Big-Data Oriented Text Analysis for the Humanities: Pedagogical Use of the HathiTrust+Bookworm Tool

Sayan Bhattacharyya
sayanb@sas.upenn.edu
University of Pennsylvania, United States of America

Christi Merrill
merrillc@umich.edu
University of Michigan, United States of America

Peter Organisciak
organis2@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Benjamin Schmidt
b.schmidt@northeastern.edu
Northeastern University, United States of America

Loretta Auvil
lauvil@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Erez Lieberman Aiden
erez@erez.com
Rice University, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

The HathiTrust Bookworm (HT+Bookworm) is an interactive tool for visualizing content from the HathiTrust Digital Library, which contains almost 15 million volumes of digitized text (Auvil et al. 2015). Our poster describes the application of HT+Bookworm in teaching. Studies show that the complexity of integrating into pedagogical practice text analysis tools operating over large datasets is a significant barrier to their uptake (Green et al. 2016), and that the best solution is to use a pre-populated text analysis tool (Sinclair and Rockwell 2012, Rockwell et al. 2010). HT+Bookworm accomplishes this by facilitating exploration of the "big" textual data of the HathiTrust Digital Library's collection without requiring mastery of complex technology. We recently used HT+Bookworm in class sessions co-taught by two of us as part of literature classes at the University of Michigan, Ann Arbor to undergraduate students who had no prior familiarity with quantitative approaches to text analysis. One of our objectives was to help students discover how the meanings of words can vary — both when word meanings change over time, and when the same word, when borrowed from one discipline or domain and applied to a different discipline or domain (or simply applied independently in two different domains), takes on separate meanings.

Rens Bod argues that it is only when humanistic disciplines are compared on a large scale that patterns across them become visible (Bod 2013). HT+Bookworm enables the discovery of such patterns by facilitating comparison across categories of knowledge. While search engines are adept at finding individual texts within a digital library, Bookworm performs a different task — abstracting across categories within the library and visualizing those abstractions. These abstractions, which are a form of 'distant reading' (Moretti 2013), are generated in the context of a specific textual fragment (for example, a word or phrase), through the variation of some attribute of the manifestation of that text fragment across the categories defined by some categorization scheme. A typical categorization scheme is the organization of the digital library by Library of Congress (LoC) classes as metadata, and a typical attribute of the manifestation of a word or phrase across categories is its normalized frequency of occurrence across those categories. HT+Bookworm works with both discrete sets of categories such as LoC classes as well as with continuous categories such as time. A time-series plot of the normalized frequency of occurrence of a word over a chronological range, within certain specified LoC classes, provides a sense of how the relative occurrence of that word or phrase has varied across those LoC classes over the specified time range. Students generate visualizations consisting of layered time-series plots (stacked area charts) for the relative frequency of their words of interest, within determinate categories of interest in the HathiTrust Digital Library collection. These categories correspond to the layers (stacks) of the plot, with each layer encompassing an LoC class and time represented along the x-axis. The HT+Bookworm tool also provides, at each point in the plot, a subset list of volumes that contribute the most to the attribute being plotted. This list, accessible by mouse-click at the requisite point, serves as the gateway to the digitized text of the individual volumes in the list. This affordance helps students bridge the gap between the abstraction of distant reading and the discovery of specific texts that they can then investigate further through close reading.

Our poster includes instances of the kinds of exploration HT+Bookworm made possible for students. An example follows. The concept of "fidelity" (an important word to explore in connection with translation studies, a topic of the classes) shows different characteristics when explored in English (in which the concept maps onto the two words "fidelity" and "faithfulness") and in Spanish (where the concept maps onto the single word "fidelidad"). Investigating the occurrence of the word by LoC category allows students to explore hypotheses such as whether the greater strength, historically speaking, of religious tradition in the Spanish-speaking world in comparison with the Anglophone world affects the relative prevalence of this word in different domains of use. Another example is a stacked area chart for a word of a kind for which HT+Bookworm helps provide an understanding of the word's differentiated meanings in different use categories (for example, in the case of the word "depression", in the use category of psychology and medicine versus that of economics). HT+Bookworm accomplishes this by abstracting separately across different LoC classes, while aiding in the indication of the points in time at which, for each class, the word entered widespread usage.

Bibliography

Auvil, L., Aiden, E. L., Downie, J.S., Schmidt, B., Bhattacharyya, S. and Organisciak, P. (2015). "Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project." Digital Humanities 2015 (DH 2015) Conference, Sydney, Australia. 29 June - 3 July 2015.

Bod, R. (2013). A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present. New York: Oxford University Press.

Green, H., Dickson, E. and Bhattacharyya, S. (2016). "Scholarly Requirements for Large Scale Text Analysis: A User Needs Assessment by the HathiTrust Research Center." Digital Humanities 2016 (DH 2016) Conference, Krakow, Poland. July 2016.

Moretti, F. (2013). Distant Reading. London: Verso.

Sinclair, S. and Rockwell, G. (2012). "Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies." In Brett D. Hirsch (ed.), Digital Humanities Pedagogy: Practices, Principles and Politics. Cambridge, U.K.: OpenBook Publishers, pp. 241-64.

Rockwell, G, Sinclair, S., Ruecker, S. and Organisciak, P. (2010). "Ubiquitous Text Analysis." paj: The Journal of the Initiative for Digital Humanities, Media, and Culture, Vol. 2, No. 1.