Clustering Metadata Across Slavery Collections

Emily Higgs, MSIS

Research Questions
• How are historical records related to slavery in the United States described across cultural heritage institutions?
• Can an algorithm detect a difference between “critical” and “traditional” metadata approaches?
• How are these approaches reflected in counts and features?

Data Set
• 15,000 metadata records for documents that relate to slavery in the US
• 35 contributing institutions
• DPLA API (JSON output)

Data Cleaning
• Python: JSON to CSV, pulling select fields
  • Institution, collection, title, creator, description, subjects, number of subjects
• Lack of uniformity is huge roadblock for use of data
• Lost linked data; 15,000 -> 14,309 records

Results: Consistent Across Attempts
• 1 large cluster, 82% of records, with median of 4 subject headings (the “standard”)
• 2 small outlier clusters of records from single collection
• 2 medium clusters of records from majority single institution
• 2 medium clusters with no obvious internal consistency to human eye, medians of 11 and 12 subject headings

Insights and Future Directions
• Machine-recognized “standard” descriptive practice with meaningful deviations
• Do the discrepancies between the standard and deviations match the way librarians and archivists talk about metadata creation?

Most Python code adapted from Steve McLaughlin, TA for course (and the rest written with his help)

Class project for Professor Tanya Clement’s “Critical Perspectives in Cultural Data Analysis” course at the UT Austin iSchool, Fall 2017