Machine Learning For Big Text

A Tutorial On Using Predictive Coding Tools To Process Large Archival Datasets

Brent West
University of Illinois
United States
bmwest@illinois.edu
https://orcid.org/0000-0001-6961-9141

Joanne Kaczmarek
University of Illinois
United States
jkaczmar@illinois.edu
https://orcid.org/0000-0001-8081-4570

Abstract – Big datasets can be a rich source of history, yet they pose many challenges to archivists. They can be difficult to acquire and process due to the varied formats and sheer volume of files. Sensitive content must be identified in advance of making materials publicly available. These challenges inhibit access for research purposes and often dissuade archivists from acquiring big datasets. Predictive coding can alleviate these challenges by using supervised machine learning to: augment appraisal decisions, identify and prioritize sensitive content for review and redaction, and generate descriptive metadata of themes and trends. Following the authors' previous work processing Capstone email, participants will learn about innovative and effective practices to enable digital preservation of large textual datasets at scale. Hands-on experience with specific tools is provided.

Keywords – access, active learning, appraisal, automatic classification, descriptive metadata, digital archives, digital humanities, digital preservation, e-discovery, email archiving, ingest, natural language processing, PII, preserving email, privacy, redaction, restricted records, scalability, sustainability, software-as-a-service, supervised learning, technical infrastructure, technology-assisted review, text mining, unstructured data

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation; Exploring New Horizons.

I. Introduction

The Records and Information Management Services (RIMS) office of the University of Illinois, in conjunction with the University Library and the Illinois State Archives (ISA), is nearing completion on a project to acquire, process, and provide access to a collection of email messages from senior government officials of the State of Illinois [1]. The project is generously funded by a three year grant through the National Historical Publications and Records Commission (NHPRC). A unique aspect of this project is the application of commercial tools to efficiently process this large dataset. In particular, the project leverages tools developed by the legal community for electronic discovery (e-discovery) to augment the preliminary archival review and increase processing output. This tutorial provides direct, hands-on access to the tools used by the project team [2] so that participants gain practical experience.

II. Description

This tutorial provides an introduction to predictive coding, a subset of machine learning, and its potential to help archivists make appraisal decisions about large textual datasets. Facilitators will describe in detail, and demonstrate, specific tools used for a project to appraise and make available a large dataset of government email. Participants will be given access to the tools and a dataset prepared in advance for their use during the tutorial as part of a hands-on exercise. Participants may optionally bring their own dataset to use with the tools. During the exercise, participants will explore visual display features, conduct faceted searches, and actively train the tool's predictive coding model to see how the training process works. They will also learn about the limitations of predictive coding tools in this setting, and how to calculate costs of manual review versus computer-assisted review. Participants will

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.
Copyright held by the author(s). The text of this paper is published under a CC BY-SA license (https://creativecommons.org/licenses/by/4.0/).

engage in an in-depth discussion, driven by their own experiences, about challenges facing archivists looking to appraise, process, and make available to the public large datasets. This tutorial is anticipated to last 3-4 hours

A. Target Audience

This tutorial is designed for Archivists, Digital Curators, and Collections Managers who currently, or may in the future, work with large textual datasets. No prior knowledge is necessary, other than a general familiarity with archival appraisal concepts and general familiarity with personal computers, as would be expected for most conference attendee. Familiarity with The Future of Email Archives [3] is beneficial.

B. Learning Goals

Participants will:

1. Gain a basic understanding of machine learning generally, and predictive coding in particular.
2. Identify challenges associated with appraising and processing large textual datasets and learn how predictive coding may help remediate those challenges.
3. Practice working with machine learning tools to prepare large textual datasets for public access.

C. Agenda

1. Introductions and project overview
2. Discussion of participants' experiences and challenges with big data
3. Predictive coding methodology overview
4. Demonstration and hands-on lab
5. Wrap-up discussion of lessons learned and potential use cases

References

[1] University of Illinois Records and Information Management Services. 2018. Processing Capstone Email Using Predictive Coding. http://go.uillinois.edu/capstone.

[2] Kaczmarek, J. and West, B. 2018. Email Preservation at Scale: Preliminary Findings Supporting the Use of Predictive Coding. In *15th International Conference on Digital Preservation*, (Boston, USA). https://osf.io/yau3c/.

[3] Council on Library and Information Resources. 2018. The Future of Email Archives. https://www.clir.org/pubs/reports/pub175/.

Copyright held by the author(s). The text of this paper is published under a CC BY-SA license (https://creativecommons.org/licenses/by/4.0/).