Towards Cultural-Scale Models of Full-Text project

Overview

This project deploys an improved infrastructure for robust corpus building and modeling tools within the HTRC Data Capsule framework to answer research questions requiring large-scale computational experiments on the HTDL. Our research questions depend on the capacity to randomly sample from full text data to train semantic models from large worksets extracted from the HTDL. This project prototypes a system for testing and visualizing topic models using worksets selected according to the Library of Congress Subject Headings (LCSH) hierarchy.

Project report can be found at http://arxiv.org/abs/1512.05004 Please refer to project report for technical details, administrative, and community impact details.

Personnel

Colin Allen, Jaimie Murdock (Indiana University)

Jiaan Zeng (HTRC)

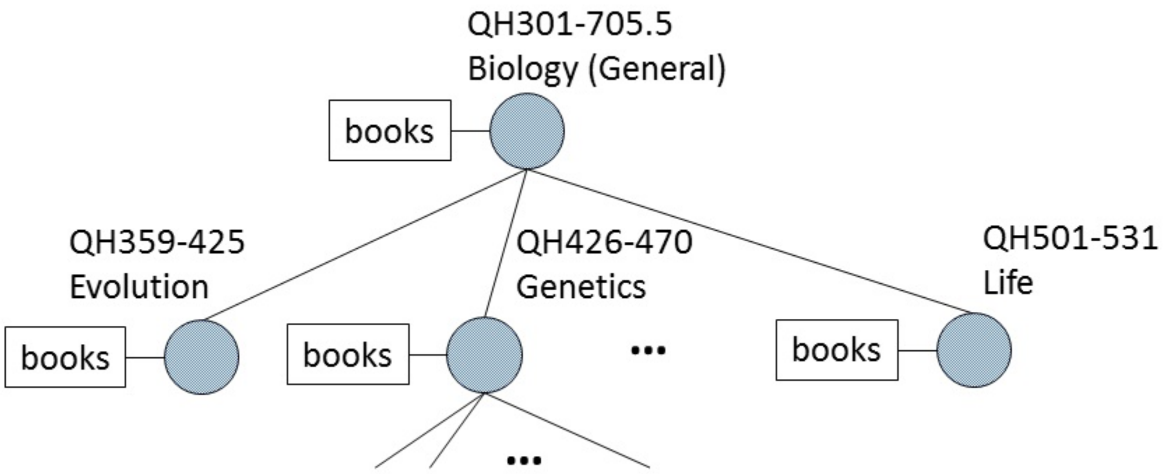
Motivation

Large-scale digital libraries, such as the HathiTrust, give a window into a much greater quantity of textual data than ever before (Michel, 2011). These data raise new challenges for analysis and interpretation. The constant, dynamic addition and revision of works in digital libraries mean that any study aiming to characterize the evolution of culture using large-scale digital libraries must have an awareness of the implications of corpus sampling. Cultural-scale models of full text documents are prone to over-interpretation in the form of unintentionally strong socio-linguistic claims. Recognizing that even large digital libraries are merely samples of all the books ever produced, we aim to test the sensitivity of topic models to the sampling process. To do this, we examine the variance of topic models trained over random samples from the HathiTrust collection.

One methodology with rapid uptake in the study of cultural evolution is probabilistic topic modeling (Blei, 2012). Researchers need confidence in sampling methods used to construct topic models intended to represent very large portions of the HathiTrust collection. For example, topic modeling every book categorized under the Library of Congress Classification Outline (LCCO) as "Philosophy" (call numbers B1-5802) is impractical, as any library will be incomplete. However, if it can be shown that models built from different random samples are highly similar to one another, then the project of having a topic model that is sufficiently representative of the entire HT collection may become tractable.

Workflow

For this preliminary study, we limit our samples to within certain Library of Congress Classification Outline (LCCO) classes. A random sampler program was developed to generate stratified volume samples for a specific percentage from a given category. Below shows part of the LCCO category tree.



To anchor our comparisons, we train several models over the entire class with different random seeds. We perform a topic alignment between pairs of models and take the average distance between the topic-word distributions for each aligned topic pair. Then, we train topic models of random samples of books from the class.

Findings

When we align the models of samples to models over the whole class, we find the average topic distance can exceed the topic distance of alignments between two whole class models. Unsurprisingly, as sample size increases, average topic distance decreases. We also find that the number of topics selected by the topic alignment increases as sample size increases. However, the decomposition of these measures by sample size differs by field and by number of topics. While this study focuses on only five areas, we speculate that these measures could be used to find classes which have a common "canon" discussed among all books in the area, as shown by high topic overlap and low topic distance even in small sample sizes. Areas requiring greater sample sizes could indicate particularly expansive LCCO classes.

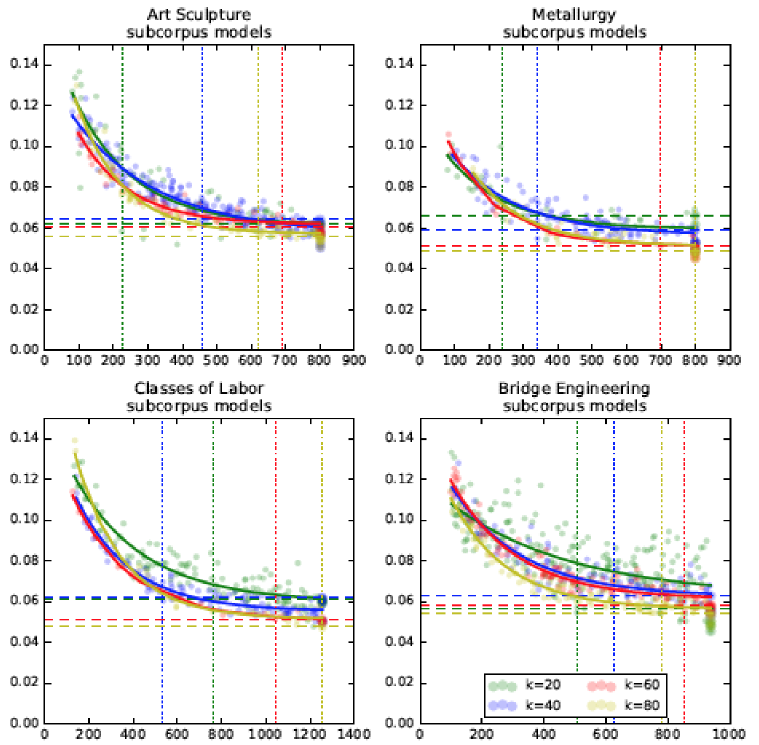


Figure. Topic Alignments. Subcorpus models (k = {20, 40, 60, 80}) for all five selected LCCO subject headings. In general, coarser models (i.e., those with lower number of topics) achieve the worst-whole-corpus-performance with a smaller subcorpus size than fine-grained models (i.e., those with higher number of topics).

Community Impact

The Topic Explorer has become a core component of the HTRC Data Capsule. In addition to our research team’s presentations at the HTRC UnCamp 2015, a tutorial at JCDL 2015, and at the HathiTrust User’s Meeting, the work was presented by the larger HTRC community at the Humanities Intensive Learning and Teaching (HILT) conference in Indianapolis in July.

In addition, the work shows continuing promise for impact. This semester an undergraduate research assistant with the InPhO lab has taken on scaling the experiments to all subject areas in the Library of Congress. On January 31, we will submit a condensed version of this report to the International Conference on Computational Social Science. Finally, we are submitting a proposal to the Institute of Museum and Library Services (IMLS) National Leadership Grants for Libraries Program for federal support to scale up this experiment. The grant pre-proposal is due on February 2nd, with announcements made in August 2016.

Resources

Jaimie Murdock, Jiaan Zeng, Colin Allen. Project report http://arxiv.org/abs/1512.05004

Jaimie Murdock, Jiaan Zeng, and Robert H McDonald. Topic Exploration with the HTRC Data Capsule for Non-Consumptive Research. In JCDL ’15 Proceedings of the 15th ACM/IEEE-CS joint conference on Digital libraries, Knoxville, Tennessee, USA, 2015. ACM Press.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, The Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331(6014):176–182, Jan 2011.

David M. Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, April 2012.