The HathiTrust+Bookworm project

Team Members:
– Current:
  J. Stephen Downie, University of Illinois, Urbana-Champaign
  Erez Lieberman Aiden, Rice University/Baylor College of Medicine
  Benjamin Schmidt, Northeastern University
  Robert McDonald, Indiana University
  Loretta Auvil, University of Illinois, Urbana-Champaign
  Peter Organisciak, University of Illinois, Urbana-Champaign
  Muhammad Shamim, Rice University/Baylor College of Medicine
  Sayan Bhattacharyya, University of Illinois, Urbana-Champaign
  Leena Unnikrishnan, Indiana University

– Past:
  Colleen Fallaw, University of Illinois, Urbana-Champaign
  Matt Nicklay, Baylor College of Medicine

Funded by a National Endowment for the Humanities (NEH) Implementation Grant (September 2014-August 2016)
Current Prototype

• The current prototype of HT+Bookworm is set up to work with ~5.2 million public domain volumes in the HathiTrust Digital Library

• Go to htrcbookworm.wordpress.com
  – Then click on “Try the prototype”
First Advantage: Worksets

• HTDL has a rich notion of workset
  – A workset is a user-determined sub-collection from the HTDL’s collection
    • so called because it is a set of books meant to be worked on by algorithms for text analysis

• HT+Bookworm leverages the workset functionality
  – HT+Bookworm will eventually:
    • query a specific workset (rather than the entire collection)
    • generate a specific workset from the result of a faceted query
First Advantage: Worksets

- Synergy with HTDL’s workset functionality

Workset creation and refinement with HT+Bookworm

Worksets are persistent subcollections of the HTDL
Second Advantage: HT Metadata!

• The HTDL has good and detailed metadata (relatively speaking)!
  – Historically, metadata was meticulously created by librarians at HTDL’s contributing libraries
  • allows for highly faceted queries
Meeting Point of Scientific and Humanistic Inquiry

Scientific Inquiry
• *Generalization* across entities
• Discovery of *patterns* across entities
• A *determinate* epistemology

Humanistic Inquiry
• Close engagement with *specific* entities
• Attending to *singular* instances among entities
• A *skeptical* epistemology
Three Use Cases

How, *across time*

1. change in social context correlates with change in preponderance of one word-concept over another

2. occurrences of related word-concepts in multiple languages/places compare with each other

3. domains/contexts of words vary
Case #1

Change in social context correlates with change in preponderance of one word-concept over another
Case #1

Change in social context correlates with change in preponderance of one word-concept over another
Case #2

How occurrences of related word-concepts in multiple languages/places compare with each other
Case #2

How occurrences of related word-concepts in multiple languages/places compare with each other
Case #3

How domains in which a word occurs vary across time

• Not yet integrated into the public-facing prototype

• To see the pilot version, go to: bit.ly/1PzkcUK
  – Enter the word you want to search for in the field marked “word limited to” and press “return”
  – Click the “Show Advanced” button
  – Do not change the contents of the “x representing”, “y representing”, etc. fields
Case #3

How domains in which a word occurs vary across time "depression" with Library of Congress categories, 1850-1923
Higher resolution: bit.ly/1QatsPs Streamgraph: bit.ly/1llytgo
Taking a Step Back

What is the overall Bookworm philosophy?

• Normal search engines are very good at finding *individual* texts in a library

• Bookworm, however, is good at *something* else:
  – Finding and understanding *categories* in a library
    • The category is plotted along the x-axis
    • The simplest use case is: plotting the usage of a *word* across *years*:
      – Goal: to understand something about the x-axis (the category) through the usage of that word
    • The x-axis can also be defined by *ordinals* (i.e. discrete categories)
      — e.g. “languages” along the x-axis — rather than *cardinals*

Adapted from Ben Schmidt, “Bookworm API Philosophy” – [bit.ly/1K3oUDF](http://bit.ly/1K3oUDF)
Plotting Against an Ordinal Category

Heatmap color represents: Words-per-million, for the word “annul”
X-axis: Four binned book-page-length categories (small, medium, large, extra-large).
Library of Congress classes “History of America” and “Law” show the most incidences of “annul”.
Law books are often thick doorstoppers!
Characterizing a Library’s Contents

Coverage of Lib. of Congress classes in the HathiTrust Library
Notice: Many classes are sparse for earlier years!
The API can be used to understand HTDL’s entire content.
Try it out at: bit.ly/1P5hwtc
Animated maps via Bookworm API

• Go to bit.ly/1Jjt1R1
• Scroll down to see three successive examples:
  – Number of books published (in the USA) in the HathiTrust Digital Library over time
    • Animation unfolding over places of publication within the USA
  – Number of books published (in Europe), in the HathiTrust Digital Library over time
    • Faceted by language
    • Choose language from the drop-down menu
  – Geography of word use:
    • Type the word in to see where-all it was being used (through displaying place of publication)
    • Click the map to access individual instances of the books
Animated maps via Bookworm API

- Number of books published (in the USA) in the HathiTrust Digital Library over time
- Animation unfolding over places of publication within the USA: bit.ly/1Jjt1R1
Underlying Data for HT+Bookworm?

• HT+Bookworm is powered by a public API that follows the Bookworm querying language

• Documentation: 
  bookworm-project.github.io/Docs/API.html