Jeffrey Tharsen tharsen@uchicago.edu
"Digital Textual Analysis for Large-Scale Repositories (HathiTrust and others)"
Research Computing Center Workshop
Winter 2019

Wikipedia's "List of Digital Library Projects": *https://en.wikipedia.org/wiki/List_of_digital_library_projects*

Sources: *http://home.uchicago.edu/~jcarlsen/Hathi-sources.zip*

I. The HathiTrust Digital Library (HTDL) vs. the HathiTrust Research Center (HTRC)

a. The HTDL is hosted at UMichigan, and can be accessed via:
https://www.hathitrust.org and/or https://babel.hathitrust.org

This is where you can build collections (in the Full-Text Search or Item View only), and also see the entire catalogue of holdings (including restricted items).

b. The HTRC is hosted at UIUC in Urbana-Champaign, and can be accessed via:
https://bookworm.htrc.illinois.edu and/or
https://analytics.hathitrust.org

II. Direct access to the HTDL from the UChicago Library Catalogue

a. In our catalogue at present, we have 1,937,210 items that link directly to the HTDL viewer; these items can also be downloaded as PDFs with embedded OCR'ed text, or as plain text only. They can also be requested *en masse* from the HTDL.
For example: https://catalog.lib.uchicago.edu/vufind/Record/ocm01693195/Holdings#tabnav
=> https://babel.hathitrust.org/cgi/pt?id=hvd.hn2bd5;view=1up;seq=7

III. Creating a Collection using the HTDL

a. Browse to https://www.hathitrust.org/
b. Log in (uses Shibboleth authentication, so your CNet & password should work here)
c. Select the "Full-Text" Search tab
d. Perform searches, and when ready, "Add Selected" to a Collection

IV. Using Worksets & Basic Algorithms in the HTRC

a. Start with Bookworm: https://bookworm.htrc.illinois.edu .

b. Browse to https://analytics.hathitrust.org/
c. Sign Up or Sign In (the HTRC requires a unique login and password)

d. For Metadata or "Bag of Words" analyses, the HTRC Extracted Features Dataset is probably the easiest to use, and can be downloaded in its entirety:
https://analytics.hathitrust.org/datasets

e. **Built-in Algorithms** ("Analyze With Algorithm") : InPhO Topic Model Explorer, Named Entity Recognizer, and Token Count and Tag Cloud Creator) can be executed directly on any list of volume IDs; at the HTRC this list is called a "Workset".

V. Using the Data Capsule: Bringing in a Collection and Running Custom Code on it

a. To create your Data Capsule, browse to: https://analytics.hathitrust.org/capsules and click "Create A Capsule". 2 VCPUs and 4GB of memory is a fine default. Click "Create Capsule". (**Demo Capsules** have access to Hathi Public Domain works only.)

b. When it's been created, go to "Capsules", click on the Data Capsule ID and/or choose "Start Capsule". It normally takes about 5-10 minutes to start.

c. It will be started in **Maintenance Mode** – in the screen you can access via the Data Capsule ID, click "Connect via Remote Desktop" and your Data Capsule VM will be displayed in the lower half of the browser window.

   Use the browser or other connections while in Maintenance Mode to set up your environment, and then switch to **Secure Mode** to bring in the sources and run your analyses.

d. The Data Capsule uses Ubuntu Linux, so much of the work is done using the Terminal.

   Here are some useful Terminal commands (**Secure Mode only**):

```
htrc export "https://babel.hathitrust.org/cgi/mb?a=listis&c=1114599640" > volume-list.txt

htrc download volume-list.txt -c -o myvolumes_fulltext
```

   (The **-c** flag here tells it to download each volume as a text file, otherwise it will be a directory in which each page is its own text file, and the **-o** flag tells it where to put them, this will create a "myvolumes_fulltext" directory and store all the text files there. You can also create a corpus in a single text file by using the "cat" command: `cat muvolumes_fulltext/* > myvolumes_fulltext.txt`).

   Documentation for the **`htrc`** package is available here:
   https://htrc.github.io/HTRC-WorksetToolkit/
   (As it now comes pre-installed, you can skip the installation instructions.)

   At this point, I would recommend either looking at your volumes using the Voyant Tools browser-based system (it takes a minute or so to spin up and will fail at first, be patient), or begin running custom analyses on your texts/corpus/corpora using python, R or other tools.

   ** When working in Secure Mode, I strongly recommend that you not shut down the Capsule or switch to Maintenance Mode, as your texts/corpus/corpora will then be deleted.*