Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing

Tim Hutchinson

University Archives and Special Collections, University of Saskatchewan,

Saskatoon, Canada

Records Management Journal, Special Issue: Technology and records management: disrupt or be disrupted? Volume 30, Issue 2, https://doi.org/10.1108/RMJ-09-2019-0055

Abstract

Purpose – This study aims to provide an overview of recent efforts relating to natural language processing (NLP) and machine learning applied to archival processing, particularly appraisal and sensitivity reviews, and propose functional requirements and workflow considerations for transitioning from experimental to operational use of these tools.

Design/methodology/approach – The paper has four main sections. 1) A short overview of the NLP and machine learning concepts referenced in the paper. 2) A review of the literature reporting on NLP and machine learning applied to archival processes. 3) An overview and commentary on key existing and developing tools that use NLP or machine learning techniques for archives. 4) This review and analysis will inform a discussion of functional requirements and workflow considerations for NLP and machine learning tools for archival processing.

Findings – Applications for processing e-mail have received the most attention so far, although most initiatives have been experimental or project based. It now seems feasible to branch out to develop more generalized tools for born-digital, unstructured records. Effective NLP and machine learning tools for archival processing should be usable, interoperable, flexible, iterative and configurable.

Originality/value – Most implementations of NLP for archives have been experimental or project based. The main exception that has moved into production is ePADD, which includes robust NLP features through its named entity recognition module. This paper takes a broader view, assessing the prospects and possible directions for integrating NLP tools and techniques into archival workflows.

Introduction

There has recently been a lot of attention within the archival community on archives as data, implemented through natural language processing (NLP) and machine learning techniques. This has particularly been the case for access to digitized collections, such as support for digital humanities, but there is also an increasing interest in applications for appraisal and related functions. An important

context for this activity is the challenge of dealing with huge volumes of digital records and making them publicly accessible.

Most implementations of NLP and machine learning for archives have been experimental or project-based. The main exception that has moved into production is ePADD, which includes robust NLP features through its named entity recognition module. Another promising development was the BitCurator NLP project, but that project is now complete and the resulting tools have a high technical barrier. Through a review of literature and software, this paper takes a broader view, assessing the prospects and possible directions for integrating NLP tools and techniques into archival workflows.

Overview of Natural Language Processing and Machine Learning

As an introduction to the literature review and discussion of software tools, we will first provide a brief overview of concepts relating to natural language processing (NLP) and machine learning, highlighting methods and terms referenced throughout the paper.

NLP focuses on text processing. Examples include named entity recognition, parts of speech tagging, and semantic role labelling.

Machine learning can be defined as "the study of computer algorithms that improve automatically through experience" (Mitchell, 1997). Examples of traditional machine learning algorithms include support vector machines (SVM), logistic regression, and Bayesian methods. Algorithms used in deep learning are generally called neural networks.

NLP and machine learning are certainly interrelated. In many cases the outputs of NLP techniques, such as text preprocessing, are used as inputs for machine learning; and vice-versa, such as applying supervised machine learning to the construction of dictionaries for named entity recognition. For a more general discussion of this interaction see Young *et al* (2018).

Text conversion can be a prerequisite for NLP and machine learning or, ideally, integrated into those processes. Text must first be extracted from the source files, which are often binary files such as Word or PDF files, or even images or audio files. Textract [1] is an open source package that brings together several tools for this purpose. Once the files to be analyzed are in text format, various types of cleanup will generally improve results, depending on the techniques to be applied. Examples include removal of stop words, removal of punctuation, and normalization such as stemming: for example, so that "archiving", "archived", and "archive" will all be interpreted as the same word (Koenig, 2019).

Application of **regular expressions** is a basic technique for NLP, and is certainly used outside that context. They enable pattern matching; syntax is available, variable dependent on the programming language or platform, to cover various combinations of matches (e.g. Van den Rul, 2019). Examples of applications for archival records include recognition of e-mail addresses, credit card numbers, URLs, and national identification numbers. Bulk Extractor [2], integrated with the BitCurator Environment, is a good example of an established tool for the application of regular expressions.

Named entity recognition (NER), or named entity extraction, is a form of supervised machine learning. Unstructured terms are classified (and tagged in context) using defined categories such as names, organizations, geographic locations, artworks, medical terms, and buildings. While it would be possible

to build a named entity model simply using existing lists of entities (e.g. name authorities), well developed systems are trained using various data sets, to allow the extractor to recognize names in context (Gupta, 2018).

Topic modelling is a type of probabilistic statistical modeling. Topics are understood as "a collection of words that have different probabilities of appearances in passages discussing the topic…. Topic modeling is a way of extrapolating backward from a collection of documents to infer the [topics] that could have generated them" (Underwood, 2012). Latent Dirichlet Allocation (LDA) is one of the most common models used for topic modelling (e.g. Debortoli *et al*, 2016). In its classic form, probabilistic topic modeling is a form of unsupervised machine learning, although there are methodologies for introducing semi-supervised or supervised learning into the process. The "topics" are represented as a series of keywords, ideally with information about frequency of terms.

Classification is a type of supervised machine learning. A classification model is trained using a collection of examples, such as documents, that have been labelled with their correct classification. This model makes predictions about the correct classification for new documents; "algorithms can learn to recognize a combination of features that is the 'fingerprint' of a given category" (Underwood, 2012).

To measure the accuracy of classification tools, the most commonly used metrics are recall and precision, defined as:

> Recall: (Number of relevant documents identified) / (Number of relevant documents in the dataset)
>
> Precision: (Number of relevant documents identified) / (Number of documents identified)

It is generally accepted that there is a trade-off between recall and precision (e.g. Buckland and Gey, 1994).

Literature Review

This review focuses on natural language processing (NLP) and machine learning in the context of archival institutions and practice, rather than the management of current records.

Elragal and Päivärinta (2017, p. 1) have articulated a framework for archives and big data to include a broad range of possible applications: "analytics-enhanced appraisal, analytics-prepared preservation, analytics-enhanced opening, and analytics-enhanced use." Using this framework, this review focuses especially on analytics-enhanced opening, and to some extent analytics-enhanced appraisal. There has been much more attention in the literature and active research projects on analytics-enhanced use: that is, access to digital archives, both digitized and born digital, with an arguable emphasis on the former. Other recent contributions delve into broader conceptual and ethical issues (e.g. Mordell, 2019; Moss *et al*, 2018), but even these have a focus on the access end of the spectrum.

Appraisal and Selection

Lee (2018) outlines a number of NLP and machine learning opportunities relating to appraisal and selection, and some of the history of related efforts. One area that remains largely untapped relates to metadata extraction:

> There are substantial opportunities to improve metadata export and timelining facilities for collections containing born-digital records, as timestamps often are automatically recorded (e.g. in email headers, filesystem attributes of files) during their production and use (Lee, 2018).

Goodman (2019) explored the topic modelling tool created by the BitCurator NLP project with a group of archivists, to test and discuss its potential application for archives and integration into processing workflows, particularly appraisal.

The New South Wales State Archives (Australia) undertook a pilot project involving the application of supervised machine learning to classifying unstructured text against a retention and disposition authority (Rolan *et al*, 2018). The project used both Multinomial Naïve Bayes and Multi-Layer Perception, with the best results from the latter (with up to 84% accuracy). The project also highlights the importance of cleaning the data, with a four to six percent improvement in accuracy using the cleaned data.

Description and Access

Cain (2016, p. 216) presents a case study of topic modelling applied to preliminary description of records. "The process I describe allows for the creation of minimal description to facilitate greater access to users, who can then make more in-depth connections with the collection." The documents analyzed were recently declassified presidential records of Bill Clinton. Cain used MALLET [3], with The Topic Modelling Tool (TMT) [4] as a user interface. The article also outlines text conversion and pre-processing such as text "scrubbing," another important set of preliminary steps for any NLP or machine learning processes; this is integrated into some tools.

Ed Summers' Fondz tool [5], a command-line system bundling MALLET and a few related tools, was an early experimental attempt to use topic modelling to generate a basic descriptive record.

Clough *et al* (2011) used NLP tools to extract geographic names, primarily from catalogue (descriptive) records, and produced a United Kingdom gazetteer as linked open data. The application of NLP to descriptive records rather than unstructured archival records also highlights opportunities to use NLP in the implementation of linked open data, as also explored by Gracy (2014). In a similar vein, Bryant *et al* (2018) developed a harvesting methodology to synchronize hierarchical archival descriptions across institutions.

An example of metadata extraction from digitized records is a project undertaken with record cards relating to Japanese-American internment during World War Two, with plans to interpret the contents using linked data methods (Underwood *et al*, 2017). Elements of the project relating to sensitivity review are further explored in Marciano *et al* (2018), referenced below.

As noted earlier, there has been a lot of activity relating to access to digital archives and 'archives as data.' This topic has been fairly well covered in recent literature (e.g. Mordell, 2019; Moss *et al*, 2018). See Lee and Woods (2017) for an overview of projects particularly in the digital humanities. There is certainly potential for tools developed for the digital humanities to have application for archival processing, especially to "identify and expose … contextual entities" (Lee, 2018). One project in the archival community, reported by Cox *et al* (2018), applied NLP techniques to the Legacy of Slavery Project in Maryland, including automation of metadata cleaning and transformation, and data visualization. "Always Already Computational: Collections as Data" [6] brought together several projects

focused on digitized records, although the overall scope was intended to include born-digital records. This project continues as "Collections as Data: Part to Whole."

It is also worth highlighting efforts to apply NLP and related techniques to web archives – and more generally to provide various access points to the massive web archives that have been gathered for over twenty years. While still focused on access considerations rather than appraisal or other processing functions, web archives are a type of born-digital records, so there may be even greater potential for broader application of tools and methods. A notable ongoing project in this area is Archives Unleashed [7].

While this review focuses on applications of machine learning and NLP to born-digital records, it is worth noting some activities relating to other formats that could have benefits for archival workflows, particularly in the description and access area.

The READ project (Recognition and Enrichment of Archival Documents) is focused on automated transcription of handwritten historical documents. The Transkribus service platform is publicly available; in November 2019, a co-operative was established to sustain and further develop the platform [8].

Facial recognition for both photographs and video has obvious applications for archives and has been the subject of testing for several years; see, for example, Banerjee and Anderson (2013), Ramanan *et al* (2007), England *et al* (2019).

Sensitivity Review

Several studies focus on applying NLP and machine learning to sensitivity reviews, particularly to identify personal information in records. This is a growing challenge particularly for public archives subject to privacy legislation.

Jason Baron and colleagues (Baron and Borden, 2016; Payne and Baron, 2017) have set out a research agenda for developing more robust computational techniques for privacy reviews, particularly from the perspective of legal e-discovery. The 2017 paper goes into more depth about existing methods and their potential.

The TOMES project was focused on processes to transfer e-mail accounts from hosted platforms, and the development of an appraisal tool [9]. The software is billed as using NLP to tag:

> Names, locations, organizations specific to state government, sensitive personally identifiable information (PII), such as social security numbers or credit card numbers, [and] information defined as confidential by law, such as personnel information or health records. [10]

The outputs of the TOMES project include an entity dictionary, which includes some regular expression definitions, and the tool has been demonstrated at recent conferences, but it is not clear from the project documentation if or when the tool itself will be publicly available.

A study by McDonald *et al* (2014) used 1,111 UK government records, focusing on screening for international relations restrictions and personal information. The study demonstrated the effectiveness of considering named entities in conjunction with text classification:

> We found that two features, namely the number of people in specific roles of interest and a risk score for countries identified within a record, can help to identify sensitive records that risk damaging international

> relations, by improving on a text classification baseline. We further found that these features did not help to improve BAC [Balanced Accuracy] for personal information sensitivities. This illustrates the need for individual feature sets to identify different aspects of sensitivity. (p. 505)

Another study relating to automation of sensitivity reviews focused on Japanese-American WWII incarceration camp records (Marciano *et al*, 2018). The development of tools for this collection included pattern recognition, in particular the position of names and dates. This is a good example of more customized tools and techniques being more appropriate for certain collections and record types. A more general text-based machine learning process would be less likely to be successful in this case.

In a similar way, a project to analyze declassified U.S. Department of State cables developed a "computational analysis workflow … dynamically, in a manner that resembles traditional processing except that it incorporates the expertise of both the archivist and the computer scientist" (Esteva *et al*, 2013). The cables are fairly well structured, but the project team determined that more granular – limited by time period – training and testing sets yielded better results.

Sensitivity review was also an important focus of the Presidential Electronic Records Pilot System (PERPOS) developed by the Georgia Tech Research Institute (Underwood 2008, 2009, 2010). One outcome of the project was development of methods to automatically identify "speech actions." The project technical reports also highlight the fact that necessary part of any NLP or machine learning process includes extraction of computer-readable text and, related, automated identification of file formats. Both are non-trivial challenge on their own; advances on both fronts, for example, with the PRONOM file format database [11] and related tools such as DROID [12], FIDO [13], and Siegfried [14], have made NLP easier to pursue.

A number of studies explore the interplay between human reviewers and machine learning tools.

Gollins *et al* (2014), while acknowledging that a fully automated review process is unlikely to be acceptable, observed that studying how human reviewers do their work should help build automated processes:

> Our work in developing our test collection has shown the value of close observation and study of human reviewers in beginning to understand the nature of sensitivity. It also helped us to identify additional document and context features to classify for sensitivity; the application of a simple bag-of-words text classification baseline appears inadequate. The development of a learned classifier, drawing on features extracted from a representative test collection, appears to be a fruitful starting point to develop a decision support and review prioritisation tool.

Kaczmarek and West (2018), as part of a project in partnership between the Illinois State Archives and the University of Illinois, report on the deployment of commercial e-discovery tools for classification of e-mail, particularly supervised machine learning elements of those tools. The authors report that "preliminary findings support the use of predictive coding as an effective tool to enable digital preservation at scale." The tools used include Advanced eDiscovery [15], Recommind [16], Ringtail [17], and Luminoso [18].

Predictive coding is an iterative classification process. A more recent grant report (Joens and Kaczmarek, 2019) provides more details about benchmarks and success rates:

> Finding "restricted" documents using predictive coding has proven more difficult than "archival" vs "non-archival". We were able to render our desired results (95% recall with 80% accuracy) for "archival" vs" non-archival" by manually tagging only 5,300 documents. For "restricted" vs "public" it has required us to manually tag 20,800 documents and apply a "rebalancing" technique to the dataset on two separate occasions. … The need for such a greater quantity of documents to be manually tagged is due to the low volume of "restricted" content to be found in the entire corpus.

Similarly, Cormack and Grossman (2017, p. 5) propose approaches to the development of technology-assisted review systems so that "hybrid human-computer systems can improve on both the accuracy and efficiency of human review alone." Tests simulating technology-assisted review, using two large data sets, achieved higher precision and recall than reviews undertaken manually.

A report by the UK National Archives (2016) explores the viability of commercial e-discovery tools for sensitivity reviews as well as appraisal and selection. The report notes that such tools "are good enough for use in courts" (p. 5). The report outlines a "funneling" method to reduce the volume of records to be reviewed manually. The focus relating to sensitivity reviews was on personal information, which is the relevant exemption for about 75% of the Archives' access requests; it was also assumed that records including personal information "could be more easily defined (i.e. their format and length can be predicted)" (p. 10).

Techniques supporting the "funneling" approach including categorization (topic modelling), classification through supervised machine learning, e-mail visualization, regular expressions (e.g. e-mail addresses, credit card numbers), and keyword matching. The report emphasizes the importance of being able to improve results through iterative user intervention (pp. 19-20).

The report concludes that "technology-assisted review using eDiscovery software can support government departments during appraisal, selection and sensitivity review as part of a born-digital records transfer to The National Archives" (p. 25). Unfortunately, the report does not disclose any detailed results in terms of recall, precision, etc., although it discusses the definition of accuracy in this context. The specific software tested is also not disclosed: "The intended outcome of the trials was to learn about technologies and useful features, and not to choose or recommend a specific software tool or supplier" (p. 11).

McDonald *et al* (2018) tested two active learning strategies to propose a methodology for systematically incorporating feedback from human reviewers and thereby "improve upon the raw active learning strategies to develop effective sensitivity classifiers more quickly, i.e. using less reviewer effort." Taking a different approach on technology-assisted reviews, the same research group (McDonald *et al*, 2019) found that providing reviewers with sensitivity classification predictions from an automated classifier produced measureable improvements in human accuracy and speed.

The Public Record Office Victoria (Australia) has similarly undertaken a proof of concept project relating to appraisal of e-mail records, using a commercial e-discovery tool (Rolan *et al*, 2018).

The author (Hutchinson, 2018) experimented with supervised machine learning for sensitivity review, focusing on personal information in documents relating to human resources, using the open source Weka tool. There were promising results, particularly with high recall scores. As consideration for further research, we noted that more granular training sets may be more effective, similar to the conclusion by McDonald *et al* (2014).

General Considerations

Greenberg (1998) explored the concept of NLP for archives at a point when "virtually no empirical testing had been done" in this area. While the focus of the study was on indexing and accessing archives, it includes an important reminder about archival context that is worth considering in the design of any NLP or machine learning systems for use in archival processing.

> In an effort to take full advantage of NLP, archivists need to support systems with a sophisticated linking feature and a mechanism for both bottom-up and top-down indexing and accessing options. This sophisticated linking feature must permit any retrieved record to be viewed within the context of the recordkeeping system from which it emerged. That is, rather than pulling a record from the context of its recordkeeping system, a retrieved record should serve as a means of entry (a link) into its recordkeeping system. (pp. 421-422)

Similarly, we need to incorporate contextual information and take advantage of knowledge of the archivists. This is an important part of being able to apply the tools appropriately, and interpret results. For example, a participant noted in Goodman's (2019) study: "I am not sure how this [the topic modelling tool] would help me analyze the collection if I didn't already know about the collection" (p. 28). Sensitivity is also dependent on context: "who said what to whom in what circumstances" (Gollins *et al*, 2014). McDonald *et al* (2017) demonstrated a measurable improvement in automated reviews by adding such semantic analysis to the model.

This is a rapidly developing area with increasing attention from both academics and practitioners. A new project led by the University of North Carolina (Chapel Hill), RATOM: Review, Appraisal, and Triage of Mail is promising. This project is extending the functionality in the TOMES and BitCurator Environment tools. Development efforts are focusing on a software library to "produce reports describing content and metadata, and apply NLP to extract and categorize entities"; and a "selection and appraisal web application … [for] reviewing individual email messages for retention, redaction, and public release" [19]. The project team recently hosted a workshop featuring talks exploring, among other topics, workflows and interoperability for machine learning, which are important considerations in moving towards operationalizing these tools for use by archivists (Higgs, 2019).

Software Tools

Several software tools have been developed for NLP and machine learning. This overview focuses on free tools that have been developed or customized for use specifically by archives.

ePADD

ePADD was developed by Stanford University Libraries for processing e-mail archives through a range of functions. For this overview we have focused on the appraisal module. See Schneider *et al* (2019) for an overview of the software, along with current user documentation [20]. Among available NLP software dedicated to archival processes, it is the most mature, so it is worth exploring its functionality in some detail.

The key NLP feature is identification of named entities. A custom NLP toolkit was developed for ePADD "which is used for named entity extraction, disambiguation and other tasks. This toolkit uses external datasets such as Wikipedia/DBpedia, Freebase, Geonames, OCLC FAST and LC Subject Headings/LC Name Authority File" [21].

Entity identification seems quite robust. In informal testing of a set of about 1,500 e-mail messages from the author's personal e-mail account, restricted to folders relating to a community choir, the main false positives were organization names whose underlying personal names were extracted separately (e.g., Elmer Iseler instead of Elmer Iseler Singers, Harry Fox instead of Harry Fox Agency). In some cases it would be more difficult to make this distinction without user intervention, since there are also abbreviated references to these organizations.

Entities can be edited to some degree – merged through a text editor interface, and suppressed by creating a text file. It does not appear to be possible to add new entities, or move an entity to a different category.

ePADD also includes a "lexicon analysis" feature: "ePADD employs lexicon analysis to search email messages for terms associated with personal or restricted information, which might indicate the need for further review. ePADD ships with several default lexicons. The 'Sensitive' lexicon can be used to assist in the identification of email messages with the potential for confidential content" [22].

The lexicon functionality allows users to edit lexicons and add new lexicons, although not in an interactive way during browsing of identified messages.

In the documentation, lexicon analysis is not directly billed as an NLP machine feature, although this general categorization is suggested in a recent article about ePADD: "Over the past six years, ePADD has pioneered and refined the application of machine learning and natural language processing to confront the challenges inherent in donating, administering, preserving, or accessing email collections. These include screening email for confidential, restricted, or legally protected information, preparing email for preservation, and making the resulting files (which incorporate preservation actions taken by the repository) discoverable and accessible to researchers" (Schneider *et al*, 2019, p. 306). With these sorts of broad claims, there is a risk that archivists considering ePADD will misunderstand the scope of NLP and machine learning options. On closer examination, the lexicon feature does seem to be subject to the expected limitations of keyword searching, and as such is good example of the potential of NLP or machine learning for this kind of analysis.

In the test set, the precision score is generally extremely low, with keywords out of context leading to many false hits. For example, 231 messages were identified as relating to "recreational and performance enhancing drug use." This is explained by several messages including the word "fire," in the context of fire inspection or fire marshall, and in one case a concert name "Yuletide Fires." These same messages were identified by the lexicon search as being related to employment. Some messages ran afoul of the filter based on parts of names: "Adam" (correctly identified as a named entity) and "Junk" (a last name, but not picked up by the entity analysis). "Adam" also showed as part of a room name (Adam Ballroom, also not identified in named entity analysis). So a useful enhancement would be to integrate the results of the named entity recognition with the lexicon analysis. The words "spice" (in the context of spicing up promotional information), "bumped" (bumping up a date), "speeds" (speeds up the learning process), "chat" and "chatted" (conversation), and "clarity" (not drug-induced) also tripped the alarm.

There does appear to be some basic NLP processing in the form of stemming. For example:

- “glass” is in the lexicon; “glasses” got a hit
- “uppers” is in the lexicon; “upper” was matched (e.g. Upper Lounge)

As noted, these are not surprising results for keyword searching. It may also be an extreme example, since most terms in the lexicon for this topic are single words. More specific phrases may yield better results, but it does illustrate the more general problem with this approach.

For Personally Identifiable Information, precision based on the default lexicon was again very low. The keyword “ID” may be the most problematic, appearing, for example, in e-mail message IDs and URLs, and transaction IDs in payment receipts. Occurrences of “credit card” do appear, but obviously the context is not taken into consideration – there were no examples found of full credit card number; in most cases the reference was in a receipt or in general information about a company’s services (e.g. a PayPal account confirmation). The regular expression matching of Bulk Extractor and Bulk Reviewer would be a useful addition to simple keyword matching. However, the lexicon search did successfully surface messages including individuals’ dates of birth.

The Employment lexicon involves a similarly high number of false positives. In the test set relating to a community choir, “promotion” is a notable example: the e-mail in the test set discusses promoting concerts, not employees. References to the fire marshall yield similar results. And “recommendation” can obviously come up in many contexts.

Enhancing the lexicon functionality through more sophisticated machine learning models would be worth considering, and it seems that ePADD presents a good example of the opportunity to directly integrate this type of functionality rather than adding new tools into an archives’ workflow.

Indeed, a recent group of case studies about ePADD implementation notes that considerable research relating to individual collections can be needed to customize the lexicons, and identifies “a machine learning model, trained on other messages flagged with particular restrictions” as potential for future research (Schneider *et al*, 2019, p. 322).

It would be interesting to compare the effectiveness and success rates relating to developing and refining training sets as opposed to further developing lists of keywords and phrases. As discussed, a key challenge with the lexicon sets is single words with multiple meanings, along with context being difficult to establish through keyword searching. The latter is a more subtle challenge, but limiting the lexicon to unique keywords and phrases might be a way to improve precision, while potentially decreasing recall, a possibly unavoidable trade-off. A short case-study relating to the development of a lexicon for an academic administrator also notes the challenge of terms that are too broad [23].

A very useful feature in ePADD is the ability to tag messages. This can be applied to individual messages, but more importantly, also in bulk, based on the entity analysis as well as lexicon analysis; for example, to identify restrictions or records that should otherwise be excluded from processing. This seems like a fundamental feature for any NLP application applied to archival processing: making the analysis actionable, not just observable. Terms are also highlighted in each message, making the reasons for a particularly classification clear. However, when an attachment is highlighted, the only way to view the attachment seems to be to download it, in which case the context of relevant terms is not available.

There is certainly potential to extend the tools developed for ePADD to born-digital records more generally, but focus on the e-mail domain facilitates higher quality results (Lee and Woods, 2017) as well as the ability to use e-mail specific metadata, notably e-mail headers. The developers have also indicated a decision to focus limited resources on developing tools that are not already available, so they are also not duplicating tools that might otherwise be useful to integrate into ePADD, such as tools to convert various e-mail formats to the required format. As suggested by Lee (2018), making the named entity recognition module available as a reusable library could be a useful way to extend ePADD's functionality to other record types.

BitCurator NLP

BitCurator is well established as a tool for digital preservation. The BitCurator Environment is now maintained through the BitCurator Consortium; the BitCurator NLP project was a standalone project to develop new tools for natural language processing [24].

Topic modelling

The topic modelling tool works on both disc images and sets of files. Text extraction tools are embedded, but with the current version of the tool it may be more reliable to do the text extraction independently, in order to resolve errors and generate a reasonably clean set of files for topic modelling.

As with the BitCurator Environment, BitCurator NLP incorporates a number of existing open source tools and libraries. For text extraction and NLP tasks, this includes textract, textacy, spaCy, scikit-learn, and GraphLab [25]. The tool is configured and launched through the command line; the topic models are generated by gensim [26]. Once the process is complete, a browser window is launched, and the user can work interactively with the topic visualizations, using pyLDAvis [27]. Goodman (2019) provides a good overview of the functionality.

Pre-processing includes cleanup of numbers and punctuation. There is also a default stop word list, i.e. words that will be excluded from the topics. You can add to that list, particularly to apply institutional or collection context. In the case of university records, for example, common words like student, memorandum, department, committee, meeting, and budget could be added. This needs to be done through the configuration file in a text editor.

A major limitation of the BitCurator topic modelling tool is that users can't currently drill down to the document level, so that appraisal must be done on the whole disk (Goodman, 2019, p. 29). This functionality appears to be possible through the API for the underlying tool, so there is potential for further development.

Named entity recognition

Named entity recognition in BitCurator is less well developed than the topic modelling tool. This is available in two tools, to different degrees.

First, in BitCurator Web Access Tools, an entity view is available at the document level [28]. This provides a visual representation through inline tags on the document. There does not appear to be any way to adjust the results for false positives or missed entities. While this could assist with review of individual documents, without the ability to aggregate this information (and then drill down to documents), the functionality for description, appraisal, or sensitivity review seems limited.

Second, a set of command-line tools is available to identify entities both in disc images and file structures [29]. The functionality includes generation of bar graphs for the entities identified (although a current bug means that phrases are split up). In our testing we were unable to successfully populate the database in order to take advantage of the main features. As documented, there are potentially useful features through a text-based menu, such as the ability to compare pairs of documents, and to export entity lists.

ArchExtract

ArchExtract was developed by the Bancroft Library (University of California Berkeley) in 2015, primarily as a proof of concept relating to NLP for archival processing. It is no longer under development, and due to deprecated dependencies it is now more difficult to launch a working instance, but the functionality developed at the time provide some good ideas for further development, particularly with the Bancroft's focus on non-technical users. These observations are based on published presentations by the project team (Elings 2016, 2017) as well as the author's own testing done in late 2017.

ArchExtract offers pre-processing, named entity recognition, topic modelling, and keyword extraction. (In our own testing, we were unable to get the named entity recognition feature to work, nor were we able to upload a custom stop word list.)

Two important design elements are worth noting: all of ArchExtract's features can be undertaken and configured through the user interface (in this case a web interface), and at least to some extent they are connected and interdependent. For example, one of the topic modelling options is to omit terms that are identified as named entities. More fundamentally, the pre-processing supports the other functionality. As noted in the literature review, cleaning up text prior to processes like topic modelling is a crucial (and non-trivial) step. The topic models are named using the pre-processing configuration, making it easier to iteratively test the best combination of options for a particular data set.

In addition to identifying high-level topic models, the ArchExtract interface also allows the user to drill down to individual documents tagged as part of a given topic. It is also possible to extract the relevant database entries for further analysis and manipulation.

While using a web interface may not be ideal for all data sets (depending on security options), as a proof of concept, ArchExtract provides a very good example of a tool that does not rely on command line access and technical expertise.

Other software

Archivematica [30] and Bulk Reviewer do not currently include NLP features, but are examples of systems with "hooks" for such functionality, either directly or connected through workflows. For example, Archivematica has an appraisal module, and specializes in identification and normalization of

files. Bulk Reviewer is intended to help automate review of records for issues such as personal identifiers. The development roadmap includes future integration of NLP, with specific mention of lexicons similar to those available in ePADD [31]. We will further discuss some considerations relating to integration of different tools below.

Some of the studies highlighted in the literature review also include information about other available tools, both free and commercial. Examples of free software with potential for integration into archival workflows include Weka, which includes algorithms for various data mining tasks [32], and the Topic Modeling Tool, a graphical user interface billed as a "point-and-click tool" for MALLET. In a similar vein, although the documentation focuses on its command line interface, Terrier IR [33] is an open source search engine and text retrieval platform. Open Semantic Search [34], which can be installed as a virtual machine, bundles several tools including named entity recognition, document tagging, data visualization, and a semantic search engine. Other studies have highlighted commercial e-discovery software.

Commercial cloud-based platforms for machine learning would also be worth further exploration; so far there appears to have been relatively little deployment of these platforms in the archival community. Key options include AWS (Amazon) [35], Google [36], and Microsoft Azure [37]. The pay-per-use pricing model, and in some cases free options, could make these services more accessible for archives than traditional commercial services, both for testing and production-level deployment.

Design Principles and Workflow Considerations

Functional requirements and design principles surfaced through the literature and software review include:

- Usable: tools are designed an appropriate level of technical expertise.
- Interoperable: ability to integrate results of NLP processing and machine learning into other pieces of an institution's digital processing workflow; ability to share results, training models, etc. across collections and institutions.
- Flexible: ability to drill down to the document level through the user interface; but also export results for independent processing/analysis, apply visualization tools, etc.
- Iterative: ability to train and refine models, ideally through a user interface, by identifying false positives and missed items.
- Configurable: ability to refine how NLP and machine learning techniques are applied, such as data cleanup options, custom stop word lists, and statistical models employed.

Usable

Usability is obviously always an underlying goal, and a complex topic on its own. A variety of tools have been developed for digital preservation, and many require more advanced technical skills and/or training, along with more general competencies, although there are also continuing efforts to make resources for digital preservation more accessible, such as through the POWRR Professional Development Institutes [38].

However, in the context of tools for appraisal, description, sensitivity review, and other archival processes, it would be worth aiming for more usable tools that could be more readily used by staff specializing in those processes, not necessarily only digital specialists. At the most basic level, developing graphical user interfaces should be a core functional requirement for any such tools to be used in production.

Interoperable

Interoperability is a core requirement for NLP and machine learning tools to be effective, at a few different levels.

There are trade-offs to consider between developing an integrated application, incorporating all the desired functionality, as opposed to a specialized application which becomes part of a suite of several applications. While a single application may have advantages, there are also some necessary assumptions about workflow and use cases. Applications that are more specialized are likely more realistic and more sustainable in terms of software maintenance, and arguably provide more flexibility as well. A middle ground, especially with open source tools, is to bundle available tools – an approach taken with software including BitCurator and Archivematica.

An example in this context is handling the related tasks of file identification and normalization. As noted earlier, these processes are important prerequisites to any NLP or machine learning task. Tools especially for identification and normalization have been well developed, and form an important part of the workflow for digital preservation, independent of any NLP considerations. For example, the PRONOM database forms the basis for tools including DROID, FIDO, and Siegfried, while Archivematica bundles those tools and specializes in integration of normalization tools. It probably makes sense to keep these functions separate, but it would be worth exploring more integrated hand-offs, e.g. through generation of PREMIS metadata. Indeed, files might be exported from Archivematica following normalization (for NLP processing), and subsequently re-integrated into the Archivematica processing workflow. Depending on institutional structures and resources, it may also be possible to have different staff responsible for different parts of this workflow, with NLP tools integrated into workflows for appraisal, description, etc. rather than digital preservation per se.

Text conversion is similarly non-trivial. Cain (2016) outlined a process for handling this if one is starting with PDF files. The BitCurator tools integrate textract, but if there is a failure at that stage, it's difficult to troubleshoot, so we found it more effective to do the text conversion separately.

Another aspect of interoperability is the ability to share entity dictionaries, training models, and similar resources across collections and institutions. Further research and testing is likely needed to determine how effectively training sets transfer between collections. How context-dependent are the training sets? As noted by Payne and Baron (2017): "[A] training set should be representative of the solution space which the algorithm or method will be used in." As we consider options for integrating machine learning into operational workflows, it must be noted that such efforts will be less effective if new models need to be trained for most new collections. Particularly for institutional records, a functional approach might be viable, as ePADD has done to some extent with its lexicons.

Flexible

ArchExtract provides a good example of the ability to drill down to the document level through the user interface. Analysis at the aggregate level is important, but for use particularly for sensitivity review, as well as appraisal, it is necessary to understand what documents have been identified. This also relates to the next design principle, supporting iterative review. That said, no tool could meet all the potential use cases for processing and analysis, so the ability to export results for independent use (and more generally, to access the underlying data) would provide a more robust tool overall.

Iterative

This functional requirement is strongly tied to usability. As identified in several studies, it is crucial to be able to easily interact with the system to refine the models being developed. For example, identifying false positives and missed items in terms of identified documents and extracted entities. This also supports a triage or "funneling" method (as articulated in the UK National Archives study), to use machine learning and NLP methods as an important but not exclusive tool for processing, and in particular to reduce the bulk of records needing manual review. Similarly, Lee (2018) notes an early attempt to design an expert system:

> [Anne Gilliland] was unable to identify a consensus on appraisal rules or principles. This suggests that software to support appraisal should allow archivists to make individual decisions based on iterative feedback, rather than attempting to replace the human decision-maker with software. Software for selection and appraisal can take the form of targeted tools to support specific assessments or decisions, rather than necessarily being full-fledged decision-support systems.

Configurable

To use NLP and machine learning methodologies most effectively, it is important to be able to easily set options applicable to various models and NLP tasks. This includes options for pre-processing data, such as stemming, custom stop words, inclusion or exclusion of entities, and number of topics. ArchExtract provides a good sample. While pre-processing can also be undertaken independently, such as outlined by Cain (2016), integrating these steps into the tool would provide much better flexibility.

Conclusion

As outlined in our review, there have been an increasing number of projects focused on developing tools and methodologies for machine learning and NLP applied to archival processes. Creating tools for wide use now seems more viable, especially with the success of ePADD. Common elements in projects to date and other studies have given rise to basic set of functional requirements that could help inform further progress.

It is also important to recognize where customized development or specialized projects are needed and appropriate. Esteva *et al* (2013) observed that: "Necessarily, for archivists to fully understand the logic and results of this [data mining] methodology implies a learning curve. It requires learning different

computational analysis approaches and working in interdisciplinary teams." However, that approach would allow niche application at best, limited to institutions with suitable resources and expertise.

Applications for processing e-mail have received the most attention so far. It now seems feasible to branch out to develop more generalized tools for born-digital, unstructured records. It is also worth noting that many of the tools described could have broader application. For example, named entity recognition and topic modelling could be used for both appraisal and description. There are also opportunities to run these tools on descriptive records as well as digitized and born-digital archival records: a strategy worth considering in the implementation of linked open data for archives. While commercial tools are more production ready, open source tools have great potential to be further developed and integrated for broader use by the archival community, perhaps in conjunction with cloud-based commodity services. Further research in data and information sciences will continue to improve these tools and introduce new opportunities.

Notes

1. https://textract.readthedocs.io/en/stable/ (accessed 25 February 2020).

2. https://github.com/simsong/bulk_extractor (accessed 25 February 2020).

3. MALLET, http://mallet.cs.umass.edu/index.php (accessed 3 September 2019).

4. The Topic Modeling Tool – GitHub repository, https://github.com/senderle/topic-modeling-tool (accessed 3 September 2019).

5. Fondz GitHub repository, https://github.com/edsu/fondz (accessed 25 August 2019).

6. Always Already Computational: Collections as Data project website, https://collectionsasdata.github.io/ (accessed 6 August 2019).

7. Archives Unleashed project website, https://archivesunleashed.org/ (accessed 20 August 2019).

8. https://read.transkribus.eu/ (accessed 11 February 2020).

9. TOMES project site, https://www.ncdcr.gov/resources/records-management/tomes (accessed 8 September 2019).

10. TOMES software overview, https://github.com/StateArchivesOfNorthCarolina/tomes-project/blob/master/20181127_TOMESsoftwareoverview.pdf (accessed 8 September 2019).

11. The National Archives (UK), PRONOM, https://www.nationalarchives.gov.uk/PRONOM/Default.aspx (accessed 3 February 2020).

12. The National Archives (UK), DROID file format identification tool, https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ (accessed 3 February 2020).

13. FIDO: Open source Format Identification of Digital Objects, http://fido.openpreservation.org/ (accessed 3 February 2020).

14. Siegfried GitHub repository, https://github.com/richardlehane/siegfried (accessed 3 February 2020).

15. "Overview of the Advanced eDiscovery solution in Microsoft 365," available at: https://docs.microsoft.com/en-us/microsoft-365/compliance/overview-ediscovery-20 (accessed 3 February 2020).

16. Cited in Kaczmarek and West (2018). Recommind was acquired by OpenText in 2016 (https://www.opentext.com/products-and-solutions/products/opentext-product-offerings-catalog/rebranded-products/recommind, accessed 3 February 2020), and appears to have been replaced by or rebranded as OpenText Discovery (https://www.opentext.com/info/ediscovery/, accessed 3 February 2020).

17. Ricoh eDiscovery, "Ringtail eDiscovery Software," available at: https://www.ricohediscovery.com/product-sales-fti-ringtail (accessed 3 February 2020).

18. https://luminoso.com/ (accessed 3 February 2020).

19. RATOM project site, http://ratom.web.unc.edu/ (accessed 17 April 2020).

20. ePADD User Guide 7.0, https://library.stanford.edu/projects/epadd/documentation (accessed 18 August 2019).

21. ePADD website, https://library.stanford.edu/projects/epadd/documentation (accessed 18 August 2019).

22. ePADD User Guide 7.0, https://library.stanford.edu/projects/epadd/documentation (accessed 18 August 2019); see also Lexicon User Group, https://library.stanford.edu/projects/epadd/community/lexicon-working-group (accessed 18 August 2019).

23. ePADD Lexicon User Group, https://library.stanford.edu/projects/epadd/community/lexicon-working-group (accessed 18 August 2019).

24. BitCurator NLP wiki, https://github.com/BitCurator/bitcurator-nlp/wiki (accessed 3 September 2019).

25. See BitCurator NLP wiki for details.

26. gensim: Topic modeling for humans, https://radimrehurek.com/gensim/ (accessed 3 February 2020).

27. pyLDAvis GitHub repository, https://github.com/bmabey/pyLDAvis (accessed 3 February 2020).

28. BitCurator Web Access Tools, https://github.com/BitCurator/bitcurator-access/wiki/BitCurator-Access-Webtools (accessed 3 September 2019). The named entity recognition is currently a largely undocumented feature, referenced only on a general GitHub page (http://bitcurator.github.io/, accessed 3 September 2019).

29. BitCurator nlp-entspan, https://github.com/BitCurator/bitcurator-nlp-entspan (accessed 3 September 2019).

30. Archivematica, https://archivematica.org (accessed 3 February 2020).

31. Bulk Reviewer GitHub repository, https://github.com/bulk-reviewer/bulk-reviewer (accessed 3 September 2019).

32. Weka 3: Machine Learning Software in Java, https://www.cs.waikato.ac.nz/ml/weka/ (accessed 3 September 2019).

33. Terrier IR, http://terrier.org/ (accessed 3 February 2020).

34. Open Semantic Search, https://www.opensemanticsearch.org/ (accessed 3 February 2020).

35. https://aws.amazon.com/machine-learning/ (accessed 25 February 2020).

36. https://cloud.google.com/products/ai (accessed 25 February 2020).

37. https://azure.microsoft.com/en-ca/services/machine-learning/ (accessed 25 February 2020).

38. POWRR Professional Development Institutes for Digital Preservation, https://digitalpowrr.niu.edu/ (accessed 8 September 2019).

References

Banerjee, K. and Anderson, M. (2013), "Batch metadata assignment to archival photograph collections using facial recognition software," Code4Lib Journal, Issue 21, July 2013, available at: https://journal.code4lib.org/articles/8486 (accessed 11 February 2020).

Baron, J.R. and Borden, B.B. (2016), "Opening up dark digital archives through the use of analytics to identify sensitive content," *Proceedings of the 2016 IEEE International Conference on Big Data, Washington, DC, 5-8 December 2016*, pp. 3324-3229, available at: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/3.pdf (accessed 24 August 2019).

Bryant, M., Reijnhoudt, L., and Simeonov, B. (2018), "In-place synchronisation of hierarchical archival Descriptions," *Proceedings of the 2018 IEEE International Conference on Big Data, Seattle, WA: 10-13 December 2018*, pp. 2685-2688, available at: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/8.Bryant.pdf (accessed 11 February 2020).

Buckland, M. and Gey, F., "The relationship between recall and precision," *Journal of the American Society for Information Science*, Vol 45 No 1, pp. 12-19, available at: https://doi.org/10.1002/(SICI)1097-4571(199401)45:1%3C12::AID-ASI2%3E3.0.CO;2-L (accessed 25 February 2020).

Cain, J.O. (2016), "Using topic modeling to enhance access to library digital collections," *Journal of Web Librarianship*, Vol 10 No 3, pp. 210-225, available at: https://doi.org/10.1080/19322909.2016.1193455 (accessed 20 August 2019).

Clough, P., Tang, J., Hall, M., and Warner, A. (2011), "Linking archival data to location: a case study at the UK National Archives," *Aslib Proceedings: New Information Perspectives*, Vol 63 No 2/3: pp. 127-147, available at: https://doi.org/10.1108/00012531111135628 (accessed 3 February 2020).

Cormack, G.V. and Grossman, M. (2017), "Navigating imprecision in relevance assessments on the road to total recall: Roger and me," *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, August 2017*, pp. 5–14, available at: https://doi.org/10.1145/3077136.3080812 (accessed 11 February 2020).

Cox, R., Shah, S., Frederick, W., Nelson, T., Thomas, W., Jansen, G., Dibert, N., Kurtz, M., and Marciano, R. (2018), "A case study in creating transparency in using cultural big data: The Legacy of Slavery Project," *Proceedings of the 2018 IEEE International Conference on Big Data, Seattle, WA: 10-13 December 2018*, pp. 2689-2695, available at: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/12.Cox_-2.pdf (accessed 11 February 2020).

Debortoli, S., Müller, O., Junglas, I., and vom Brocke, J. (2016), "Text mining For information systems researchers: An annotated topic modeling tutorial," *Communications of the Association for Information Systems*, Vol 39, pp. 110-135, available at: https://doi.org/10.17705/1CAIS.03907 (accessed 25 February 2020).

Elings, M.W. (2016), "Using NLP to support dynamic arrangement, description, and discovery of born digital collections: The ArchExtract experiment," 26 May 2016, bloggERS, Society of American Archivists, Electronic Records Section, available at:available at: https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/ (accessed 3 September 2019).

Elings, M. (2017), "Using NLP to support dynamic arrangement, description, and discovery of born digital collections: The ArchExtract Experiment," nlp4arc 2017, BitCurator Symposium, University of North Carolina, Chapel Hill, 3 February 2017, available at: https://bitcurator.net/files/2016/12/elings.pdf (accessed 3 September 2019).

Elragal, A. and Päivärinta, T. (2017). "Opening digital archives and collections with emerging data analytics technology: A research agenda," *Tidsskriftet Arkiv*, Vol 8 No 1, available at: https://doi.org/10.7577/ta.1959 (accessed 20 August 2019).

England, C., Prud'homme, P., and Soliday, H. (2019), "Automate it: A deep learning solution for Library Archives," paper presented at 2019 Texas Conference on Digital Libraries (slides), 21-23 May 2019, Austin, Texas, available at: https://hdl.handle.net/2249.1/156417 (accessed 11 February 2020).

Esteva, M., Tang, J.F., Xu, W., and Padmanabhan, K.A. (2013), "Data mining for 'big archives' analysis: A case study," *Proceedings of the American Society for Information Science and Technology*, Vol 50 No 1:

pp. 1-10, available at: https://www.asis.org/asist2013/proceedings/submissions/papers/90paper.pdf (accessed 20 August 2019).

Gollins, T., McDonald, G, Macdonald, C., and Ounis, I. (2014), "On using information retrieval for the selection and sensitivity review of digital public records," *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security, Gold Coast, Australia, 11 July 2014*, available at: http://ceur-ws.org/Vol-1225/pir2014_submission_9.pdf (accessed 20 August 2019).

Goodman, M.M., "'What is on this disk?' An exploration of natural language processing in archival appraisal," Master's paper, Master of Science in Information Science, School of Information and Library Science, University of North Carolina at Chapel Hill, April 2019, available at: https://cdr.lib.unc.edu/downloads/wm117s91s?locale=en (accessed 6 August 2019).

Gracy, K.F. (2015), "Archival description and linked data: a preliminary study of opportunities and implementation challenges," *Archival Science*, Vol 15 No 239, available at: https://doi.org/10.1007/s10502-014-9216-2 (accessed 3 February 2020).

Greenberg, J. (1998), "The applicability of natural language processing (NLP) to archival properties and objectives," *The American Archivist*, Vol 61 No 2, pp. 400-425, available at: https://doi.org/10.17723/aarc.61.2.j3p8200745pj34v6 (accessed 25 August 2019).

Gupta, M. (2018), "A review of named entity recognition (NER) using automatic summarization of resumes," 9 July 2018, available at: https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175 (accessed 25 February 2020).

Higgs, E. (2019), "ml4arc – Machine learning, deep learning, and natural language processing applications in archives", 4 September 2019, bloggERS, Society of American Archivists, Electronic Records Section, available at: https://saaers.wordpress.com/2019/09/04/ml4arc-machine-learning-deep-learning-and-natural-language-processing-applications-in-archives/ (accessed 7 September 2019).

Hutchinson, T. (2018), "Protecting privacy in the archives: Supervised machine learning and born-digital records," *Proceedings of the 2018 IEEE International Conference on Big Data. Seattle, WA: 10-13 December 2018*, pp. 2696-2701, available at: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/5.Hutchinson.pdf (accessed 6 August 2019).

Joens, D. and Kaczmarek, J. (2019), "Processing Capstone email using predictive coding: Semi-annual performance report," NHPRC Project RG-50011-16, available at: https://www.aits.uillinois.edu/services/professional_services/rims/about_rims/projects/processing_capstone_email_using_predictive_coding/ (accessed 3 February 2020).

Kaczmarek, J. and West, B. (2018), "Email preservation at scale: Preliminary findings supporting the use of predictive coding," paper presented at iPres 2018: the 15th International Conference on Digital Preservation, Boston, 27 September 2018, available at: https://osf.io/yau3c/ (accessed 6 August 2019).

Koehrsen, W. (2018), "Beyond accuracy: Precision and recall," Towards Data Science blog, 3 March 2018, available at: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c (accessed 25 February 2020).

Koenig, R. (2019), "NLP for beginners: Cleaning & preprocessing text data," 29 July 2019, available at: https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e306bef0f (accessed 25 February 2020).

Lee, C.A. (2018). "Computer-assisted appraisal and selection of archival materials," *Proceedings of the 2018 IEEE International Conference on Big Data. Seattle, WA: 10-13 December 2018*, available at: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/6.Lee_.pdf (accessed 25 August 2019).

Lee, C.A. and Woods, K. (2017), "Diverse digital collections meet diverse uses: applying natural language processing to born-digital primary sources," paper presented at iPres 2017: the 14th International Conference on Digital Preservation, available at: https://ipres-conference.org/ipres17/ipres2017.jp/wp-content/uploads/50.pdf (accessed 6 August 2019).

Marciano, R., Underwood, W., Hanaee, M., Mullane, C., Singh, A., and Tethong, Z. (2018), "Automating the detection of personally identifiable information (PII) in Japanese-American WWII incarceration camp records," *Proceedings of the 2018 IEEE International Conference on Big Data. Seattle, WA: 10-13 December 2018*, available at: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/2.Marciano.pdf (accessed 25 August 2019).

McDonald G., Macdonald C., Ounis I., and Gollins, T. (2014) "Towards a classifier for digital sensitivity review," in *Advances in Information Retrieval*, *European Conference on Information Retrieval 2014*, Springer, Lecture Notes in Computer Science, Vol. 8416, available at: http://www.dcs.gla.ac.uk/~graham/publications/ecir2014_mcdonald.pdf (accessed 8 September 2019).

McDonald, G., Macdonald, C., and Ounis, I. (2017), "Enhancing sensitivity classification with semantic features using word embeddings," in *Advances in Information Retrieval, 39th European Conference on IR Research, Aberdeen, UK, 8–13 April 2017*, ed. Joemon M. Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Springer, 2017): pp. 450–63, available at: http://eprints.gla.ac.uk/135030/ (accessed 8 September 2019).

McDonald, G., Macdonald, C., and Ounis, I. (2018), "Active learning strategies for technology assisted sensitivity review," in Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (Eds.), *Advances in Information Retrieval*, ECIR 2018 (Lecture Notes in Computer Science, Vol 10772). Springer, Cham, available at: https://doi.org/10.1007/978-3-319-76941-7_33 (accessed 3 February 2020).

McDonald, G., Macdonald, C., and Ounis, I. (2019), "How sensitivity classification effectiveness impacts reviewers in technology-assisted sensitivity review," *CHIIR '19: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, March 2019: pp. 337-341, available at: https://doi.org/10.1145/3295750.3298962 (accessed 3 February 2020)

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill, New York, NY, available at: http://www.cs.cmu.edu/~tom/mlbook.html (accessed 17 April 2020).

Mordell, D. (2019), "Critical questions for archives as (big) data," *Archivaria* 87 (Spring 2019): pp. 140-161, available at: https://muse.jhu.edu/article/724731 (accessed 17 April 2020).

Moss, M., Thomas, D., and Gollins, T. (2018), "The reconfiguration of the archive as data to be mined," *Archivaria* 86 (Fall 2018): pp. 118-151, available at: https://muse.jhu.edu/article/711160 (accessed 17 April 2020).

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., and Varner, S. (2019). "Always Already Computational: Collections as Data – Final report," available at: https://doi.org/10.5281/zenodo.3152935 (accessed 20 August 2019).

Payne, N. and Baron, J. (2017), "Auto-categorization & future access to digital archives," *Proceedings of the 2017 IEEE International Conference on Big Data. Boston, MA: 11-14 December 2017*, available at: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Payne.pdf (accessed 8 September 2019).

Ramanan, D., Baker, S., and Kakade, S. (20070, "Leveraging archival video for building face datasets," *2007 IEEE 11th International Conference on Computer Vision*, *Rio de Janeiro*, pp. 1-8, available at: https://ttic.uchicago.edu/~ramanan/papers/faces.pdf (accessed 11 February 2020).

Rolan, G., Humphries, G., Jeffrey, L., Samaras, L., Antsoupova, T., and Stuart, K. (2018), "More human than human? Artificial intelligence in the archive," *Archives and Manuscripts*, Vol 47 No 2, pp. 179-203.

Schneider, J., Adams, C., DeBauche, S., Echols, R., McKean, C., Moran, J., and Waugh, D. (2019), "Appraising, processing, and providing access to email in contemporary literary archives," *Archives and Manuscripts*, Vol. 47 No. 3: pp. 305-326, available at: https://www.tandfonline.com/doi/full/10.1080/01576895.2019.1622138 (accessed 18 August 2019).

Underwood, T. (2012), "Topic modeling made just simple enough," 7 April 2012, available at: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ (accessed 25 February 2020).

Underwood, T. (2015), "Where to start with text mining," updated 8 June 2015, available at: https://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/ (accessed 25 February 2020).

Underwood, W. (2008), "Recognizing speech acts in presidential e-records," Technical Report ITTL/CSITD 08-03, Georgia Tech Research Institute, http://perpos.gtri.gatech.edu/publications/TR%2008-03.pdf (accessed 20 August 2019).

Underwood, W., Hayslett, M., Isbell, S., Laib, S., Sherrill, S., and Underwood, M. (2009), "Advanced decision support for archival processing of presidential electronic records: Final scientific and technical

report," Technical Report ITTL/CSITD 09-05, Georgia Tech Research Institute, available at: http://perpos.gtri.gatech.edu/publications/TR%2009-05-Final%20Report.pdf (accessed 20 August 2019).

Underwood, W. (2010), "Grammar-based recognition of documentary forms and extraction of metadata," *International Journal of Digital Curation*, Vol 5, No 1, available at: http://perpos.gtri.gatech.edu/publications/149-680-1-PB.pdf (accessed 20 August 2019).

Underwood, W., Marciano, R., and Laib, S. (2017), "Computational curation of a digitized record series of WWII Japanese-American internment," *Proceedings of the 2017 IEEE International Conference on Big Data. Boston, MA: 11-14 December 2017*, pp. 2251-2255, available at: https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Underwood.pdf (accessed 11 February 2020).

United Kingdom National Archives (2016), "The application of technology-assisted review to born-digital records transfer, Inquiries and beyond," March 2016, available at: http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf (accessed 6 August 2019).

Van de Rul, C. (2019), "[NLP] basics: Understanding regular expressions," 30 November 2019, available at: https://towardsdatascience.com/nlp-basics-understanding-regular-expressions-fc7c7746bc70 (accessed 25 February 2020).

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018), "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, Vol 13, No 3, pp. 55-75, available at: https://ieeexplore.ieee.org/abstract/document/8416973 (accessed 25 February 2020).