Facilitating Large-scale analysis of Scholarly Archives: Hathi Trust Research Center

Beth Plale, Director Data To Insight Center, Indiana University; Chair, HTRC Executive Management Team

HathiTrust is an online repository dedicated to the provision of access to a comprehensive body of published works for scholarship and education (http://www.hathitrust.org). Over 60 universities belong to the HathiTrust community and over 10 million volumes have been ingested into its digital archive from sources including Google Books, member university libraries, the Internet Archive, and numerous private collections. The HathiTrust Research Center (http://www.hathitrust.org/htrc) is dedicated to facilitating scholarship using this enormous corpus through enabling computational access to the corpus, developing research tools, fostering research projects and communities, and providing additional resources such as enhanced metadata and indices that will assist scholars to more easily exploit the HathiTrust corpus.

Why HTRC? Of the 10.2 million volumes in the HathiTrust corpus, 62% are subject to copyright laws and the remaining 38% in the public domain are subject to Google terms of access. With the access controls that exist on the HathiTrust volumes and will exist for some time, complicated legal agreements have to be executed just to get access to the public domain works. HTRC has entered into these agreements with HathiTrust and Google and can provide researchers with computational access to the volumes, saving the researcher lengthy legal delays. Additionally, HTRC has laid groundwork through funding from the Alfred P. Sloan Foundation for research into non-consumptive research, that is, providing ways in which text mining and information retrieval can be carried out on copyrighted materials of the corpus that does not violate fair use terms. Finally, HTRC is working on agreements for sizeable access to hardware resources so that researchers can run their algorithms efficiently and at modest cost using parallel computing in a way that scales well beyond what they can do on their desktop.

HTRC follows the cloud computing model where richer forms of interaction with the HathiTrust corpus can be had when a researcher is willing to move their digital analysis algorithms and tools to the data, thus opening a window to new discoveries. HTRC infrastructure is built using a noSQL volume store cluster with SAN disk access, SOLR indexes, Hadoop, REST interfaces, identity management service (from WSO2), and secure data capsules. It is using compute resources of FutureGrid and at NCSA (Univ of Illinois).

The kinds of applications that will be executed against the HathiTrust corpus are captured at the top level of Figure 1, that is search, trend tracking, network graph algorithms, classification, and simple statistics. Users want to do topic modeling,

A challenge in supporting search is in discovering the correct subset of the data that meets a scholar's research criteria. Traditional searching methodologies, including full text and bibliographic data, have limitation. Full text search is limited by the quality of the underlying data. Uncorrected OCR, while a cost effective technology, can be error-filled. Bibliographic data, while useful in many types of searches such as title, author, and publication details, lacks important information that many humanities researchers need. Gender of author, reliable description genre of materials, and major themes of the works are some of the major pieces of information missing in bibliographic data. Two internal areas of investigation in HTRC are enhancing the amount of metadata available to researchers, and strategies for OCR error correction. The latter is discussed briefly.

OCR error detection study: We undertook a study to quantify OCR errors in the HathiTrust corpus. Scholars are interested in doing quality text analysis, but results can be confounded by OCR errors. Information on which books (or pages) in the collection have significant rates of OCR errors could help. The HTRC explored a couple of approaches to OCR error detection and have results for one approach that uses machine-generated and expert-evaluated rules. Starting with a

large dictionary of correctly spelled words, HTRC members identified outlier words that were in the HathiTrust corpus but not in the dictionary.

As a check on identified words, the rules by which outliers were detected were verified by a human expert. Using this approach, HTRC formulated 48,308 rules that identified outlier words and provided corrections. HTRC members applied the rules to 256,000 non-Google digitized volumes from HathiTrust, which took 4 hours using the National Center for Supercomputing Applications (NCSA) Ember supercomputer. The results showed that the probability of a word having an OCR error (detected by the rule set) was 20%. The average number of errors per page was 0.57. The average number of errors per volume was 156. The probability that a page had one or more errors on it was 11%. The probability that any volume had one or more errors was 84.9%. Overall, 217,754 of the 256,416 volumes had one or more OCR errors and 7,745,034 of the 69,297,000 pages had one or more errors.

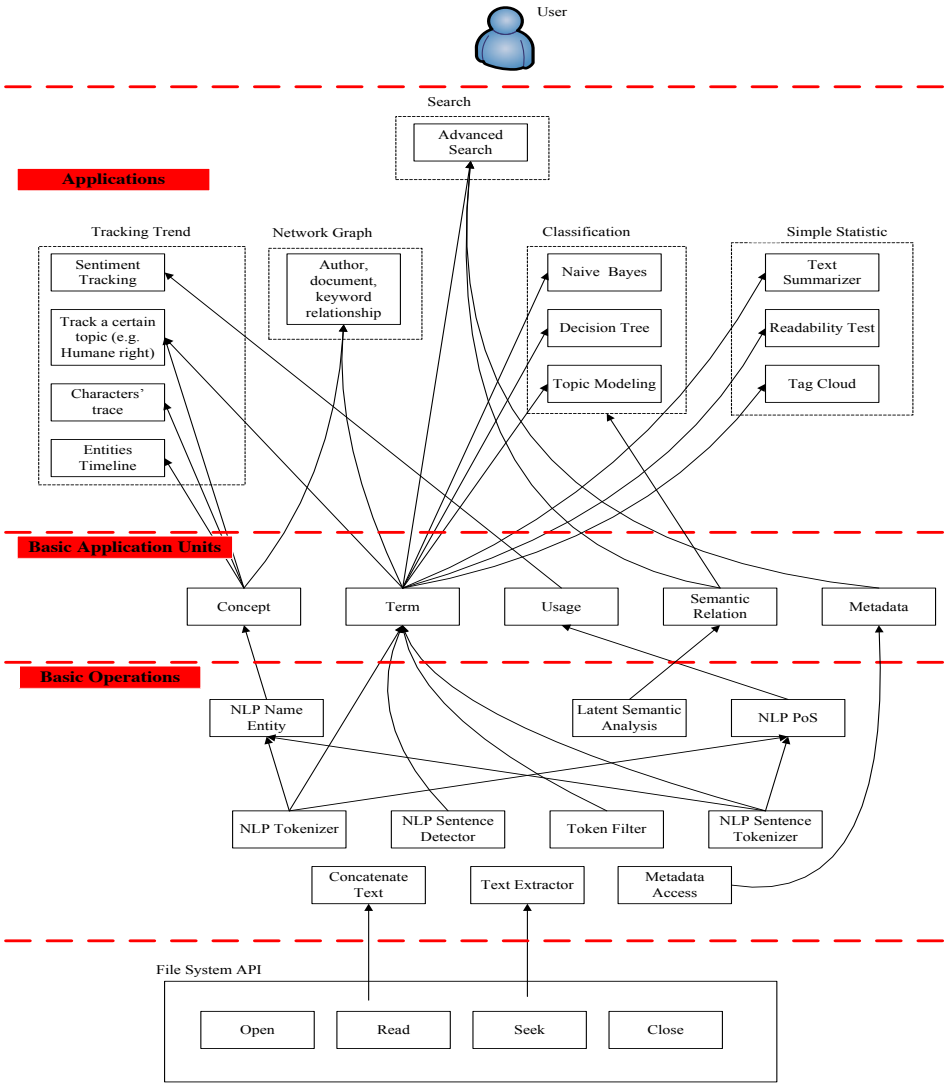

Figure 1. Categorization of Applications across the top of the figure include trend tracking, search, network graph, classification, and simple statistics.