LG-86-18-0061-18

UNIVERSITY of DENVER

Text Duplication and Similarity in the HathiTrust Digital Library

Summary: Content-based analysis in digital libraries provides a new domain for cultural and historic insights, but its effectiveness is limited by issues of duplication and text variants. The University of Denver, in cooperation with the HathiTrust Research Center at the University of Illinois, requests $262,968 to develop tools to identify multiple levels of work duplication, reveal archetypal 'best' copies of each work, and recommend non-duplicate works and authors. Grant funds will be supplemented by $67,847 of voluntary cost-share.

Introduction: Massive digital libraries have enabled a class of collection questions that can only be asked of our aggregate published history. Yet, models seeking to learn from them are frustrated by a key issue accompanying scale in digital libraries: duplication and text variants. We propose a content-based study of text duplication and similarity in massive digital collections. This project will produce volume-level relationship datasets, of immediate value in public and academic libraries for supporting access services such as recommendation and retrieval. At the same time, it will tackle a critical hurdle for text mining scholars seeking new historical and cultural insights from massive digital libraries.

Objective: The outcome of this research project is curatorial: to learn more about the works of the 15 million volume HathiTrust Digital Library (HTDL) and provide methods to transfer the results to other digital libraries. The goal is to learn more about the individual volumes by modeling them and their relationships at a collection-wide scale, then taking that learning back to the volume level. This study will focus on three research questions:

1. How do you accurately disambiguate varying levels of duplication in the HTDL? Recent work has found that repeating text causes issues for information modeling[1], but controlling for the issue is surprisingly complex. While scans of identical print runs (duplicates of the same manifestation[2]) can be inferred from metadata, the challenge is amplified when considering variant expressions - e.g. aligning works that exist in multiple versions or editions - or variant manifestations - e.g. multi-volume sets or parts of a collected anthology. Scanning errors add further complications.

2. What is the canonical copy of each work in the HTDL, the 'best' copy for access or analysis? For retrieval, access, and analysis, not all digital volumes are equally valuable. We seek to identify archetypal copies of works by identifying the volume that most resembles the text of all the duplicates, with the smallest amount of deviation from that model.

3. What are the most similar volumes and authors to each volume in the HTDL? In looking at the fuzzy gradients of what a duplicate work is, we also seek to find the volumes just beyond those bounds: those that are not the same work but very similar in content. Recommending similar books by content-based analysis serves as a good complement to metadata-based or expert recommendation, providing an access vector that is good at surfacing lesser-known works.

Outcomes: Results will be released in an open dataset. For each work in the HTDL, the dataset will enumerate: topically similar authors and works; items (i.e. a scan of a volume) grouped by related realizations and manifestations, including multi-book and partial book identification; and ranked recommendations for the items most representative of an archetype of the work. This dataset will be the most broadly useful of a set of deliverables, accompanying reporting on our research, a release of the best performing models and code to use them for other digital libraries, and evaluation datasets for future scholarship.

[1] Schofield, Alexandra, Laure Thompson, David Mimno. 2017. "Quantifying the Effects of Text Duplication on Semantic Models."

[2] Here, we use FRBR vocabulary, which describes the distinct intellectual creation as the work, realized through one or more expressions, each of which is embodied by one or more manifestations, each of which is exemplified by one or more items.

Impact: This project is of national benefit on two levels. On the micro scale, this research infers more about the individual works in library collections, producing methods to reduce noise in digital libraries and data on relationships between works and authors. On the macro scale, it greatly improves scholars' ability to research history, culture, language, and style through massive-scale digital libraries.

Consider the scope of the problem through an example. A reader searching the 'natural selection' subject area of the HTDL can retrieve 518 relevant volumes - 97 of which are various copies, editions, volumes, or remixes of Darwin's Origin of Species. A library that can group these copies in their catalogue search will provide better access to the relevant materials beyond Darwin. A scholar applying text mining to learn overarching themes in natural selection work will be equally challenged by the book's popularity; providing a single archetypal copy for analysis would avoid the bias. In other areas, past scholarship already complements our proposed work to the benefit of library collections, such as research finding that the most effective way to estimate a date of first publication is to use the earliest date of known duplicates in the HTDL[3]. Our work will improve this approach and immediately allow libraries to improve publish date metadata.

Design: We anticipate a two-year timeline, from May 2018 to April 2020. The first year will focus on defining the problem and preparing our methods for a sound evaluation, including annotating corpora for evaluation, training transformations for the book data, and scaling our methods to HTDL sizes. In year 2 we will complete, evaluate, and document our methods, and prepare maximally useful deliverables, including applying our best approach over the full collection, annotating the models, and generalizing the code assets.

In exploratory work[4], we approached this problem on a 101k volume subset of the HTDL, modeling lower-dimensional vector representations of each book, then applying distance metrics to cluster volumes with their partners. Best-copy was determined by identifying the work closest to a duplicate centroid – an averaged point in vector space. The proposed project will improve on this foundation. The exploratory approach is not sensitive to deviations from 'one work per book'; we will pursuing page-level fingerprinting methods in response. We also intend to make the best copy selection more robust to outliers, and assess it with carefully researched evaluation corpora. Finally, the existing methods simply list notably similar works - *candidates* for duplication - without knowing the reasons for the relationship. Learning how these measures align with conceptual boundaries and acting on them is a core concern for our forthcoming work.

Personnel: Peter Organisciak is an assistant professor at the University of Denver (DU), with expertise in large-scale text analysis. He has worked on information access for multiple large consortial collections, including IMLS Digital Collections and Content and the HathiTrust Research Center (HTRC), and will lead the intellectual and technical effort on the project. A developer and domain expert at the HTRC will assist in collecting, cleaning, and interpreting the source data. Developing evaluation corpora, and researching and describing problem edge cases will be aided by a DU library science student, and additional theoretical and methodological assistance will be provided by a DU research methods student.

Budget Summary: This project requests $262,968 in grant funds, to be supplemented by $67,847 of voluntary cost-share for a budget of $330,815. The requested funds include $163,886 for staff salaries and benefits, including the PI and research assistants at DU, as well as consultant and developer time through a sub-award ($20,366) to the HTRC at the University of Illinois. Other costs include travel ($5,605), computation and storage costs ($11,000), and indirect costs ($82,177).

[3] Bamman, Carney, Gillick, Hennesy, and Sridhar. 2017. "Estimating the Date of First Publication in a Large-Scale Digital Library".

[4] Organisciak, Capitanu, Underwood, and Downie. 2017. "Access to Billions of Pages for Large-Scale Text Analysis".