Text analysis tools in progress from the HathiTrust Research Center

Mellon Digital Humanities Seminar, Price Lab for Digital Humanities, University of Pennsylvania.

October 10, 2016.

Sayan Bhattacharyya will briefly describe two ongoing tool-building initiatives at the HathiTrust Research Center (HTRC), the research wing of the HathiTrust Digital Library: the HTRC Bookworm and the HTRC "Extracted Features" functionality. The first tool, the HTRC Bookworm, consists of the generic Bookworm tool, developed by Erez Aiden Lieberman and Ben Schmidt, integrated with the HathiTrust Digital Library (HTDL). This hookup leverages the extensive metadata that enriches the HTDL, enabling motivated visualizations of facets of a corpus. One such motivation, though by no means the only one, is the tracing of individual words over facets across time or across other dimensions. The second tool, HTRC's "Extracted Features", provides users with bags of words (and some other information) per page. This is not only useful for performing text analysis on those texts which cannot be made available as linear, sequential streams of words because of copyright restrictions, but also lowers the cost of processing for those texts which can be. He will argue that, in addition to their utilitarian value, tools such as these may also help problematize such notions as "text" and "reading".