MODULE 4.1 Performing Text Analysis: Using Off-the-Shelf Tools

KEY TOOLS

HTRC algorithms
A set of off-the-shelf text analysis algorithms provided via HTRC Analytics for users to analyze their worksets, such as algorithms for extracting named entities and doing topic modeling.

ACTIVITY: Review algorithm descriptions

☞ ***Slide M4.1 – 13***

Link to algorithm descriptions: https://wiki.htrc.illinois.edu/x/HoJnAQ

In pairs or small groups, please read the descriptions for the following the HTRC algorithms. *Can you explain to one another what each of them does? What kind of research questions might they help answer?*

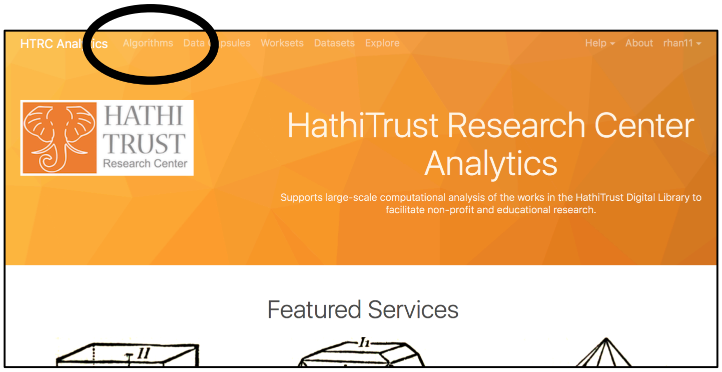
Tool	What does it do?	Example research question?
Token Count and Tag Cloud Creator		
Named Entity Recognizer		
InPhO Topic Model Explorer		

ACTIVITY: Running an algorithm in HTRC Analytics

☞ ***Slide M4.1 - 15***

Let's try performing a popular text analysis method, topic modeling, using a web-based tool.

1. From the homepage of HTRC Analytics, click "Algorithms."

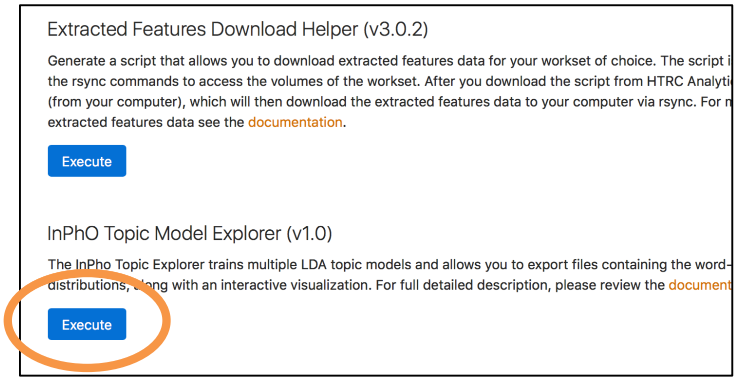


HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

Featured Services

2. Click on the "Execute" button under the name and description of the algorithm you want to run. Select "InPhO Topic Model Explorer (v1.0)" for this activity.



Extracted Features Download Helper (v3.0.2)

Generate a script that allows you to download extracted features data for your workset of choice. The script i
the rsync commands to access the volumes of the workset. After you download the script from HTRC Analyti
(from your computer), which will then download the extracted features data to your computer via rsync. For n
extracted features data see the documentation.

InPhO Topic Model Explorer (v1.0)

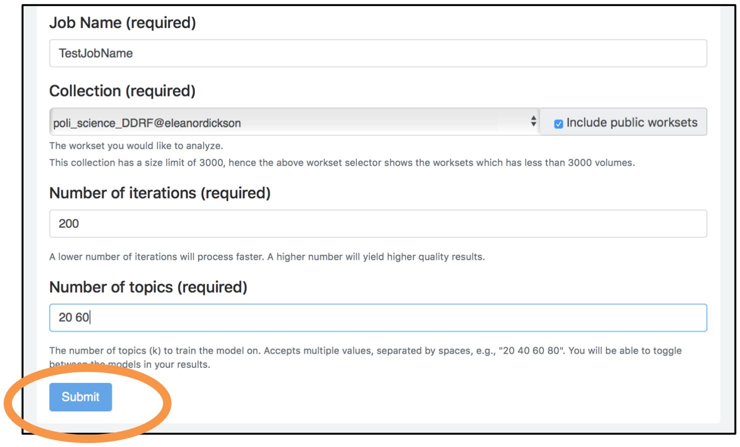
The InPho Topic Explorer trains multiple LDA topic models and allows you to export files containing the word-
distributions, along with an interactive visualization. For full detailed description, please review the document

3. Choose a workset from either all worksets or just your private worksets.

4. For this example exercise, check the "Include public worksets" option and select "**poli_science_DDRF@eleanordickson**".

5. To navigate to the workset more quickly, after clicking on the arrow button to expand the list of worksets, type "EF" and the down arrow and the workset that we need will appear at the bottom of the list.

6. Enter a name for your job, type "200" for the number of iterations, and type "20 60" for the number of topics to be created. Click "Submit."



Job Name (required)

TestJobName

Collection (required)

poli_science_DDRF@eleanordickson ☑ Include public worksets

The workset you would like to analyze.

This collection has a size limit of 3000, hence the above workset selector shows the worksets which has less than 3000 volumes.

Number of iterations (required)

200

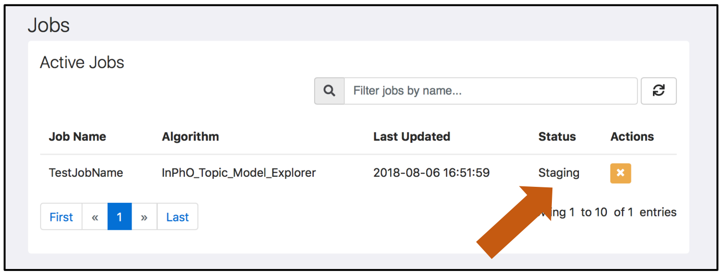
A lower number of iterations will process faster. A higher number will yield higher quality results.

Number of topics (required)

20 60

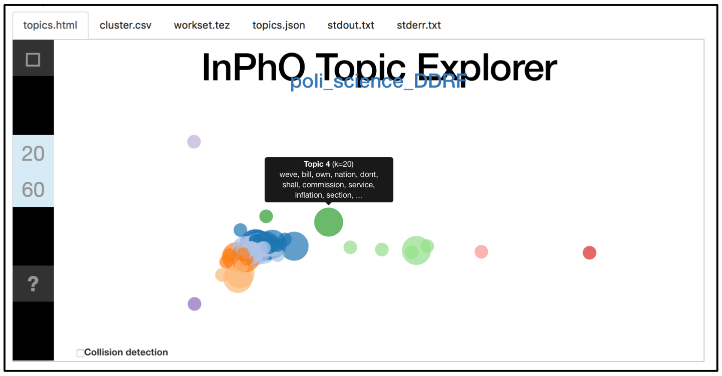
The number of topics (k) to train the model on. Accepts multiple values, separated by spaces, e.g., "20 40 60 80". You will be able to toggle between the models in your results.

7. See the current job in “Active Jobs” and refresh your screen to see the status change.
8. You may have to be patient while it finishes, especially if the workset is large.
9. Once the job is done, it will be listed under “Completed Jobs.”

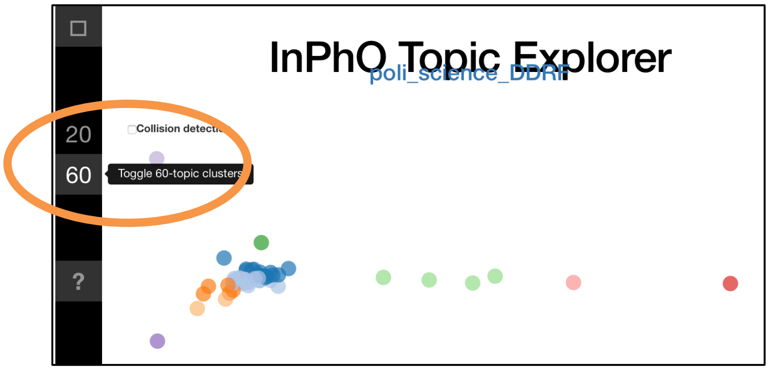


Job Name	Algorithm	Last Updated	Status	Actions
TestJobName	InPhO_Topic_Model_Explorer	2018-08-06 16:51:59	Staging	✖

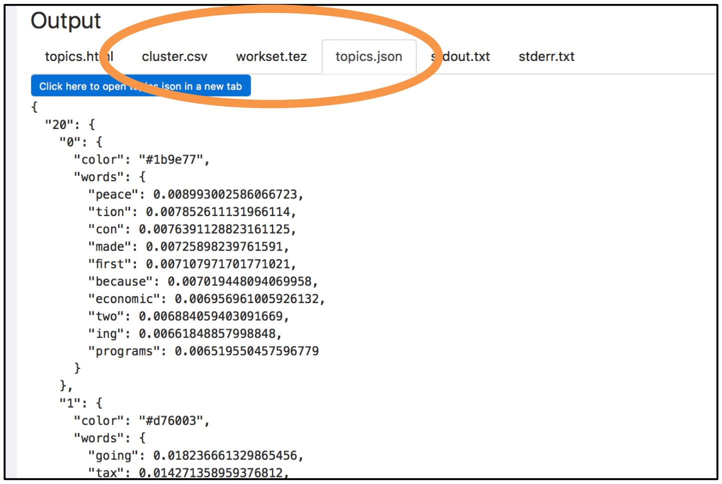
10. Click on the job name to see the results. Scroll to the “output” area to see the bubble visualization of the generated topics. Hover over a bubble to see the top terms in a topic.



11. The numbers on the side relate to the number of topics generated, as do the size of the bubbles. Toggle the display of the n-topic clusters by clicking on those numbers.



12. You can also view and download 3 results files: topics.json, cluster.csv, and workset.tez. These files can be used to play with the visualization in more depth outside HTRC Analytics.



```
{
  "20": {
    "0": {
      "color": "#1b9e77",
      "words": {
        "peace": 0.008993002586066723,
        "tion": 0.007852611131966114,
        "con": 0.0076391128823161125,
        "made": 0.00725898239761591,
        "first": 0.007107971701771021,
        "because": 0.007019448094069958,
        "economic": 0.006956961005926132,
        "two": 0.006884059403091669,
        "ing": 0.00661848857998848,
        "programs": 0.006519550457596779
      }
    },
    "1": {
      "color": "#d76003",
      "words": {
        "going": 0.018236661329865456,
        "tax": 0.014271358959376812,
```