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

ABSTRACT

In this tutorial, participants will learn about and gain hands-on experience with products of the BitCurator NLP project, which is developing software for libraries, archives and museums (LAMs) to extract and expose features (e.g. people, places, organizations, events, relationships, topics) in text extracted from born-digital materials. The services and methods can be used by LAM professionals for appraisal and description, as well as facilitating a wider range of access and use scenarios.

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 entity recognition; text processing; topic modeling

INTRODUCTION

Libraries, archives and museums (LAMs) are increasingly called upon to move born-digital materials from their original locations 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, mounting 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, investigated 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]

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 for identifying, extracting and exposing contextual entities from the wide diversity of born-digital materials that LAMs already hold and continue to receive. This includes helping to identify and explore information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers.

There are many existing mature open source natural language processing platforms, including platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Production-quality open source software toolkits for natural language processing include OpenNLP (Java-based) and NLTK, Pattern, and spaCy (Python-based).

Our target use cases differ from previous work in two fundamental ways. First, disk images are internally complex and require a significant software dependency stack that is already available through the BitCurator environment and BCA 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 their data and metadata.

A second distinguishing factor is that disks may contain a broad range of file types and data encodings, requiring substantial

[1] https://wiki.bitcurator.net/

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

pre-processing to extract content so that it can be processed by NLP tools and organized into meaningful reports, access points and visualizations. BitCurator NLP is building from a variety of existing tools and initiatives to provide services that LAMs can be run independently or integrate into existing software environments and access portals via simple application programming interfaces (APIs).

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 from 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 the various elements of the BitCurator environment available as self-contained software installers (software packages that may be installed in Ubuntu and Debian Linux environments), so users can selectively install and update them as they find most useful. Institutions could load all 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.

TUTORIAL OUTLINE

This half-day tutorial will have the following structure:

- Brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts
- Demonstration and hands-on exercises that demonstrate specific tools and methods, including text extraction, entity and entity relationship extraction and topic segmentation and recognition
- Brief summary of targeted visualization scenarios using e.g. the scattertext python library[4]
- Concluding discussion about implications for participant's institutional practices

INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for acquiring or transferring collections of digital materials. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how data generated by the BitCurator NLP tools can complement and potentially be integrated with data generated by other tools and systems.

EXPECTED LEARNING OUTCOMES

This tutorial will prepare participants to use the open-source BitCurator NLP tools to process, investigate and visualize born-digital collections. This will include text extraction from heterogeneous collections of file and executing NLP tasks such as part-of-speech tagging, entity and entity relationship extraction, and topic segmentation and recognition.

Upon completion of this tutorial, participants should understand several of the major motivations and uses cases for applying the BitCurator NLP tools to meet digital preservation and access objectives. They should also recognize the limitations and how they might be used in combination with other tools and methods.

Participants will also become aware of the resources that are available for learning more about the software and engage with other users after completion of the tutorial.

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.

[4] https://github.com/JasonKessler/scattertext