Karen Spilman, "MPLP @ 5: more access, less backlog?" Journal of Archival Organization vol. 8, issue 2 (2010), pp. 110-133; see especially footnote 5.

[3] Jeremy Mohr, "An evaluation of More Product Less Process (MPLP) processing methods at the Provincial Archives of Saskatchewan," MPA project, School of Public Administration, University of Victoria, August 2016. URI: http://hdl.handle.net/1828/7715, accessed 10 October 2017.

[4] Shira Peltzman, Sally DeBauche, Erin Faulder, Kate Tasker, and Dorothy Waugh, "What we talk about when we talk about processing born-digital: building a framework for shared practice," Annual Meeting of the Society of American Archivists, 27 July 2017, Porland, Oregon.

[5] Ted Underwood, "Topic modeling made just simple enough," https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/, accessed 16 August 2017.

[6] BitCurator NLP project website, https://www.bitcurator.net/bitcurator-nlp//, accessed 26 September 2017

[7] BigCurator NLP project wiki, https://wiki.bitcurator.net/index.php?title=BitCurator_NLP, accessed 26 September 2017

[8] BitCurator NLP topic modeling repository, https://github.com/BitCurator/bitcurator-nlp-gentm, accessed 8 September 2017.

[9] [ArchExtract repository, https://github.com/j9recurses/archextract, accessed 11 September 2017.

[10] Mary Elings, "Using NLP to support dynamic arrangement, description, and discovery of born Ddgital collections: The ArchExtract experiment," May 2016, https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/, accessed 26 September 2017). See also Mary Elings, "Using NLP to support dynamic arrangement, description, and discovery of born digital collections: The ArchExtract experiment," nlp4arc 2017 (BitCurator workshop), University of North Carolina, February 2017. https://www.bitcurator.net/wp-content/uploads/2016/12/elings.pdf, accessed 26 September 2017.

[11] Fondz repository, https://github.com/edsu/fondz, accessed 6 October 2017.

[12] Thomas Padilla, "Topic Modeling Archival Materials: Guest Post," 1 November 2013, http://e-records.chrisprom.com/topic-modeling-archival-materials/, accessed 26 September 2017, cited in [10]. See also poster presentation, http://www.thomaspadilla.org/projects/tm/woese_poster.pdf (accessed 26 September 2017).

[13] Google Code Archive: Topic Modeling Tool, https://code.google.com/archive/p/topic-modeling-tool/, accessed 26 September 2017.

[14] Ian Milligan and Shawn Graham, "Review of MALLET, produced by Andrew Kachites McCallum," Journal of Digital Humanities, Vol. 2, No. 1 (Winter 2012). http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham/, accessed 26 September 2017.

[15] Simon Hengchen, Mathias Coeckelbergs, and Seth van Hooland, "Exploring archives with probabilistic models: Topic modeling for the valorization of digitised archives of the European Commission," 2016 Workshop on Computational Archival Science (IEEE International Conference on Big Data), December 2016. http://dcicblog.umd.edu/cas/ieee_big_data_2016_cas-workshop/, accessed 6 October 2017.

[16] TOME white paper, http://4humwhatevery1says.pbworks.com/w/file/fetch/104296913/TOMEwhitepaper.pdf, accessed 26 September 2017 (cited in [10]).

[17] Miles Efron, Peter Organisciak, and Katrina Fenlon, "Building topic models in a federated digital library through selective document exclusion," ASIST 2011, New Orleans, October 2011, http://people.ischool.illinois.edu/~mefron/papers/ASIST2011.pdf, accessed 26 September 2017.

[18] Robert K. Nelson, "Mining the Dispatch", http://dsl.richmond.edu/dispatch/, accessed 26 September 2017, cited in [14].

[19] "Topic modeling Martha Ballard's diary," http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/, accessed 26 September 2017, cited in [14].

[20] Collections as Data project site, https://collectionsasdata.github.io/, accessed 6 October 2017.

[21] Jason R. Baron and Bennett B. Borden, "Opening up dark digital archives through the use of analytics to identify sensitive content," 2016 Workshop on Computational Archival Science (IEEE International Conference on Big Data), Washington, DC, December 2016. http://dcicblog.umd.edu/cas/ieee_big_data_2016_cas-workshop/, accessed 6 October 2017.