Computational Support for 'Reading Chicago Reading'

HTRC ACS Final Report (August 2018)

Principal Investigators: Robin Burke, John Shanahan, Ana Lucic (DePaul University)

Support: Eleanor Dickson and Boris Capitanu

Project objectives

Our primary object of research is a small set of fiction and non-fiction texts: the selections by the Chicago Public Library (CPL) for the *One Book One Chicago* (OBOC) program since 2011. In the past, we have obtained the circulation data (time-stamped checkouts, check-ins, holds, etc.) for the seven most recent selections (2011-18), but also for a larger set of nearly 350 works made from the CPL staff recommended works for each selection (i.e. "if you liked ... you might also like"). Nearly every one of these almost 350 texts, like the seven official program selections, is in copyright.

Given that the majority of works in our original and comparators set are also under copyright, our project would have been much more difficult to accomplish without access to the HathiTrust Research Center (HTRC) secure environment of the data capsule. While conducting textual analysis of works in copyright is not prohibited, technical and legal barriers make it a challenge to build sets of copyrighted works for text mining. Having access to HathiTrust digital library resources that are specifically built for non-consumptive analysis of works as well as the assistance of HathiTrust staff Eleanor Dickson and Boris Capitanu made the process of analyzing works in copyright much easier. We are grateful to the HTRC for the opportunity to work with this set of texts in the data capsule.

The Advanced Collaborative Support grant from the HathiTrust Research Center allowed us to expand our analysis from the seven most recent *One Book One Chicago* selections to make them part of a larger comparator set made up of CPL "recommended" works associated with the program's official selection. With the help of HathiTrust staff and the staff of the DePaul University Library we were able to:

(1) identify which of our recommended works are available in the HathiTrust library. The assistance of Eleanor Dickinson was instrumental with this process. Out of our initial set of 309 recommended books, we ended up with a set of 70 available in the HathiTrust digital library for use as a comparator set. The HathiTrust library, although vast, does not include scanned copies of many of the CPL recommended works. (Many of the works are new, and almost all are still under copyright. In addition, many are in genres unlikely to be scanned; for instance, because of their connection to the theme of the 2016-17

Animal, Vegetable, Miracle program, the set of recommended texts include a picture books as well cookbooks.)

(2) We can also note that the complete HTRC sets for the CPL recommended readings were not equally distributed for our OBOC selection seasons, but we do have have a number of recommended works for each of the seven official OBOC selections.

(3) Out of seven OBOC selections, three were in the HathiTrust digital library at the start of the project and the remaining four were added recently with the help of the DePaul University Library: *Animal, Vegetable, Miracle* by Barbara Kingsolver, *The Third Coast* by Thomas Dyja, *Gold Boy, Emerald Girl* by Yiyun Li and *The Warmth of Other Suns* by Isabel Wilkinson. Having all seven selections in the data capsule has made it easier for us to conduct analysis and run the same processes on both the original and comparator sets. Although the full ingest of the four works mentioned above is not yet complete, we do have access to the scanned copies of the works in the capsule.

Workset building tool

Building worksets in the HathiTrust digital library took up a significant amount of time throughout the duration of the project. This is not surprising given that building a dataset of work represents a critical task for textual (or any other type of) analysis.

Because Chicago Public Library patrons could read the selected book in any number of editions and formats (print/electronic), any edition -- except an abridged one -- of the selection and the recommended works would have been eligible for inclusion in our dataset. We were agnostic about this though we did hope to have the same edition as those chosen by the CPL library branches.

Establishing which manifestations of a particular work, in the FRBR [Functional Requirements for the Bibliographic Records] sense of the word, was not an easy task. It implied having access to and being able to retrieve all the book work IDs (for example, OCLC owis) and all the book *manifestations* (again in the FRBR sense of the word) of a particular work, and, in addition, establishing the HathiTrust volume ID for a particular manifestation. Two programmers associated with the Reading Chicago Reading project, Jeremy Erwin and Dan Aasland, created a python program that can establish the OCLC owis for a work, the book manifestations for a particular work using the OCLC APIs and, finally, by querying the HathiTrust digital library identify the HathiTrust IDs for the manifestations of a particular work. We don’t know of another tool that has such functionality. Being able to see the HathiTrust IDs in relation to the OCLC owis and IDs was tremendously helpful in drawing the boundaries of our dataset and querying the HathiTrust.

We believe a tool of this type will be of significant interest and use to other scholars working with HathiTrust resources as well. We plan to make this program publicly available so that others can use it to query the HathiTrust digital library.

Boundaries of the main text

Building the OBOC set and the comparator set was an essential but time-consuming task that preceded conducting text mining on the two sets together. The next problem that we tackled was establishing the boundaries of the main content in the two sets. Our earlier work with the seven most recent OBOC selections pointed us in the direction of excluding paratext from our analysis. Our original set of seven works contains three non-fiction works that included front matter and back matter (index, etc.). The frame for the main text with all of its elements (acknowledgement page, foreword, table of contents, etc.), especially the index and bibliography, can influence the number of extracted locations as well as the type of location extracted from works if we are not careful to eliminate them during the preprocessing stage. Given that location extraction is one type of information we are keen to retrieve from our worksets, we were interested in delimiting the analysis to the main content and excluding the front and back matter in this process.

And yet, however, there is no *a priori* or error-free way to establish where the main text of a work begins and ends when working with separate files of scanned individual pages of a book. In fact, this is an open problem for any digital library that contains scanned copies of the text unless the main content is accurately and consistently reflected in the adjoining metadata files. The METS files that we have access to in the data capsule contain structural metadata elements that indicate whether a page is the beginning of a chapter, but we established that the quality and consistency of METS files is not guaranteed, which ultimately makes it an unreliable source of information for the boundaries of the main text.[1] For this reason, by default any page inside the secure volume folder is potentially a content page. The question we are dealing with is whether a page that has text on it within the scanned volume belongs to the main text or not. This problem has been tackled, simply but bluntly, for instance, by Ted Underwood, who has eliminated the first 10% of pages as well as last 5% of pages in his corpus.[2] We are keen to find out how easy or how difficult it is to identify the start and end of the main text of a book using automated methods. Eventually, we plan to compare our results to Underwood's results.

Automated method

Co-PI Ana Lucic's automated method of inferring the main content of the book consisted of establishing cosine similarity of the pages in a work and then calculating distance between the

[1] Tom Burton West's blog https://www.hathitrust.org/blogs/large-scale-search/challenges#_edn3

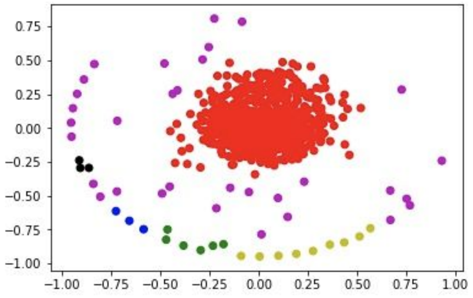
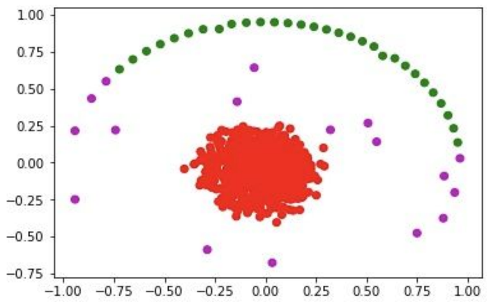
[2] See Underwood's preprint of a paper for the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature: "The Historical Significance of Textual Distances", arXiv:1807.00181, p. 3.

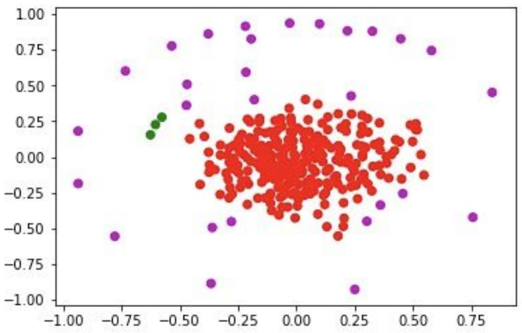
pages. To visualize the distance, multidimensional scaling was used. Both ‘bag of words’ and tf-idf document-term matrices (where individual pages of a work in this instance correspond to documents) were used as input to cosine similarity and the scaled distances between pages served as the input to a DBSCAN clustering algorithm. Sickit-learn python library was used in the background for these tasks. To establish the accuracy of our method we also had to manually identify the start and end pages of the main content. The true start and end pages of the main content in the set were documented manually by graduate student researcher Mihaela Stoica and co-PI John Shanahan.

Preliminary results

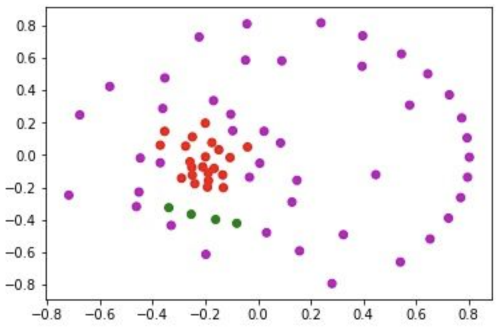
Our tests on the four public domain works included in our list of recommended works (from the comparator set) revealed that we were able to infer correctly the boundaries of *The Adventures of Tom Sawyer* (Fig 1). The least accurate of the four results was the clustering of the content for *The Rhyme of the Ancient Mariner* (Fig 4) -- not surprising given that the edition of the *Rhyme of the Ancient Mariner* in the HTRC contains a long critical introduction in addition to other paratextual elements that can blur the boundaries of the main text. This suggests to us a basic soundness of our methods so far.

The red cluster in each of the figures below indicates the main content for the four books as established by the clustering algorithm. The results thus far look promising in the sense that the main cluster identified by the algorithm corresponds roughly and with varying accuracy to the main text of a work (Figs 1-4). Future work will examine how accurately other clusters identified are able to infer paratext elements of a work. The code that we used to conduct this analysis will be posted on GitHub.

 *David Copperfield* by Charles Dickens https://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t6zw1qm9k	 *The House of Mirth* by Edith Wharton https://babel.hathitrust.org/cgi/pt?id=aeu.ark:/13960/t77s91d3g



Tom Sawyer by Mark Twain
https://babel.hathitrust.org/cgi/pt?id=ucw.ark:/13960/t07w6r028



The Rhyme of the Ancient Mariner by Samuel Coleridge
https://babel.hathitrust.org/cgi/pt?id=coo1.ark:/13960/t7cr6ch7n

Next phase

The next phase of this project will use the established boundaries of the main content to extract the features from the original and the comparator set and analyze the differences. We plan to do the following analyses:

- Location extraction
- Sentiment extraction
- Lexical measures extraction
- Part of speech extraction

Once we have all the measures we will be able to analyze the differences between the main and the comparator sets. These features will serve as input to a predictive model in development.

Summary of project outcomes

- A tool to establish the HathiTrust ID for a given work and find out which manifestation of a work is available in the HathiTrust library

- Comparison of several methods of establishing boundaries of the main text

- Extracting features from the main and comparator set to serve as input to the model

We plan to share the results of this work at the conferences such as JCDL, Digital Humanities, ASIS&T, IConference and in journals such as *Digital Humanities Quarterly*, *Digital Scholarship in the Humanities*, *JASIST*, and *Library Quarterly*.