Topic Exploration with the HTRC Data Capsule for Non-Consumptive Research

Jaimie Murdock
Indiana University
jammurdo@indiana.edu

Jiaan Zeng
Indiana University
jiaazeng@indiana.edu

Robert H. McDonald
Indiana University
rhmcdona@indiana.edu

ABSTRACT

In this half-day tutorial, we will show 1) how the HathiTrust Research Center (HTRC) Data Capsule can be used for non-consumptive research over collection of texts and 2) how integrated tools for LDA topic modeling and visualization can be used to drive formulation of new research questions. Participants will be given an account in the HTRC Data Capsule and taught how to use the workset manager to create a corpus, and then use the VM's secure mode to download texts and analyze their contents.

Categories and Subject Descriptors

C.2.4 [**Computer Communication Systems**]: Distributed systems; H.3.7 [**Information Systems**]: Digital libraries; I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

topic modeling, non-consupmtive use, data capsules

1. A NON-CONSUMPTIVE WORKFLOW

As large-scale digitization projects have grown to include both copyrighted and public domain works, legal consensus was built for the "non-expressive use" of text, including text mining over copyrighted works to produce analysis of a large corpus (see *Authors Guild v. HathiTrust*). The HathiTrust Research Center (HTRC)[1] Data Capsule [4] enables this "non-consumptive" use. In this tutorial, we present a five-stage research pipeline for non-consumptive textual analysis: from initial corpus curation to modeling and visualization with tools developed by the InPhO Project[2].

First, a collection of HathiTrust IDs is created using either the plain-text bibliography parser of the InPhO Corpus Builder or the search capacities of the HathiTrust Workset Creator (Figure 1, blue boxes). Then, the HTRC Data Capsule is switched to *secure mode* in which network and file system access is highly constrained to protect copyrighted texts (Figure 1, yellow boxes). At this point, the volumes may be downloaded to the Data Capsule. These are modeled by Latent Dirichlet Allocation (LDA) [1] and visualized through the InPhO Topic Explorer [2] (Figure 1, top). Additional programmatic access to the models is provided by automatically-generated IPython/Jupyter Notebooks [3].

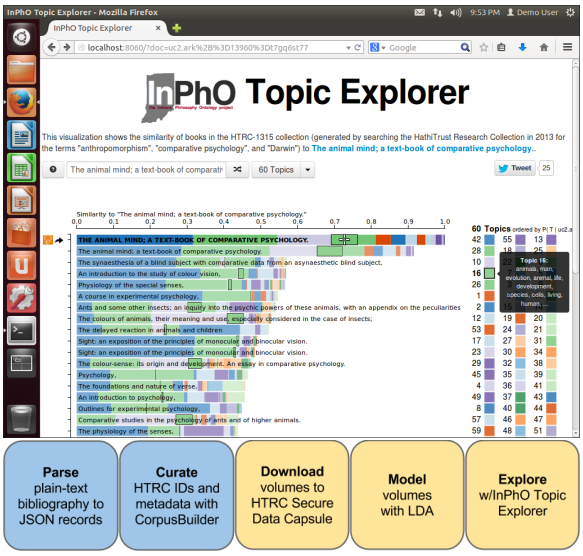

Figure 1: *Top* — The InPhO Topic Explorer in the HTRC Data Capsule. *Bottom* — Topic exploration workflow: blue nodes may be completed in maintenance mode; yellow nodes require secure mode.

[1] `http://hathitrust.org/htrc/`
[2] `http://inpho.cogs.indiana.edu/`

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.
ACM 978-1-4503-3594-2/15/06.
http://dx.doi.org/10.1145/2756406.2756929.

2. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] J. Murdock and C. Allen. Visualization techniques for topic model checking. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[3] F. Pérez and B. E. Granger. IPython: a system for interactive scientific computing, May 2007.

[4] J. Zeng, G. Ruan, A. Crowell, A. Prakash, and B. Plale. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM Workshop on Scientific Cloud Computing*, ScienceCloud '14, pages 9–16, 2014.