AUDIOVISUAL METADATA PLATFORM

Lightning talk: Commercial ML Tools in Metadata Production

- Challenge:
  - Growing quantity of digitized and born-digital AV media in library and archival collections
  - Lack of metadata for Discovery, Identification, Navigation, Rights, Accessibility
  - Institutions lack resources for large cataloging/transcription/inventory/rights clearance projects
- Proposed solution:
  - Leverage machine learning together with human expertise to produce more efficient workflows
  - Workflow pipeline for “Metadata Generation Mechanisms” - can be automated or human
- Goals of current project phase:
  - Design and build workflow system
  - Evaluate and integrate commercial and open source MGMs
  - Test using collections from Indiana University and NYPL

Jon Dunn / Indiana University / @jwdunn
Shawn Averkamp / AVP / @saverkamp

New York Public Library

THE ANDREW W. MELLON FOUNDATION

How to select accurate, low barrier tools for a production environment?

Evaluation criteria | *accuracy, cost, ease of implementation, social impact, computing resources, processing time, privacy/data reuse, model training needs*

To date | speech-to-text (STT), named entity recognition (NER), video OCR, speaker diarization, and speech/music/silence detection

Next | music analysis (genre detection, instrument identification), object detection, entity reconciliation

Commercial ML tools

PROS	CONS
ease of use	black box
high accuracy	no visibility into training data
custom vocabularies	limited trainability
integration into cloud infrastructure	vague terms of service for data reuse

Findings and issues

- Need for diverse sample data to evaluate black box algorithms
  - Quantifying accuracy is difficult with few representative samples and large project scope
  - Ground truth samples should be selected to draw out anticipated biases
- Terms of service often require opt-out of data reuse rather than opt-in
  - AWS requires written request to opt-out of Amazon use of data. Vague wording in TOS about employees who have access to your data
  - Google offers discounted pricing for opt-in data logging
- Commercial algorithms are a moving target
  - Evaluation will need to be ongoing to monitor accuracy and hidden bias
- Commercial tools are not always available or may need training
  - Some tasks not addressed by off-the-shelf tools will require open-source solutions and/or training expertise

Project wiki: https://go.iu.edu/amppd
Twitter: @AVMetadata