Diverse Digital Collections Meet Diverse Uses: Applying Natural Language Processing to Born-Digital Primary Sources

Christopher A. Lee
University of North Carolina at Chapel Hill
callee@unc.edu

Kam Woods
University of North Carolina at Chapel Hill
kamwoods@unc.edu

ABSTRACT

Use of primary sources often focuses on identifying and tracking entities (e.g. people, places, organizations, events) and other values (e.g. dates and times) across documents. There are many existing open-source natural language processing (NLP) tools that can identify and report on named entities, and projects in the digital humanities have previously demonstrated the scholarly value of NLP approaches when working with digitized materials. To date, there has been relatively little adoption of NLP tools for the analysis of born-digital materials by libraries, archives and museums (LAMs). There are a variety of challenges associated with applying NLP tools to born-digital primary source collections, including those forensically acquired from removable media. Many of the challenges relate to the diversity of materials and potential use cases. This paper reports on the BitCurator NLP project, which is developing software for LAMs to extract and expose features in text extracted from such materials. The resulting services and methods can be used by LAM professionals and the users they serve.

CCS CONCEPTS

· Information systems~Digital libraries and archives
· Computing methodologies~Natural language processing
· Security and privacy~Data anonymization and sanitization

ADDITIONAL KEYWORDS AND PHRASES

BitCurator NLP; archival processing; named entities; text processing

INTRODUCTION

A growing body of materials with significant cultural value are "born digital." Libraries, archives and museums (LAMs) are increasingly called upon to move born-digital materials from their original locations – whether those are networked environments or removable media (e.g. floppy disks, flash drives, CD-ROMs, hard drives) – into more sustainable preservation environments. Information professionals must be prepared to extract digital materials from removable media in ways that reflect the rich metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data.

There has been a significant shift in recent years toward the adoption of digital forensics tools and methods by LAMs in order to meet the above goals. This process has been facilitated by the BitCurator project (2011-2014), funded by the Andrew W. Mellon Foundation, which has packaged and disseminated an open-source software environment[1] that allows users to create disk images; extract data and metadata from disks or directories; scan bitstreams for the presence of potentially sensitive data values; characterize the contents of disks; and perform other practical tasks, such as scanning for viruses, finding duplicate files, creating and working with forensically packaged disk images, generating cryptographic hashes, and viewing hexadecimal representations of bitstreams.

The BitCurator Access project (2014-2016), also funded by the Andrew W. Mellon Foundation, has further advanced these activities by investigating mechanisms for providing access to forensically-acquired data. A major product of the project has been BitCurator Access Webtools, which allows users to dynamically navigate filesystems of disk images, as well as searching over the content of many common files types contained within the images.[2] The project also created BitCurator Access Redaction Tools to redact strings and byte sequences identified in disk images.[3] This includes the ability to overwrite specific strings or regular expression matches, or byte runs that match specific files or directory entries.

Members of the BitCurator user community and other interested LAM parties have expressed a need for tools to help in identifying and exploring information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers. These needs can be addressed by building existing natural language processing (NLP) and information visualization tools on top of the existing BitCurator environment and BitCurator Access Webtools. This combination of functions can beneficial both LAM staff and a variety of end users of digital collections.

RATIONALE

[1] https://wiki.bitcurator.net/
[2] https://github.com/bitcurator/bitcurator-access-webtools
[3] https://github.com/BitCurator/bitcurator-access-redaction

One of the primary motivations for using the BitCurator and BitCurator Access software is to capture and provide access to contextual information. For example, the original filesystem attributes associated with files (e.g. directory paths, timestamps) can be essential to understanding their provenance and original order. There are many other types of contextual information that can be vital to making sense and meaningful use of digital objects. Lee's "Framework for Contextual Information in Digital Collections" identifies nine classes of contextual entities: object, agent, occurrence, purpose, time, place, form of expression, concept/abstraction and relationship [1]. In a study of reference questions submitted to archives, Duff and Johnson found that most information requests were based on "proper names, dates, places, subject, form, and, occasionally, events when composing their information request" [2]. In their study of genealogists, Duff and Johnson identified information seeking practices that were focused primarily on names, places and time periods [3].

PREVIOUS WORK

There have been several recent initiatives to better exploit specific types of contextual information from within archival descriptions. Much of this work has focused on what Lee's framework would classify as agents: individuals, families and organizations. Two standardization efforts have focused on characterization of such contextual information: Encoded Archival Context (EAC) [4] and the International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR (CPF)) [5]. The Social Networks and Archival Context (SNAC) project has extended the work of EAC-CPF by exploring methods to better create, combine and disseminate name records for persons, families and corporate bodies. While SNAC has been valuable in exploiting and exposing metadata from within human-generated archival descriptions, it has focused on one specific set of contextual entities and it has not addressed the automatic extraction of entities from the content of digital objects themselves.

Open source natural language processing platforms have matured rapidly during the past decade. These include platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Popular open-source toolkits for natural language processing include OpenNLP, NLTK, Pattern, and spaCy. Some of these platforms have been used in projects specifically targeted at LAMs, but the use cases are often narrow, and none include facilities specifically designed to process content from disk images.

One project that incorporates NLP functionality is ePADD (email Processing, Appraisal, Discovery, and Delivery), developed by Stanford University's Special Collections and University Archives [7]. The ePADD software allows LAMs to process collections of email by using a customized Named Entity Recognition (NER) engine to identify correspondents within email. The NLP functions developed for ePADD have been customized to the domain of email materials to ensure high-quality results.

Another project with goals closely related to BitCurator NLP was ArchExtract, conducted (2014-2015) at the University of California, Berkeley. Bancroft Library staff worked with a student at the Berkeley iSchool, Janine Heiser, to develop a prototype for applying topic modeling, named entity extraction, and analysis of other common terms to collections of text documents, which Heiser tested against a collection of digitized materials. Topic modeling, in this instance, is a text mining technique used to cluster words identified in the corpus into abstract "topics." These clusters may include entities extracted using NER, and can be cross-referenced to entities of interest to the researcher. The ArchExtract software, available through GitHub,[4] is a proof of concept and not currently being actively maintained.

In the digital humanities, there have been many years of work on applying NLP to the content of primary sources. Projects in the field often focus on specific areas of NLP, such as NER and topic modeling to provide researchers with meaningful views of the people, organizations, and events described within a formal collection or data gathered from the Web. There is great potential to apply these methods more widely to LAM collections in order to identify and expose the sorts of contextual entities discussed above. Examples of such digital humanities projects include Perseus,[5] WordHoard,[6] Nora,[7] Metadata Offer New Knowledge (MONK),[8] and the Software Environment for the Advancement of Scholarly Research (SEASR)[9] being used by the HathiTrust Research Center (HTRC).[10] SEASR provides a modular analytics approach to improve scholars' access to digital research materials through a text mining application capable of summarization, an ngram viewer, named entity extraction, and entity-to-graph visualization. This project is no longer active; the last major release was published in 2013.

Other projects have focused on specific types of contextual entities, such as PeriodO (Periods, Organized), which is creating a gazetteer of scholarly assertions about the spatial and temporal extents of historical periods in order to facilitate linking across data sources [8]. These activities have contributed to the available infrastructure for processing texts to facilitate research, but they have been applied predominately to digitized primary sources, rather than to born-digital sources.

BITCURATOR NLP PROJECT

BitCurator NLP (2016-2018), funded by the Andrew W. Mellon Foundation and led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS), is developing and disseminating software that responds directly to the user needs expressed above by identifying, extracting and exposing contextual entities from the wide diversity of born-digital materials that LAMs already hold and continue to receive.

[4] https://github.com/j9recurses/archextract
[5] http://www.perseus.tufts.edu/
[6] http://wordhoard.northwestern.edu/
[7] http://web.archive.org/web/20070123113233/http://www.noraproject.org/
[8] http://monk.library.illinois.edu/
[9] http://www.seasr.org/
[10] https://www.hathitrust.org/htrc_collections_tools

The BitCurator and BitCurator Access projects have defined and tested support for a digital curation workflow that begins at the point of encountering holdings that reside on removable media – either new acquisitions or materials that are within a repository's existing holdings – and extends this to the point of interaction with end users, providing and supporting a variety of access mechanisms. BitCurator NLP is extending this earlier work by helping LAMs to identify and explore information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers.

Our target use cases differ from previous work in two fundamental ways. First, disk images are internally complex (containing filesystems and sometimes bootable operating systems) and require a significant software dependency stack that is already provided by the BitCurator environment and BitCurator Access Webtools. These include the ability to read, mount and provide access to the contents of various filesystems, as well as extracting, presenting and reporting on files and metadata.

A second distinguishing factor in the application of NLP to disk images in LAM collections is that disks may contain a broad range of file types and data encodings, requiring substantial pre-processing to extract content so that it can be processed by NLP tools and organized into meaningful reports, access points and visualizations. By contrast, most previous applications of NLP methods have focused on more "well-behaved" collections or corpora with more consistent types of content. BitCurator NLP is building from and extending a variety of tools and initiatives discussed above to provide services that LAMs can be run independently or integrate into existing software environments and access portals via simple application programming interfaces (APIs).

Some LAM access systems in use today incorporate NLP to describe the contents of collections, but this technology is often tightly coupled to the platform being used or is applied strictly to file types that tend to share common structures and metadata. An example of this is email, which contains raw text but also significant structured metadata in the header that can be used to cue identification of persons and organizations and describe their relationships. Identifying entities, relationships, and other features of interest by processing open text from heterogeneous collections of files (such as those extracted from a disk image) is inherently "noisier," as the extracted text will often contain patterns of features (such as persons, places, and organizations) common to a wide range of devices and production environments (e.g., documentation of system files).

BitCurator NLP is exploring approaches that focus on improving the utility of reports produced about the contents of born-digital collections. Using data extracted from open text using NLP tools, along with techniques described in recent digital forensics research to eliminate or deemphasize those that appear to be irrelevant or common to the system rather than the documents themselves (e.g., names and email addresses of developers or organizations that created the software used to produce a given document), the project team will also develop guidelines describing how to apply the tools in ways that support common access and research use cases.

The BitCurator NLP team is ensuring close integration between the existing functionality of the BitCurator environment, BitCurator Access Webtools and the software developed by the BitCurator NLP project. For example, we are increasingly making elements of the BitCurator environment available as self-contained software installers (specifically software packages that may be installed in Ubuntu and Debian Linux environments using existing package managers), so users can selectively install and update them as they find most useful. Institutions could load each of the access tools onto the same machine (or virtual machine) as the one they are using for the initial processing, or they could instead decide to run those tasks in different environments in order to better manage and allocate their resources.

CONCLUSION

Given the numerous projects that have focused on application of NLP to enable scholarship on older primary sources, it is striking how little investigation there has been into use of NLP to the growing volume of more contemporary born-digital materials acquired by LAMs. While these professionals may have access to some tools that apply a subset of NLP techniques (such as topic modeling, NER, and word or phrase frequency analysis) to specific collections or file types, they generally lack software allowing them to execute these techniques in aggregate for large sets of heterogeneous files (including complex files such as disk images). This is particularly important for materials that contain thousands or hundreds of thousands of files, when it is intractable to manually inspect materials to determine which of the files are relevant preservation targets and what relationships hold between the files that could be exposed in archival description. Providing mechanisms that allow LAM professionals to conduct this type of analysis in the existing BitCurator environment and in BitCurator Access Webtools will streamline this process. The BitCurator NLP software is enabling LAMs to incorporate (or improve) basic NLP capabilities within existing access environments (or establish them as a dedicated service) by providing a service layer that reads files from a digital archival store and produces reports for end users on demand.

ACKNOWLEDGEMENTS

This work has been funded by the Andrew W. Mellon Foundation. The BitCurator NLP team is composed of Jacob Hill, Christopher A. Lee, Sunitha Misra and Kam Woods. We have also benefited from contributions of the project Advisory Board.

REFERENCES

[1] Lee, C. A. Lee, "A Framework for Contextual Information in Digital Collections." Journal of Documentation 67, no.1 (2011): 95-143.

[2] Duff, W. M. Johnson, C. A. A Virtual Expression of Need: An Analysis of E-Mail Reference Questions. American Archivist 64 (2001): 43-60.

[3] Duff, W.M. and Johnson, C. A. Johnson. Where is the List with All the Names? Information-Seeking Behavior of Genealogists. American Archivist 66 (2003): 79-95.
[6] Pitti, D., Hu, R. Larson, R., Tingle, B. and Turner A. Social Networks and Archival Context: From Project to Cooperative Archival Program. Journal of Archival Organization 12 (2015): 77-97.
[7] Hangal, S., Piratia, V., Manovit, C., Chan, P., Edwards, G. and Lam, M. S. Historical Research Using Email Archives. Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. ACM Press, New York, 2015.
[8] Golden, P. Shaw, R. Nanopublication Beyond the Sciences: The PeriodO Period Gazetteer. PeerJ Computer Science 2:e44 (2016).