BitCurator NLP: Natural Language Processing for the Rest of Us

CHRISTOPHER (CAL) LEE

Abstract:
This poster will present the rationale and preliminary products of 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). The project 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, including on physical media. This includes helping to identify and explore information based on specific entities (e.g. people, places, organizations, events) using natural language processing (NLP).

Our target use cases differ from previous NLP work in two fundamental ways. First, disk images are internally complex and require a significant software dependency stack that is already available through BitCurator software. A second factor 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.

We are 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 digital forensics techniques 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), we will also develop guidelines describing how to apply the tools to support common access and research use cases.

About the Author:

Christopher (Cal) Lee is Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches archival administration; records management; digital curation; understanding information technology for managing digital collections; and digital forensics. His primary area of research is the long-term curation of digital collections. Cal edited and provided several chapters to *I, Digital: Personal Collections in the Digital Era*, published by SAA. Cal is Principal Investigator of BitCurator NLP and was Principal Investigator of the BitCurator and BitCurator Access projects. He has served as Co-PI on several digital curation education projects with Helen Tibbo, and he is currently Co-PI for OSSArcflow.