Getting Started With the HathiTrust Research Center
Francesca Giannetti, Digital Humanities Librarian

The goal of this workshop is to learn about the HathiTrust Research Center (HTRC) and familiarize yourself with the tools and resources of the HTRC portal.

LEARNING OBJECTIVES

1. You will know how to create and edit a workset - a custom collection of HathiTrust works - using the Workset Builder.
2. You will learn the functions of some of the algorithms, or analytic tools, in the portal, and how to run them against your worksets.
3. You will explore the HT+Bookworm interface for the purposes of generating research questions, or for deepening your understanding of the outputs of the algorithms.

Logging In

Please go to the HTRC portal at https://htrc2.pti.indiana.edu/ and sign in to follow along with the demos in the first part of today's workshop.

If you haven’t already created your personal account, please go to the portal at https://htrc2.pti.indiana.edu/ and click Sign Up. Use an institutional e-mail address to create your account.

HTRC Portal

As you’ve noticed, the site of the HTRC Portal is separate from www.hathitrust.org, but the main site does link to the portal at http://www.hathitrust.org/htrc.

There are two main parts to working in the portal. The Workset Builder is where you create the collection of texts that you will analyze. Currently, you can select from a corpus of 3 million public domain volumes. The Algorithms are the analytical tools you will use to generate computational analyses of your worksets.

Algorithms

Let’s click on Algorithms in the top navigation bar. Select the algorithm Meandre_Tagcloud by clicking on Execute. I suggest that you pick a workset that is not too big (2-10 texts), like Collins_Wilkie@fgiannetti, so that you can get results sooner.

Discussion: Noisy? How to reduce that noise?
Next let’s try Meandre_Tagcloud_with_Cleaning on Collins_Wilkie@fgiannetti. What does cleaning mean? In this case:

1. Header/Footer removal
2. Joining hyphens across lines of text
3. Upper/lowercase normalization
4. OCR cleaning
5. British-American and period spelling normalization
6. Removal of stop words

And in other cases: removal of front matter, indexes, commentary and/or footnotes.

If the concept of a stop word is new to you, open this English stop word list to explore: http://repository.seasr.org/Datasets/Text/common_words.txt. You will notice that Meandre_Tagcloud_with_Cleaning has a parameter for choosing a custom stop word list. How might you improve this tag cloud by editing the list of stop words?
Topic Modeling is useful for getting a sense of the contents of your workset. This algorithm creates a list of “topics” from the workset. (A “topic” is simply a set of words that have a high probability of co-occurring within the material of the workset.) The topics are displayed as tag clouds.

Demo: I assembled a workset of early (published in or before 1800) medical texts that contain the keyword tea. (Alas, I was really hopeful my favorite caffeinated beverage would migrate to the top of one of those topics.) Apply the topic modeling algorithm to the tea_medicine_early@fgiannetti workset. Note that this is a larger workset, and so it is unlikely that the algorithm will run to completion in a timely manner for you.

Discussion: What topic clusters emerge? Are there surprises? Points of interest? Objects of further study?

Are there areas of further refinement? Strategies for this?

Can we take a stab at topic labeling? Are there junk topics?

This algorithm identifies words that are significantly over- and under-represented in one workset (the “analysis” workset) as compared to another workset (the “reference” workset). It is useful for comparing and contrasting two worksets. The resulting sets of words are visualized as tag clouds.

Demo: Compare works of music history and criticism published in the United States to works of music history and criticism published in the United Kingdom in the period 1750-1923.

Use us_music_hist_crit@fgiannetti as the analysis workset and uk_music_hist_crit@fgiannetti as the reference workset. Then switch them!
On the left are the words that are over-represented in works published in the US, while on the right are words much more likely to appear in works published in the UK.

What are some observations or speculations we could make based on these two word clouds?

Further investigation of observed phenomena using HTRC Bookworm at http://bookworm.htrc.illinois.edu/

Why is it that the British are more likely to use first person plural cases (e.g. “we,” “us,” “our,” but also “them”)?

Why and in what contexts do the Americans emphasize pedagogy so much (e.g. “technic,” “conservatory,” “teacher,” “studied”)?

By the way, Ben Schmidt offers a terrific and fascinating explanation of Dunning’s log-likelihood statistic using Library of Congress classes E and F (history) in a blog post.

Meandre_OpenNLP_Date_Entities_To_Simile

Time permitting, we’ll also look at this timeline algorithm (ha). This algorithm scans your workset for date entities and uses the SIMILE API to display them in a timeline. The timeline displays the volume ID, page number and sentence number in which the date entity appears.

The HathiTrust Solr API provides a quick way to retrieve the rest of the volume metadata, so you understand in which work the date appears. For example, typing this URL in the address bar of your browser and clicking return will pull a condensed version of the metadata of volume nyp.33433014054591.

http://chinkapin.pti.indiana.edu:9994/solr/meta/select/?q=id:nyp.33433014054591
Workset Builder

Exercise: Now it's time to build your own workset! Go back to the portal at https://htrc2.pti.indiana.edu/. Click on Worksets > Create Worksets and log in again with the same credentials you used for the portal.

Use the search interface to retrieve volumes based either on bibliographic metadata (author, subject, place of publication, etc.) or full text using keywords or key phrases. Select the items you want either by manually clicking on the Select box, or by clicking on “Select All Items On the Page” or “Select All Items.” Then navigate over to “Selected Items” in the upper right menu. In that window, you should see the option to “Create/Update Workset.” Click on it. Later, you can add to your workset as you find more texts.

Exercise: Next, let’s run an algorithm on your workset. Go back to the HTRC portal (https://htrc2.pti.indiana.edu/) and click on Algorithms in the upper navigation menu.

Click on the desired algorithm. I suggest Meandre_Tagcloud or Meandre_Topic_Modeling. Enter a name of your choosing in the blank field for Job Name. This is the same name that will appear when you are looking at the results.

How did you decide to focus your workset? Did you have any challenges? What algorithm did you execute? What did the results reveal about your texts? Was there noise? What do you think it's from?
Part of a topic model of Ann Radcliffe’s novels and short stories.

Questions?

Francesca Giannetti, Digital Humanities Librarian, francesca.giannetti@rutgers.edu

Get Involved!

HTRC Announcements:
htrc-announce-l@list.indiana.edu

HTRC User Group:
htrc-usergroup-l@list.indiana.edu

Acknowledgements

I am especially grateful to Harriett Green and Sayan Bhattacharyya of the University of Illinois at Urbana Champagne, and to Michelle Paolillo of Cornell University for their generosity in sharing their instructional materials.