BitCurator NLP: Natural Language Processing for Archivists

CHRISTOPHER (CAL) LEE

Abstract: This presentation will describe the rationale and products of BitCurator NLP, a two-year project funded by the Andrew W. Mellon Foundation that ends in September 2018. We have developed software to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections. We are using existing natural language processing (NLP) software to identify and report on items likely to be relevant to ongoing preservation, information organization, and access activities. These may include entities (e.g., persons, places, and organizations), potential relationships among entities (e.g., for example, entities that appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents. The project has also made significant enhancements to BitCurator Access Webtools, a system that allows users to search, navigate and examine the contents of disk images.

This project differs 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 pre-processing to extract content so that it can be processed by NLP tools.

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 courses and workshops in archives and records management; understanding information technology for managing digital collections; and digital forensics. His primary research focus is the long-term curation of digital collections. Cal edited and provided several chapters to *I, Digital: Personal Collections in the Digital Era*. He has served as Principal Investigator of the BitCurator, BitCurator Access, and BitCurator NLP, and Co-PI on OSSArcFlow projects, as well as several projects (including DigCCurr and ESOPI) focused on digital curation education. Cal is a Fellow of the Society of American Archivists, and he serves as editor of *American Archivist*. Established in 1938, *American Archivist* is the most widely circulated English-language archival journal.