Morgan M Goodman. “What is on this disk?” An Exploration of Natural Language Processing in Archival Appraisal. A Master’s Paper for the M.S. in IS degree. April, 2019. 45 pages. Advisor: Cal Lee

This paper explores current processes in archival appraisal and selection and investigates the potential uses of automation in the processes. Through an exploration of the BitCurator NLP topic modeling tool, bitcurator-nlp-gentm, I evaluated reactions by participants who agreed to an interview and exploration of the tool. I conclude that topic modeling can assist archivists through identification of like-collections and possible duplication within hybrid collections. Outside of appraisal, topic modeling tools may have uses for archival description and arrangement. Researchers and those with subject matter expertise may also benefit from these tools. This paper points to areas where topic modeling is effective and offers suggestions for making NLP and topic modeling more universally practical in archival workflows.

Headings:

Natural Language Processing

Digital Forensics

Archival Appraisal

Topic Modeling

BitCurator

“WHAT IS ON THIS DISK?” AN EXPLORATION OF NATURAL LANGUAGE PROCESSING IN ARCHIVAL APPRAISAL

by
Morgan M Goodman

A Master’s paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2019

Approved by

____

Christopher A. Lee

Table of Contents

1. Introduction

We live in an era where machine learning and artificial intelligence tools are widely used in a variety of contexts. Natural Language Processing (NLP) is a type of artificial intelligence used in consumer applications such as Siri and Alexa. Natural Language is defined as language used by humans that differs from computer languages, such a computer code. The processing of natural language requires algorithms (made by humans) that can reflect how humans use their language and instruct the machine to interpret data sets.[1] For instance, digital forensics workflows utilize natural language processing to extract targeted information from large data files. Another possible application of natural language processing is within the archival profession. NLP tools have the ability to identify named entities which can help archivists for arrangement and description. The ability to cluster topics can assist archivists in grouping like-collections, and provide a quick glance at the overall content, saving time and improving accuracy.

This paper explores the BitCurator topic modeling tool, bitcurator-nlp-gentm, and its possible application within archival practices. BitCurator is a software environment originally created by a team at the School of Information and Library Science (SILS) at the University of North Carolina in Chapel Hill, and the University of Maryland's Institution of Technologies and Humanities, through two grants from the Andrew W.

[1] This is a paraphrase of information found on Wikipedia. There are some deeper explanations of NLP and its functions there, which I am not covering in this paper. https://en.wikipedia.org/wiki/Natural_language_processing

Mellon Foundation. The goal of the BitCurator project was to incorporate digital forensics tools into existing workflows for collecting institutions; particularly born-digital collection materials. The BitCurator environment is a specialized Ubuntu Linux operating system (Lee et al, 2012). The BitCurator Consortium, which is a community-led membership association, continues to promote the use of these digital forensics tools and provides administrative, user, and community support. The Andrew W. Mellon Foundation has since funded two additional projects: BitCurator Access and BitCurator NLP, both of which focused on better preservation, management, and access to born-digital collections using open-source software.

Born-digital materials are created in a digital form, rather than having been digitized from an analog form. Examples of born-digital items are email, word documents, and data sets. All files stored on a computer have assigned metadata -- for example, timestamps and files names -- that can be easily analyzed by archivists. However, analysis of the contents of the files may first require extraction of the text and then further processing on that text.

BitCurator has a robust set of digital forensics tools currently in use within archival institutions, and in 2018 and 2019 introduced new natural language processing tools. The tools provide a way for archives, libraries, and museums (LAM) to perform NLP tasks within heterogeneous collections. Functions of natural language processing that may be helpful to LAM institutions are named entity recognition, topic modeling, and document summarization (BitCurator GitHub). Within the archival community, bitcurator-nlp-gentm is one of the first tools of its kind and it is open for exploration.

The BitCurator project has demonstrated the value of digital forensics tools within archives and collecting institutions. Libraries, archives, and museums are receiving large amounts of born-digital materials that need to be transferred into more sustainable environments for preservation. They may have files in their collection that need maintenance, or they may receive storage media from donors or other record creators. However, LAMs often do not retain all materials; they must engage in selection and appraisal processes that can be challenging and time-consuming.

In this exploration of the bitcurator-nlp-gentm tool I sat down with users of the BitCurator software and explored ways in which these NLP functions could fit into their current appraisal and selection workflow of born-digital materials. The study is not intended to determine a workflow or rules and procedures in archival appraisal and selection. This exploratory study investigates natural language processing as a tool to support the decision-making process. Through semi-structured interviews and think-aloud sessions, I shadowed representatives from four different collecting institutions as they investigated the bitcurator-nlp-gentm software and discussed its possible uses for their appraisal process. This discussion aims to determine if the NLP software functionality could support their work. A study like this has not previously been done, and the findings will inform the LAM community of possible software support for these processes while also helping the BitCurator team to make adjustments or improvements to the software. It is meant to encourage follow-up investigations into workflows in general as archives move to adopt new NLP technologies.

2. Literature Review

In order to understand what functionalities would support archival appraisal and selection, we must first understand the decision-making processes currently in place. Section 2.1 will walk through a brief summary of archival appraisal. Natural language processing is a natural progression from the standard digital forensic tools. It is important to first understand the nature of an archive's digital collection and digital forensics' role. Section 2.2 will discuss BitCurator and the project's background in digital forensics tools, and how these tools are helpful for archival institutions. Section 2.3 provides information on natural language processing and its place in modern archives. Section 2.4 explains the functions of the BitCurator NLP topic modeling tool and details the system's output, which uses pyLDAvis visualization.

2.1 Summary of Appraisal Literature

According to the Glossary of the Society of American Archivists, appraisal is the "process of determining whether records and other materials have permanent (archival) value," while selection is defined as "identifying materials to be preserved because of their enduring value" (Pearce-Moses, 2005). This is essentially stating that before anything is accepted into a collecting repository, it must first be evaluated. If the acquired materials fit within the institutions collecting policy and proves to have archival value, it will be kept.

Appraisal methods have shifted over the years. Several of the most influential authors[2] in the early English-language archival appraisal literature were Muller, Feith, and Fruin (1898)[3]; Sir Hilary Jenkinson (1922)[4] and Theodore R. Schellenberg (1956).

Schellenberg argues that appraisal is an essential part of the archivists' duties, especially during his time when the volume of records being produced was growing rapidly. In his work, *Modern Archives*, he stresses the need to reduce the amount of bulk by evaluation and selection of the materials. He advocated keeping only the records that held value – either primary value, which holds evidentiary value for the creator, or secondary value, which pertains to the historical and societal function of the record (Schellenberg, 1956).

Gerald F. Ham followed up several years later in agreement with Schellenberg, stating that trying to "preserve everything of possible value…would be irresponsible". He goes on to encourage the development of social archives, state archives, and urban archives to ensure the capture of valuable documents, across various sectors in the form of oral stories, survey data, and ephemera, to name a few (Ham, 1975).

It wasn't until the late 1970's that appraisal of records stored on computers began to be discussed. Charles Dollar wrote a piece for the American Archivist in 1978 on appraisal of Machine-Readable Records (MARC records), and stated: "it is likely that

[2] There are many works authored by these men, but I am referring to specific items for which I am including the citation in the footnotes.

[3] Muller, Samuel; Fruin, R.; Feith, Johan Adriaan (1940). Manual for the arrangement and description of archives: drawn up by direction of the Netherlands Association of Archivists. New York: The H. W. Wilson company

[4] Jenkinson, Hilary. A Manual of Archival Administration. London: Percy Lund, Humphries & Co., 1966

current practices and standards will be obsolete and irrelevant within a decade." In this paper he discusses the appraisal considerations of these records: can the item be read? If it can be read what are the "legal, evidential, and informational value of the records" (Dollar, 1978)? Validation of the data must be determined by the archivist, who should also consider costs for preservation.

In the 1990's several influencers of current archival thinking made their voices heard. Frank Boles and Julia Mark Young worked to establish a structure to the appraisal of records. They created a model that based the appraisal decision on three criteria: value of information, cost of retention, and implications of the appraisal recommendations (i.e. political considerations and procedural precedents) (Boles, Young, 1991).

Helen Samuels suggested that better appraisal decisions would be made if they were based on an understanding of the "phenomenon or institution" that is being documented. She suggested a shift toward better understanding and documentation of institutions rather than attempting to predict future research (Samuels, 1992). Greene and Daniels-Howell believed appraisal decisions should be based on the repository's specific goals and resources. This strategy is called the "Minnesota Method", and instructs repositories to do an analysis of their current collections as well as the institutions from which they collect, to produce "rational and efficient" selection criteria (Greene, Daniels-Howell, 1997). The Minnesota Method is "a strategy for appraising materials that combines aspects of collection analysis, documentation strategy, appraisal, and functional analysis" (Pearce-Moses, 2005).

What we have seen is a fluidity in archival appraisal across time, and yet a lack of consistency in practice. Ann Gilliland's 1995 dissertation highlights the apparent lack of consensus on appraisal rules and principles at the time of writing. Today, archives are inundated with digital materials. Since the introduction of electronic records in archives there have been attempts to manage their appraisal. As described by Tyler O. Walters, “instead of producing distinct series of records from a particular function of an organizational unit, electronic records can result from broader information systems reflecting many functions which may even cut across several organizations” adding a whole new layer to the appraisal process (Walters, 1996). Walters suggests that part of the necessity for Terry Cook to develop his macro-appraisal theory and strategy was to “cope with the appraisal of electronic records” (Walters, 1996).

While we have not seen any implementation of software to support archivists’ decision-making for appraisal and selection of digital records, there have been attempts to envision the use of such software. Harvey and Thompson’s 2010 paper on automating the appraisal of digital records argues that if humans continue to use manual methods for appraisal and technical re-appraisal it will be “unsustainable and unproductive in the long term” (Harvey et al. 2010). When the paper was written in 2010, automated appraisal was just an idea. They point out that, while it is a good idea and worth pursuing, there are obstacles. They explain that the “cost of designing, developing and implementing [automated technology] may be prohibitive in some situations” (Harvey et al, 2010). One of the biggest issues hindering automation in archival appraisal is the lack of standard methods in this area.

2.2 BitCurator and Digital Forensics in Archives

Understanding natural language processing as it is used by BitCurator requires an understanding of Digital Forensics. Digital Forensics is a term that is often associated with criminal investigations and cybersecurity, and law enforcement agencies have been using digital forensics tools and methods for many years. There are similarities between the work done in criminal investigations and archival workflows. Archivists are tasked with investigating the contents of any donated or acquired collection. Archivists cannot get access to digital information "without mediation through complex instrumentation or layers of interpretative software" (Kirschenbaum et al, 2010). Digital forensics tools can help them examine those layers. In a criminal investigation, digital forensics aims to uncover the data about the person who owns the media (floppy disks, hard drives, or whole computers) to gain knowledge about that person's actions and produce evidence. Some of the traces left on media may be unknown to the creators or original users of electronic records. For example, the content of deleted files is not immediately overwritten; the location on the storage disk is simply marked as 'available' for re-write. This means that deleted files have the potential to be uncovered and restored.

The idea of implementing these same practices in archives is not a new. There have been discussions about utilizing digital forensics tools since the late 1990's. Seamus Ross and Ann Gow's 1999 paper *Digital Archaeology: Rescuing Neglected and Damaged Data Resources* emphasized digital forensics as a tool for recovering data from media that is obsolete or becoming obsolete. Modern users of Digital Forensics in law enforcement are typically working with the latest technologies, such as mobile devices and computers with modern Operating Systems, while archives must often deal with older media that are quickly becoming obsolete. Kirschenbaum et al. (2010) touch on this in their report based on discussions from the Maryland symposium titled *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, suggesting that digital forensics is important for Archivists not only to recover data from storage media, but also to make sense of it. In other words, they must engage in both “data recovery and data intelligibility” (Kirschenbaum, 2010).

One implementation of digital forensics tools in archives is disk imaging. As Christopher (Cal) Lee describes, disk images often should be treated as “basic units of acquisition” (Lee, 2014). Disk images capture much more than a traditional file copy because digital files have a multitude of storage sectors. Disk imaging captures the information on all of those sectors, along with associated metadata in the filesystem, not just the content of the file as a bitstream. This helps archivists ensure the value of the record through verification of provenance, original order and chain of custody (Lee, 2014). Additionally, archivists are able to use digital forensics tools to parse out personal identifying information (PII). Responsible access to records is a primary task of the archivist, and running a disk image ensures that personal or sensitive information is properly handled before access is granted. In addition to disk imaging, digital forensics tools can also help to locate PII.

Martin Gengenbach’s 2012 Master’s Paper explored which digital forensics tools were currently in use within archives, and how they were implemented into their workflows. Archivists perform a variety of tasks, including creating authentic copies of materials, identifying original order, identifying sensitive information, and exporting contents of the disks for inclusion in Archival Information Packages (AIP) and

Dissemination Information Packages (DIP). Prior to the BitCurator environment, archivists completed these tasks using tools such as Fiwalk to extract metadata from disk images, and the Forensic ToolKit (FTK) which is a more robust processing tool used for arrangement and description.

Lee points out that traditional digital forensics tools were not built with archival functions in mind. This was the intention of the BitCurator Project when the project began. The BitCurator Project combines a set of digital forensics tools in a package that compliments the workflows of collecting institutions. Lee states that the project intended to "bridge this gap through engagement with digital forensics, library and archives professionals, as well as dissemination of tools and documentation that are appropriate to the needs of memory institutions" (Lee, 2014). Through the creation of the open-source software, BitCurator has incorporated digital forensics into archival institutions. The features of BitCurator included:

- Pre-imaging data triage
- Forensic disk imaging
- File system analysis and reporting
- Identification of private and individually identifying information
- Export of technical and other metadata (BitCurator.net).

This robust functionality has encouraged archives to implement the software and it is becoming widely used in archival practice. However, one tool that has not yet been widely adopted in archives is natural language processing.

2.3 Natural Language Processing

2.3.1 What is NLP?

As earlier defined, natural language processing is the computer programs ability to understand human-created data sets and make sense of them in a way that a human would. Jane Greenberg uses Tamas E. Doszkocs' definition of natural language processing: "… intelligent analysis, understanding and expression of 'meaning' as exemplified in natural language…". NLP is divided into basic, rudimentary AI, and full AI. Basic NLP is the ability to search within the text of a document. While more advanced NLP can do an array of things such as syntactic and semantic processing (Greenberg, 1998). Two of the main functions of BitCurator NLP are: named entity identification and topic modeling.

NLP software can identify a variety of entity types, including Persons, Places, and Organizations. As one can imagine, these elements appear frequently throughout written documents. By identifying these entities, the content is given a structural framework for further discovery. Topic modeling is a method of providing insight into how concepts are naturally clustered. Topics are groupings of words that are likely to appear in the same document, and through the process of topic modeling, patterns arise that show the main topics that organize a collection. This paper will solely focus on the topic modeling function.

2.3.2 NLP Tools in Archives

BitCurator is not the first instance of a natural language processing tool used in an archival setting. In 2015 ePADD was released by Standford University's Special Collections and University Archives. It is an open-source tool developed to support the “appraisal, processing, preservation, discovery, and delivery of historical email archives” (Stanford Libraries). It uses natural language processing to search the corpus by named entities and allows users to search by person, organization, or location (Schneider, 2016).

Other topic modeling tools have been developed and discontinued over the last several years: ArchExtracts is a tool that was used at the University of California, Berkley for archival arrangement and description (Hutchinson, 2017) but is no longer being maintained. This application used MALLET which can produce up to 100 topics. Another experimental tool, called Fondz, was intended for archival description. It used natural language processing to auto-generate an archival description from a bag or series of bags. This tool is also no longer in development (Hutchinson, 2017).

Until now digital forensics tools supporting archival appraisal have been underrepresented. As these limited uses of NLP have shown, few applications have yet targeted appraisal within an archival setting. This paper explores the possible uses of BitCurator NLP and how it might assist archives in their appraisal workflows.

2.4 BitCurator NLP

The BitCurator NLP project has adopted several open-source tools: Textract, Textacy, SpaCy, Scikit-learn, and GraphLab. In order to run NLP tools on the contents of files, one must first extract those contents as text. Textract is a tool to extract text from a variety of file types (e.g. Microsoft Word, PowerPoint, PDF) (Textract.readthedocs.io). This is essential because archives may receive multiple born-digital record formats as part of one collection. Textacy is a python library that performs NLP tasks. It uses the SpaCy library, which is capable of processing large “dumps” of information and performing extraction tasks. (Textacy GitHub). The BitCurator NLP has used SpaCy because it has "good pre-trained models for entity and item recognition" (BitCurator). It is relatively simple and integrates easily with machine learning platforms. Scikit-learn is a machine-learning tool, also in python, that performs classification, clustering, and model selection (scikit-learn). Classification takes the extracted data and determines which category these identified entities belong to. It then clusters them into similar groups. The visualization output, pyLDAvis, is the final step which graphically presents the modeled topics in a standard web browser. See figure 1 below for a chart of the bitcurator-nlp-gentm flow.

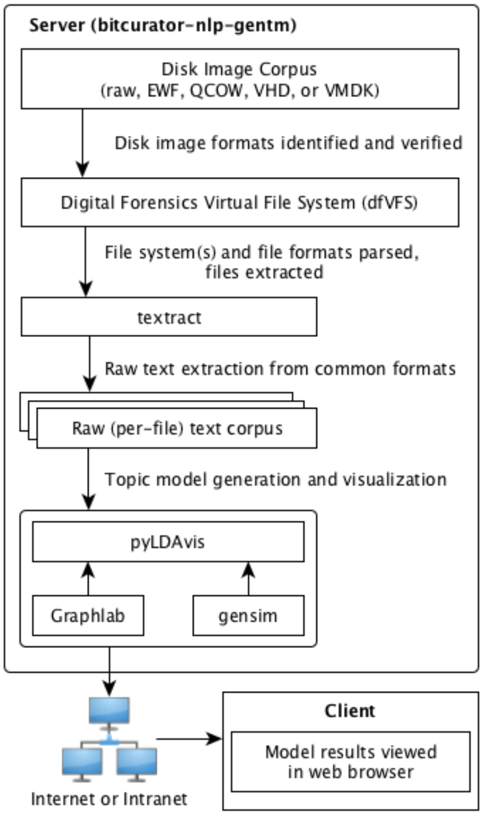
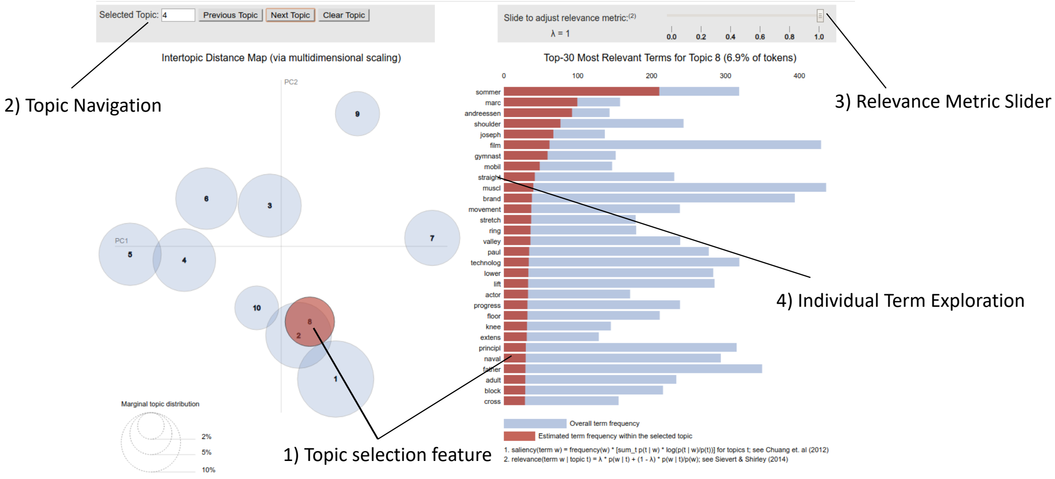

Server (bitcurator-nlp-gentm)
Disk Image Corpus
(raw, EWF, QCOW, VHD, or VMDK)
Disk image formats identified and verified
Digital Forensics Virtual File System (dfVFS)
File system(s) and file formats parsed,
files extracted
textract
Raw text extraction from common formats
Raw (per-file) text corpus
Topic model generation and visualization
pyLDAvis
Graphlab
gensim
Internet or Intranet
Client
Model results viewed
in web browser

Figure 1. Text extraction process though use of pyLDAvis (Github,BitCurator NLP)

Archives rely heavily on human interpretation of documents because, until now, only humans have been able to detect the nuances within the contents of the documents. BitCurator NLP uses topic modeling to determine the meaning of the data that is extracted. However, without the inclusion of SpaCy and scikit-learn, the software would not be able to accurately interpret and organize the information. Lastly, pyLDAvis, a python adaptation of the original LDAvis, converts the data into a single page visualization with navigation for use by the archivist. With the introduction of BitCurator NLP, the aim is to automate some of the time-consuming processes that would normally have been done by humans.

2.4.1 pyLDAvis

To understand the features and functions of the bitcurator-nlp-gentm the output of the visualizations should be briefly described. The visualization consists of four main features for manipulation by the user: 1) topic selection, 2) topic navigation, 3) relevance metric slider, and 4) individual term exploration. Refer to figure 2 for an annotated image highlighting these features.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

FIGURE 2 Example of Pyldavis Output (Angelov, 2018)

The left side view of the visualization has ten topic circles displayed in relative sizes and plotted on a two-dimensional plane. While ten topics is the default, the number of topics can be changed in the configuration settings. Each topic is labeled with a number in decreasing order of prevalence. The size of the circle depends on the prevalence of the top 30 terms extracted from the text. The space between topics is determined by the relation of the topics, i.e. topics that are less related are plotted further away from each other. Clicking on a topic circle will alter the right-side view of the terms list by adding a red bar for each term. The red bar indicates the frequency at which a term appears in that particular topic in relation to the overall corpus.

Grouping related terms is one of the main features of this visualization. However, this alone will not help the user to determine what the topic is about. The Metric Relevance Slider is a complex feature meant to assist in deciphering topics. Essentially, when the slider is all the way to the right at "1", the most prevalent terms for the entire corpus are listed. When a topic is selected and the slider is at 1, it shows all the prevalent terms for the corpus in relation to that particular topic. Alternatively, the closer to 0 the slider moves, the less corpus-wide terms appear and the more topic specific terms are displayed. The relationship between the two is meant to assist the user in deciphering the topic. The ideal setting for the slider may vary depending on the content, although a study done by Sievert and Shirley in the paper *"LDAvis: A method for Visualizing and Interpreting Topics*" determined that it has the best results at about .6.

The navigation feature lets the user move from one topic to the next or clear the topic. Another helpful feature is the individual term view. When hovering over any term in the term list the visualization of the circles will change to show where the term appears across topics. This may increase the size of a topic circle or drop a circle from the view entirely if that term is not present in the topic. This is the only time that all ten circles may not appear on the screen. More information can be found by reading the aforementioned paper by Sievert and Shirley.

3. Methods

This study aims to answer the following questions:

1) Could the BitCurator NLP software be useful for archivists in their appraisal/selection process for born-digital materials?
2) What functions does NLP have for archival selection/appraisal in an institution that relies heavily on human analysis of materials?

A qualitative approach was used to collect data in this study. Each session consisted of an unstructured interview and a brief think-aloud session while using the BitCurator NLP software application. For the purposes of this study I retained a neutral stance toward the BitCurator NLP tool. This approach ensured that my own biases did not influence the participants as they explored the tool and their reactions could be accurately gauged. To prepare for the session I learned how to use the tool and created a set of instructions on how to install the application, which was provided to the participants before the meeting. My general questions can be found in the appendix.

The following sections detail how this study was conducted: 3.1 includes information about the sample, 3.2 describes the process by which I collected and analyzed the data, and 3.3 addresses the limitations of the study and ethical issues.

3.1 Sample

The sample set for this study is a convenience sample of four institutions in the Research Triangle, North Carolina. These institutions were selected because they are members of the BitCurator Consortium and are therefore familiar with the process of creating disk images, and would likely have images to work with in the study.

All four institutions agreed to participate and there was one session per institution. Three of the four sessions had two participants, resulting in a total of seven participants across the four sessions. The institutions are: The State Archives of North Carolina, Duke University, North Carolina State University, and Wilson Library at the University of North Carolina, Chapel Hill. The participants are anonymized and will be referenced by number moving forward (i.e.: Participant 1 (P1); Participant 2 (P2)). None of the participants were familiar with the pyLDAvis output prior to this study, nor had they used the bitcurator-nlp-gentm tool.

3.2 Data Collection and Analysis

I recruited the participants by email and after receiving confirmation from the institutions, a meeting time was arranged. The sessions were conducted on site at each of the institutions where they had installed the NLP tool and prepared a set of disk images to be viewed. It should be noted that in some sessions the sample set of data was used instead of the institution’s own collection due to technical issues that were unable to be resolved. The sample set is a preconfigured sample Expert Witness Format disk image from the bitcurator-nlp-gentm repository and, unless reconfigured, the tool will use the files extracted from this disk for the topic modeling and visualization (GitHub, Bitcurator-nlp-gentm).

Before the session began, I received verbal approval from the participants to record the audio of the session on an iPhone using the voice memo application. Each session consisted of a semi-structured interview and a think aloud session. To begin the semi-structured interview, I posed questions regarding their current appraisal processes and the types of born-digital materials they encounter. After the initial interview we began using bitcurator-nlp-gentm. I briefly explained what they were seeing in the visualization and the process by which the output was created. Throughout the session we discussed their impression of the tool and how the tool's features might fit into their appraisal processes as well as the overall appraisal workflow. These exchanges were intended to identify functions of the NLP software that are useful in appraisal, determine if there is functionality that is missing, and what other applications these features might have for archival processes. Each session lasted approximately one hour.

For my data analysis I transcribed the audio recording from each session. Once transcribed I built a matrix that connected each of my questions to the appropriate answers and labeled the answers by participant. This helped structure my findings and see all of the answers together in one place to make comparisons. I also created some additional sections to the matrix that were reserved for unanticipated themes observed during the conversations.

3.3 Limitations / Ethical Concerns

There are limitations in this study which stem from the small sample size. Another limitation is a lack of diversity amongst the type of institutions. Three of the institutions are academic and one is a government archive. Institutions such as museums, libraries, and smaller community archives are not represented.

Think-aloud sessions also come with their own limitations. Results rely on rich participation by the subject. Subjects may feel insecure sharing all of their thoughts, and their thoughts may contain biases. There was also a constraint on time as each session was limited to approximately one hour which included setup time. Therefore, some topics were not discussed to the extent they may have been had the session duration been longer.

There are no special ethical concerns in this study. Participation in the session was voluntary, and participants were not persuaded to adopt the software permanently. It is up to them what they do with the tool after the conclusion of the session.

4. Findings

Each interview began with questions about the appraisal process and the questions that are asked by the archivist when making selections for collections. The participants explained their process for acquisition of born-digital materials and the steps of their appraisal processes. The findings in this paper are structured using direct quotes from the participants, referenced as Participant 1 (P1); Participant 2 (P2), etc. Section 4.1 details the findings about what types of materials are collected and where in the processes the appraisal decisions are made. Section 4.2 reveals the answers the participants provided when asked if they have any experience using NLP within their archives and what they hope NLP can do for them.

Sections 4.3 reviews the features and determines if they are able to help the archivists answer their appraisal questions in their decision-making process. In an analysis of the data collected from the interview sessions and my observations, each feature of the pyLDAvis visualization was assessed. The visualization itself does not provide the user with an explanation of how to manipulate the visualization. It is up to the user to click, hover, and slide to get an idea of what the underlying algorithm has grouped as topics. Section 4.4 reveals what features are missing or would make the tool better. It should be noted that the discussions around the tool focus primarily on appraisal; however, the primary uses of this tool are likely to be in other areas such as arrangement and description. Section 4.5 discloses the participants thoughts on where this tool might fit in appraisal, but the flow of conversation also exposed other possible areas of use. All of these are discussed in this section. Finally, section 4.6 discusses the implications of the study.

4.1 The Born-Digital Appraisal Process

Each institution represented in this study is unique in the nature and contents of its collections, with over 30 collecting areas across the four institutions. Naturally, they vary in their appraisal processes and the decisions they make, but all have several phases of their process in common. When asked about their steps in the process these are the findings from their responses:

First, acquisition of materials depends on their source. The accessioning of the materials depends on both the donor and if the materials fit with the institutions collection policy. Curators or an acquisition team will take into consideration the source, and the decision to acquire an item might happen before the materials hit the desk of the archivist. Of course, the first question is: what is it?

> *"Basically, we are looking for things that are records and fall in line with what is described on the schedule." (P1).*
>
> *"What is its relationship to the collection that we were given? Were we expecting it from the donor? Based on what they told us they were giving us, does this thing fit?" (P3).*
>
> *"Do we want the material from this particular donor? If they meet the criteria [the curators] transfer them to us." (P4).*

Knowing who the donor is and the possible nature of the material may lead the archivist to determine if further investigation is necessary. The value of the material for the collection might easily be judged based on the donor alone. Usually when the archives are in communication with a donor, the material the donor is providing is expected in advance. In this case, the archivists may decide to keep all of the media without doing a deep dive into the content. *"It's more like a macro sort of appraisal. I think that part of the decision is made by what we can figure out might be on those disks" (P4).*

At this crossroad the appraiser may do further investigation or they may decide to keep the materials based on the information they already have. With that said, deciding to keep all media does not mean the appraisal process is over. One participant explains, "*There is also appraisal that happens as part of processing, you know as far as weeding and things like that (P1).* There may be items included that are not needed by the archives which get quickly removed.

> *"Sometimes when we get the entire hard drive or something, things like iTunes libraries or photos that they had saved on their computer are immediate don't-need-to-keep types of things" (P2).*

Further investigation of the material often depends on the media type. Some media types are easier to work with, such as optical discs which can easily be burned, and thumb drives which allow for easy mobility of files. The type of media influences how the item will be appraised. Some types of media mentioned were:

> *"…3.5" floppies, 5.25" floppies, optical, internal hard drives…" (P6)*
>
> *"It's kind of a mix of things but I think the majority are still, you know, floppy disks that we find in the collection" (P3)*

Floppy disks were mentioned frequently as an example in the conversation, possibly because these are common and may require some extra examination by the archivist. That is, of course, unless the archivist is relying on the label (if it has one at all).

> *"Most floppy disks are going to have very idiosyncratic notes and labels. But something that made sense to the creator is probably not going to make sense to*

> *us. Maybe they are very good at note taking… but the assumption is we are not going to know what's on the disk until we do a disk image."(P6)*

> *"Well, do we just rely on the label on the disk? How reliable is that? That's one decision point that would be tricky. Do we really want to commit to this whole workflow of extracting all of this?"(P3)*

Committing to an extensive workflow to find what is on a floppy disk is part of an appraisal decision made by the archivist. If it is important enough to know what is on the disk, there is justification in going through the process. Once they have begun to investigate there are two or three main things that they are looking for. *"We primarily look for viruses, privacy and duplicates. Duplicates are something that we come across fairly frequently" (P4)*. Born-digital materials are often part of hybrid collections. A question that a majority of the archivists want answered is: how much duplication is there between what is on this media and what is in the analog collection?

> *"In those cases, I think when we start looking at the born-digital we are wondering how much duplication is there between the paper and the born-digital? Is the born-digital all the drafts and then the paper is the printed final version? Do we want both?" (P3)*

> *"We get duplicates. Lots and lots of duplicates"(P2).*

4.2 Experience with NLP and Hopes for NLP

Throughout the conversations about their workflow processes, I wanted to know if they had used natural language processing in any other areas. Most of the participants have some experience in NLP through different archival avenues, such as description and arrangement. One had been using NLP to *"process emails and tag entities such as organizations and personal identifying information"(P2).* Other tools were mentioned such as ePADD and Semantic Search. These tools have assisted the archivists in finding specific terms within a collection, locating sensitive information, or picking out patterns across folders in a collection. It is unclear if these tools have been permanently adopted by the archivists as a consistent part of the workflow. *"We've done a little bit with EPADD in email… But we have not figured out what about that experience was worth retaining" (P6).*

Archivists are also considering how researchers might be able to use natural language processing tools. This is a theme that emerged that I will discuss later.

> *"We're exploring some programs that allow semantic search using NLP in various ways for researchers to access large digital collections with enhanced search capabilities than what they would just get on the operating system. So, we're doing some user testing for a program right now" (P5).*
>
> *"We had one researcher who used ePADD in the reading room. Otherwise she would have spent probably five days looking at the same thing" (P7).*

Before moving on to test the tool, I asked what they hope NLP can do for them. The answers were strikingly similar. The participants hoped to have a tool to look over a large data set and make sense of it, and alleviate the task of manually examining each file.

> *"Obviously there's human intervention as far as what to do with it, but getting [an automated way of looking at things] rather than having to look through things manually would be great." (P1).*
>
> *"I think we're hoping for anything that can help us find what we don't want to keep. Being able to search across such a large body of information is one of the things that's the best about born-digital collections because you can't do with analog." (P3).*
>
> *"Being able to get a sense of the material without having to look at every single file is something that would be useful for us"(P6).*
>
> *"I think that the generation of topics could help in instances of hybrid collections." (P4).*
>
> *It'd be great if we can figure out what is across a set of either disks or one large volume. Being able to target what parts of a volume might be worth retaining*

> *versus the entire set and not have to do further work on what we don't want would save us some time" (P6).*
>
> *"We have a bunch of regular expressions that look for FERPA type of data or HIPPA data that would require a human to look at the report and be like 'is this instance of the abbreviate GPA an innocuous instance of that or is it something that we have to actually have to put on a different server?' That slows us down. So yeah, if we can use NLP to get actual reports that are more useful than a spreadsheet of hits that might be useful." (P7).*

Saving time and unnecessary work is the primary benefit the participants are seeking in implementing NLP into their workflows. The limited applications of NLP currently in use accomplish some of these wish-list items, but there is certainly more than can be done.

4.3 Exploration of bitcurator-nlp-gentm Features

Based on the interviews, archivists are looking to answer two general questions while performing the appraisal process: 1) what is on this disk? And 2) is this something that provides value that we want to keep? The key aim of this study is to explore whether or not the features provided by the tool assist in answering these questions.

After providing a brief explanation of the output, I observed the participants as they moved through the features. Multiple topic selection and changing between topics were rare; they would select one topic to explore and would largely ignore the others. This could be due to the limited time frame available for exploration, but regardless, attention was generally directed toward other features. Overall, the initial exploration of the clustering of topics in the visualization returned these comments:

> *"The separation into topics doesn't make as much logical sense to me" (P5).*
>
> *"... I mean, I feel like in an appraisal situation where I don't know what's on the disk it's hard for me to know what's on the disk just from looking at these topics"(P4).*

> *"I am not sure how this would help me analyze the collection if I didn't already know about the collection" (P3).*

The Metric Slider was the most-used feature throughout the session. The functionality of the slider is not immediately intuitive, so the participants had a lot of questions about how it worked. I provided a high-level explanation and occasionally pointed the participant to the article explaining the probabilistic mathematical functions for which the slider creation was based. The slider got mixed reactions. Participants who were more comfortable with the feature liked playing around with it. *"I do like the slider a lot. If I am looking at topics I am more interested in the unique words in these topics rather than knowing that 'Birds' is in all of them" (P5).* For others, the slider seemed to be a bit cumbersome, even after using it for a few minutes. *"I still don't quite get the slide, I have to be honest with you" (P1).*

Term frequency and the slider feature go hand in hand. Term frequency was the element that participants were most interested in, as some participants use term frequency regularly in their analysis of documents. The slider provides a novel method of viewing term frequency; the terms are weighted and the movement of the slider adjusts the weights and therefore adjusts terms in the view. Even with this general understanding, there were still questions about how to analyze the view: *"…it kind of looks like the similar ones are staying at the top. Which maybe are the most important terms in that topic?"(P2)*

The slider did succeed in showing the uniqueness of each topic. In some instances, the top 30 most prevalent terms for the entire corpus fell away, leaving a few terms that appeared solely in that topic. However, after further exploration of the topics, the participants remained unable to define precisely what each topic represented.

Using the topic clustering and term frequency features, the participants were able to make certain assumptions. However, by adding all of the possible views enabled by the slider, interpretation of the topics became more difficult. It was easier for the user to make a connection between four or five of the words, and ignore the rest. Some participants expressed concern about being able to accurately conclude what the topic is about without the topics being clearly labeled for them.

> *"I always struggle with labeling the algorithmic output of anything because it's going to be highly affected by what my bias already is, or what I am expecting to see on this disk" (P5).*

4.4 What is missing?

The participants were unanimous in their desire for a feature not currently present in the tool: a way to identify which individual files pertain to which topics. The ability to look across a corpus and get an overall idea of the topics within it was deemed useful, but limited. Without a way to locate where the topics reside within the corpus, archivists are unable pick the items to keep or to remove. Instead, the archivist can determine if they want to keep the whole disk or divest of it.

> *"...if you are using it for appraisal or processing then you need to be able to act on it. You need to be able to go locate those things." (P1)*
>
> *"The one thing that is missing, from the appraisal point of view is being able to get back to those items that are on the disk images. To know what files actually relate to what you're seeing here in terms of prevalent topics because my goal would also be to get back to those items and be able to keep or not keep them" (P3).*
>
> *"The interface doesn't support 'I want to see what the files are from topic 1' click on those, and then take you to the files. So, there is no interaction there at all" (P7).*

Additionally, participants were concerned about the possible uneven distribution of the topics across the files. There may be one enormous file pertaining to a topic that does not have relevance to the collection, but the term frequency of that topic overshadows the more relevant topics.

> *"It would be cool, even if you have a group of stuff, to be able to see that this word appears this many times across this group, but then knowing 'is it just in one file?'" (P2).*
>
> *"I could imagine a case where you've got a giant pile of office memos or something like that and two thirds of them relate to some controversy that you are interested in and the other ones don't. How do I get a sense of that? I don't know if that would be clear from something like this" (P6).*
>
> *"If I click on topic 2 here, it would be nice to know that 20% of the files that are represented in here relate to that topic" (P7).*

There was also one other suggestion about how to improve upon the tool:

> *"I think it would benefit from having multiple modes of communicating this information. A trend graph especially would be useful. If there is some way to extract 'date' for example and then have a trend graph of these topics over time that would be useful, I think"(P5).*

4.5 Possible Applications

As previously mentioned, it is important to note that appraisal is not the only possible application for NLP and topic modeling tools. This study primarily focuses on how topic modeling might assist in the appraisal process and does not exhaust all likely uses of the software.

4.5.1 Appraisal

The use of a topic modeling tool for born-digital materials in appraisal would depend on the collection. Considerations may include the size of the collection, the media, the donor, and the return on investment for committing materials to an intensive workflow. Extracting files can be a time-consuming process, so if a decision can be made about whether or not to accept the media as is, that would be most beneficial.

> *"We want to be able to make some decision about the media without having to go through the whole process of the workflow to get stuff off of the media. Like, do we really want to commit to this whole workflow of extracting all of this and then find that none of this is useful? But maybe that's unavoidable and we just have to take the time to do it"(P3).*

> *"We don't really have a holding pin for files that have come off media that we're not going to proceed with. We're not going to run through the whole workflow with this content... So, for me this is gonna be like 15 MB of data with some reports or it's gonna be nothing" (P4).*

One reason an archivist may not image a floppy disk is because the disk is labeled and they assume that the label is correct. This is part of an appraisal process that would limit the amount of time and energy examining each individual file. Where the NLP topic modeling tool would come into play is in the scenario where there is effectively no way to know what is contained in the media. At this point the archivist has determined that the media cannot be appraised without making a disk image. The donor or creator will also influence whether or not the media is investigated at this point. If the donor is not considered "high-profile" or a higher priority, these mystery disks may take a back seat to more pressing appraisal matters. It is unlikely that a low-priority disk would be imaged unless there was an immediate need.

> *"I think if we were to use something like this in processing it would be on collections where we had just a ton of information that we needed to figure out how to triage in some way. It would be a pretty high value collection" (P6).*

This topic modeling tool can be useful when trying to parse out topics that seem related, but the archivists does not know the subject matter very well. Being able to determine that some topics seem related across multiple disk images can be a positive application for this tool.

> *"If we had multiple disk images from one collection it might actually be really useful…we have 10 disk images, we look at them all and we see how related are they. Are they similar types of content or are they very different? That might help us prioritize what to keep" (P3).*

Using a topic modeling tool for this type of exploration could save time, especially when considering the effort involved with looking through ten different disk images one-by-one and trying to make connections between them.

> *"If we're thinking about a workflow perspective, it would be quicker to do this with a whole bunch of disk images than to mount each one or to export text files for each one" (P7).*
>
> *"I could see if we had no idea what was on a set of disks, topic modeling having some relevance for trying to figure out what things might be. If we had 50 floppies and group them by...saying these disks probably are related to X… or these disks are more related to Y" (P6).*
>
> *"If [the curator] said 'I still don't know what this stuff is', then I could see doing this. But it feels like that would be a very special sort of use case because it's that much more work to do" (P4).*

It would be a unique scenario to move ahead imaging the disk and looking at the topics to decide if the whole thing would be kept. One participant succinctly described it, *"it would be kind of a tool in search of a use case rather than a use case in search of a tool"(P7).*

4.5.2 Research

One question that arose out of the sessions was whether it would be more beneficial for the researcher to use the tool instead of the archivist? Some archivists expressed uncertainly over whether it is the researchers' role to figure out what is on a disk. Researchers will often have very specific questions about a collection and it is not necessarily the archivists' job to be a subject matter expert on any given topic of research. *"...this kind of blurs the line between what are we going to do and what do we expect patrons to do"(P7).* Another participant commented, "*...from a research point of view, this is exactly the thing that would help with research" (P3).*

Perhaps the topic modeling tool could be run after a query from a researcher.

> *"If it's a clear question like 'I'm really interested in this authors archives that you have and I want to know if they corresponded with these people' we could probably run some sort of NLP thing on it and at least point them in the right direction" (P6).*

4.5.3 Other Applications

Other areas discussed by the participants were indexing and description:

> *"It could be used for indexing or subject guides" (P1).*

> *"But I think if I knew more about the collection and was writing description this might help me think about what the main topics are. It can give some direction about which topics are talked about a lot. So, we can include that in the description" (P3).*

Description would also require time spent on analysis of the topics, but topic modeling can assist in the process that would normally be done without automation.

> *"If I were using it to enhance description on the aggregate level, talking about the whole disk or the whole collection, I would have to make judgements about what these topics actually refer to" (P5).*

> *"We cannot really automate description because you need people to be able to do description well. But you know, this is a good starting point" (P3).*

4.6 Implications

The implications of this study are to enhance archivists understanding of how natural language processing tools could fit into archival selection and appraisal processes. As a result of the study institutions might be encouraged to adopt similar tools. It may be beneficial for information schools to incorporate more digital forensics classes into their curriculum. This paper may also assist software developers by identifying areas of need within real-world applications of appraisal and selection, and provide insight into usability and feature adoption. This is intended to be a jumping off point for future progress in archival automation.

5. Discussion

In this exploration I sought to answer the following questions:

1) Could the BitCurator NLP software be useful for archivists in their appraisal/selection process for born-digital materials?
2) What functions does NLP have for archival selection/appraisal in an institution that relies heavily on human analysis of materials?

Following completion of the study and analysis of the findings, I am prepared to provide answers to these questions. BitCurator-nlp-gentm could be useful for archival appraisal and selection, but only in certain instances. It has proven to be effective in producing a quick, at-a-glance view of the subject matter in any given corpus. The tool is most effective at searching a large body of materials, potentially across multiple media. Many archivists take an all-or-nothing approach to their appraisal; if the media seems to be even partially relevant, all of it will be kept. Not all archivists take this stance – some may want to get down to the item level and "weed" from there. In either case the tool did not offer the study's participants a way to identify those individual items to be ingested or divested.

The findings revealed that archivists are interested in looking for viruses, sensitive information, and duplication. BitCurator-nlp-gentm is not meant to act as a virus checker or to parse through and find PII – those are the domain of other tools. The bitcurator-nlp-gentm tool may however succeed in identifying duplication within a collection that contains born-digital materials and analog materials. This would require extensive interpretation on behalf of the archivist. While not able to compare individual files, the tool can suggest that topics seem to be repetitive across the collection and may be duplications.

To address the second question this study set out to answer, there are multiple functions that have relevance within an archival institution that relies heavily on human analysis of materials. The primary function is that NLP can analyze large amounts of data fairly quickly, using algorithms designed to emulate human thinking. The topic modeler divides terms into topics meant to make sense to the user. Missing is a function that explicitly tells the user what the topic is, which may be seen as a positive or a negative. Labeling the topics would save time in analysis, but trust in algorithmic outputs does not necessarily come naturally. Unlabeled topics do however give the archivist an opportunity to explore the topics on their own. The other major function of this tool helpful for archivists is the ability to compare topics with each other, and the overall corpus.

The first impressions from the participants were that the visualization was “cool” but they would not understand it without assistance. One person commented, *“It takes a little bit of decoding. If you weren’t here I would probably struggle with what this visualization is doing.”* Participants had trouble with deciphering what the visualization was showing as well as understanding the inner-workings of the natural language processing as it divided topics into a modeled output. To paraphrase another participant, the process was little bit “headier” than they are comfortable with.

One of the biggest pain points in appraisal is how to make decisions when the contents of the disks are unknown. The process of investigating the disk can be time consuming and may often be less fruitful than expected. Archivists are looking for a tool that will help figure out what is on the media without going through a painstaking workflow process. The option to feed whole disk images or a set of files into an NLP tool for quick analysis has the potential to alleviate some of these pain points.

6. Conclusion

The overall response to this tool is positive. The archival world seems ready to apply these types of tools into their workflows although appraisal may not be where the need for a tool like this currently lies. A lot of the conversation drifted toward description and the desires for a natural language processing tool to assist in that area. What this topic modeler does well is provide an overview of what the disk contains and makes quick work of the analysis of the disk's topics. For appraisal, archivists need to act on this information and as the tool currently is not able to point to specific files it is limited in assisting the archivist. In practice, this tool will likely be used as a way to group disks in a large collection. In a scenario where a large collection was donated and spread over a variety of media, topic modeling would help the archivist group like items.

A possible use for this item could extend beyond the archivist to the researcher. There may come a time when it makes sense to utilize these tools in the reading room. Topic modeling can be useful to researchers who have subject matter expertise on particular topics and can do a more effective job interpreting and analyzing the algorithmic output. However, the everyday researchers would not immediately know how to use it and they require training. The next logical step is to educate archivists on the workings of NLP tools, and promote teachings of the tools to their patrons. In this case, researchers who have the subject matter expertise of the collections they seek and would get real value from an at-a-glance view of the materials.

Generally, training in natural language processing is essential for all. If it was better understood there may be more desire to implement these tools into their workflows. Understanding NLP tools may also help to alleviate some of the stresses about how interpret an algorithmic output. Human bias will always be a concern but that comes with the territory. I am of the opinion that tools such as this will be making an entrance permanently into archives in the near future.

Bibliography

Angelov, B. (2018, December 15). What do successful people talk about? [Digital image]. Retrieved from https://towardsdatascience.com/what-do-successful-people-talk-about-a-machine-learning-analysis-of-the-tim-ferris-show-161fc7ed4394

BitCurator NLP (n.d.) Project website. Retrieved from, https://www.bitcurator.net/bitcurator-nlp/

BitCurator NLP (n.d.) project wiki. Retrieved from https://github.com/BitCurator/bitcurator-nlp/wiki

BitCurator (n.d.) bitcurator-nlp-gentm. Retrieved from https://github.com/BitCurator/bitcurator-nlp-gentm

Gilliland-Swetland, Anne. "Development of an Expert Assistant for Archival Appraisal of Electronic Communications: An Exploratory Study," PhD Dissertation, *University of Michigan*

GitHub (n.d.) NLP before and after SpaCy. Retrieved from https://github.com/chartbeat-labs/textacy

Gengenbach, M. (2012). The way we do it here: Mapping digital forensics workflows in collecting institutions. Unpublished master's thesis, *The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

Greenberg, J. (1998). The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives. *The American Archivist*, 61(2), 400–425.

Greene, M.A., & Daniels-Howell, T.J. (1997). Documentation with an attitude: A pragmatist's guide to the selection and acquisition of modern business records. In J.M. O'Toole (Ed.), *The Records of American Business* (pp. 161-230). Chicago, IL: Society of American Archivists.

Ham, F. (1975). The archival edge. *The American Archivist*, 38(1), 5-13.

Hutchinson, T. (2017). Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2251–2255). Boston, MA, USA: IEEE. https://doi.org/10.1109/BigData.2017.8258177

Bibliography Continued

Kirschenbaum, M. G., Ovenden, R., Redwine, G., & Donahue, R. (2010). Digital forensics and born-digital content in cultural heritage collections. *Washington, D.C: Council on Library and Information Resources.*

Lee, C. A., Chassanoff, A., Woods, K., Kirschenbaum, M., & Olsen, P. (2012). BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine*, 18(5/6). https://doi.org/10.1045/may2012-lee

Harvey, R., & Thompson, D. (2010). Automating the appraisal of digital materials. *Library Hi Tech*, 28(2), 313–322. https://doi.org/10.1108/07378831011047703

Ross, S., & Gow, A. (1999). Digital archaeology: Rescuing neglected and damaged data resources. *A JISC/NPO study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials.*

Samuels, Helen Willa.(1992). *Varsity Letters: Documenting Modern Colleges and Universities.* Chicago, IL: Society of American Archivists

Scikit-learn (n.d.) Home page. Retrieved from https://scikit-learn.org/stable/

Schellenberg, Theodore R. (1956). *Modern Archives: Principles and Techniques*. Chicago, IL: University of Chicago Press

Schneider, J. (2016). ePADD: Supporting Archival Appraisal, Processing, and Research for E-mail Collections. *MAC Newsletter*, 43(3), 8.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *In Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Standford Libraries (n.d.) ePADD. Retrieved from https://library.stanford.edu/projects/epadd

Pearce-Moses, R., & Baty, L. A. (2005). *A glossary of archival and records terminology* (Vol. 2013). Chicago, IL: Society of American Archivists.

Textract (n.d.) Textract. Retrieved from https://textract.readthedocs.io/en/stable/

Bibliography Continued

Walters, T. O. (1996). Contemporary Archival Appraisal Methods and Preservation Decision-Making. *The American Archivist*, 59(3), 322–338.

Appendix

Semi-Structured Interview Questions:

1) Tell me about your collections. What do you collect here?

3) What role do you have in Appraisal/Selection

4) Can you talk me through your appraisal process?
   a) What are you looking for when appraising records?
   b) What decisions do you make throughout the process?
   c) How long does this take?

5) How Familiar are you with NLP?

6) Do you have a vision for how NLP might help you in your work?

At this point we introduce bitcurator-nlp-gentm

7) Overall, what is your impression?

8) How might this tool help in your appraisal process?

9) Is there anything missing?