**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***

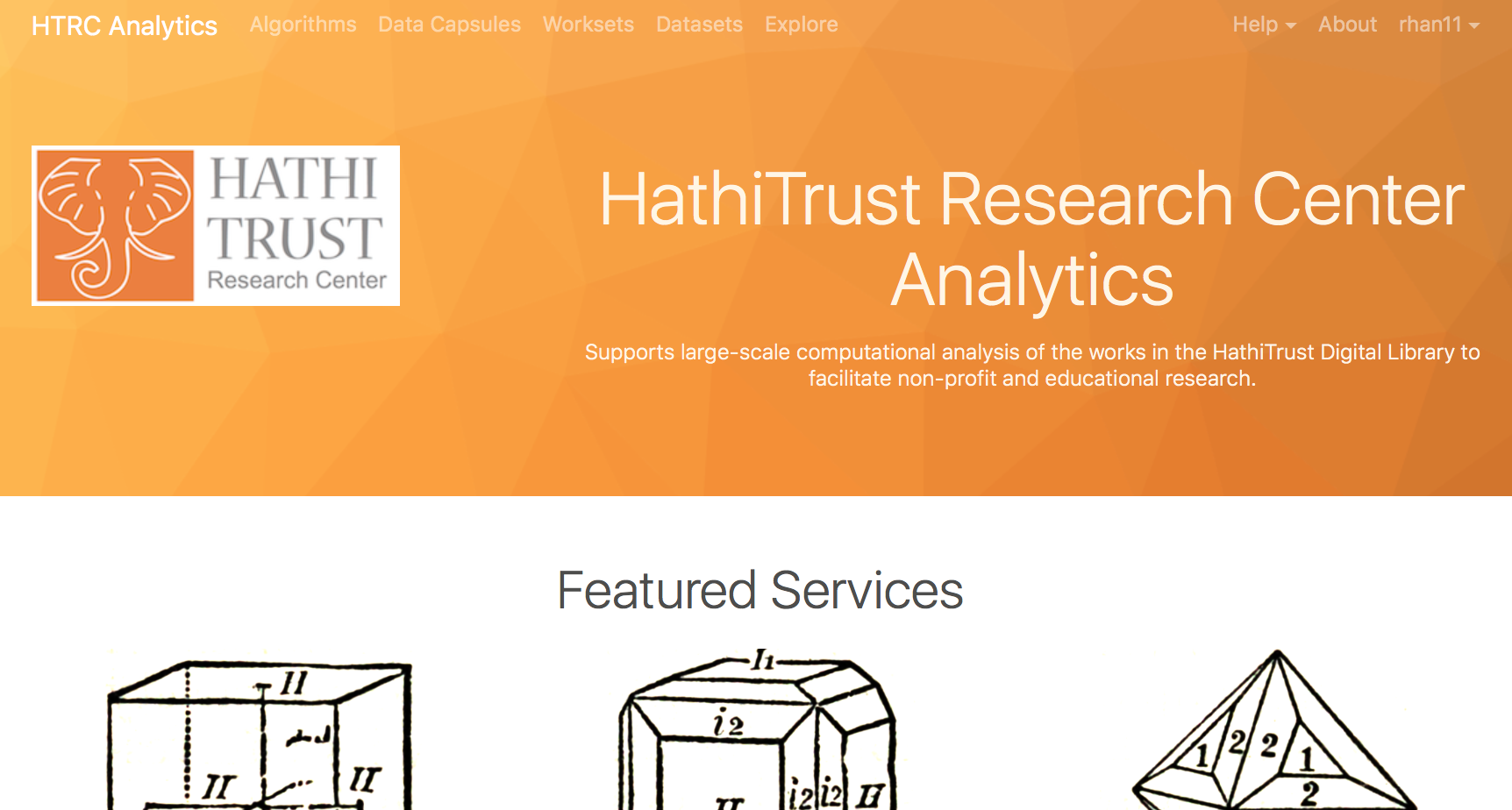
Link to algorithm descriptions: <https://wiki.htrc.illinois.edu/x/HoJnAQ>

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?*

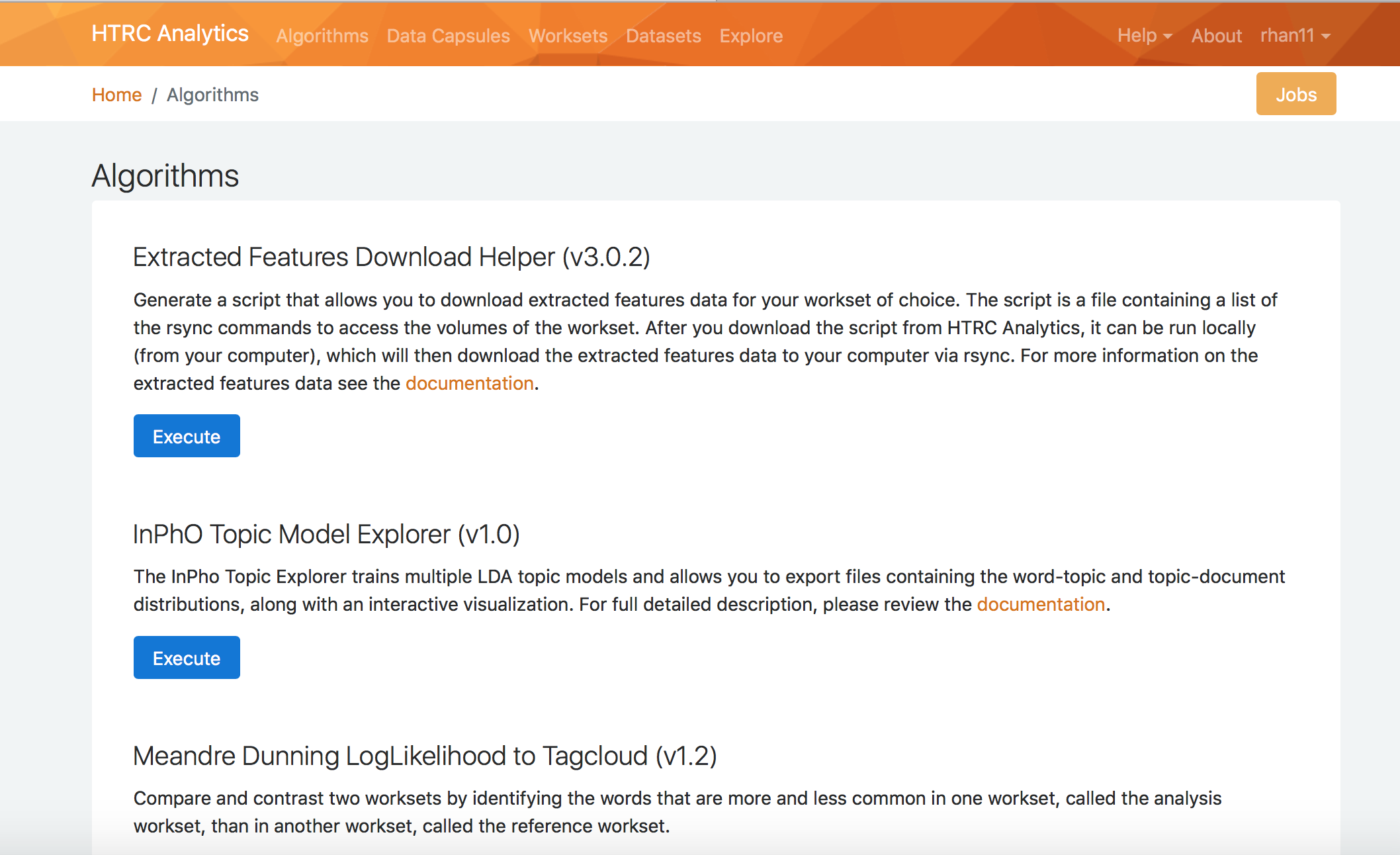
|  |  |  |
| --- | --- | --- |
| **Tool** | **What does it do?** | **Example research question?** |
| **Token Count and Tag Cloud Creator** | **Creates a tag cloud of the most frequently occurring words.** | **Identify themes.** |
| **Named Entity Recognizer** | **Generates a list of people, places, dates, times, percentages, monetary terms.** | **Determine geographic extent of a text. Genealogy, social network overlap.** |
| **InPhO Topic Model Explorer** | **Generates "topics" based on words that have a high probability of occurring close together in the text.** | **Identify topics in discussions on social media** |

**ACTIVITY: Running an algorithm in HTRC Analytics** ☞ ***Slide M4.1 - 15***

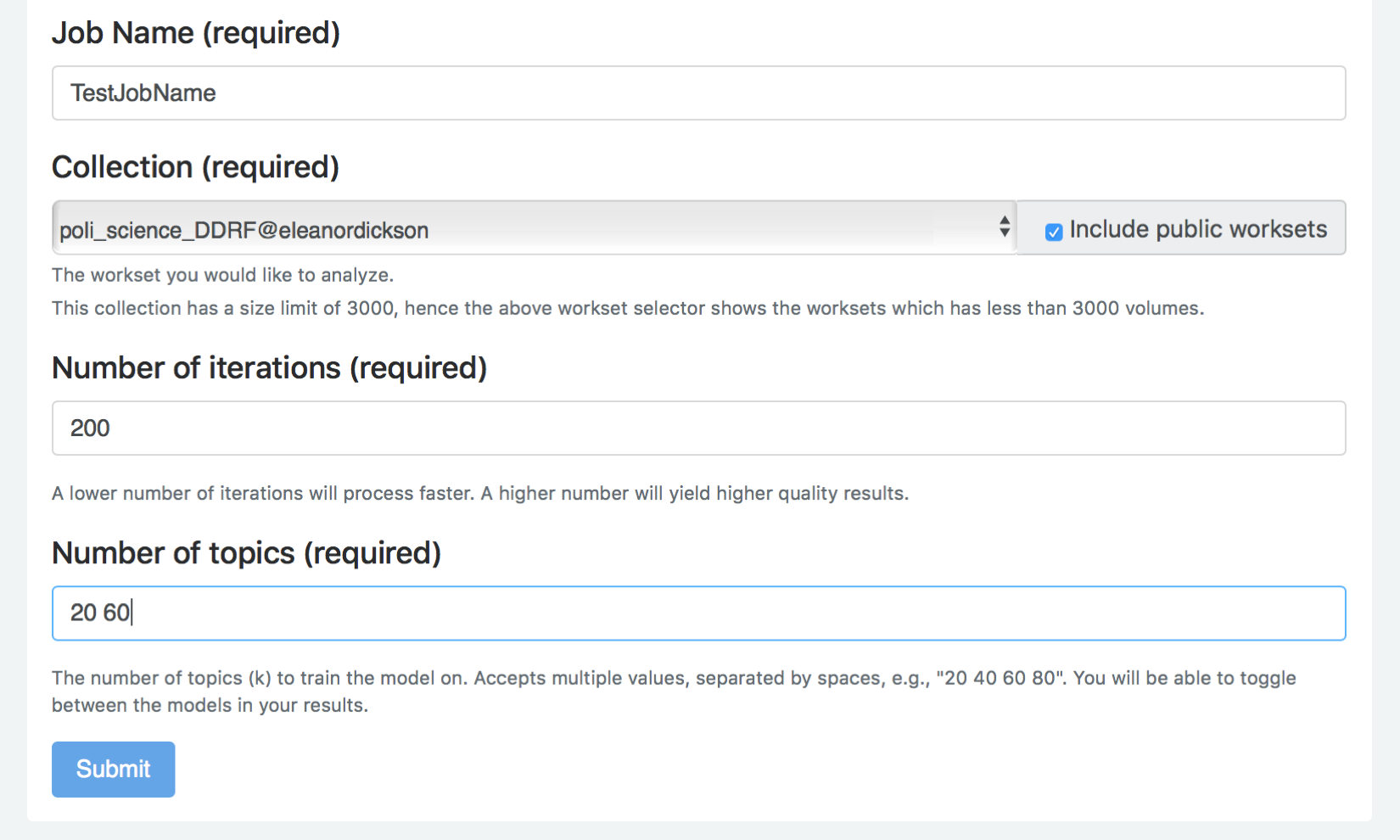
*Let’s try performing a popular text analysis method, topic modeling, using a web-based tool.*



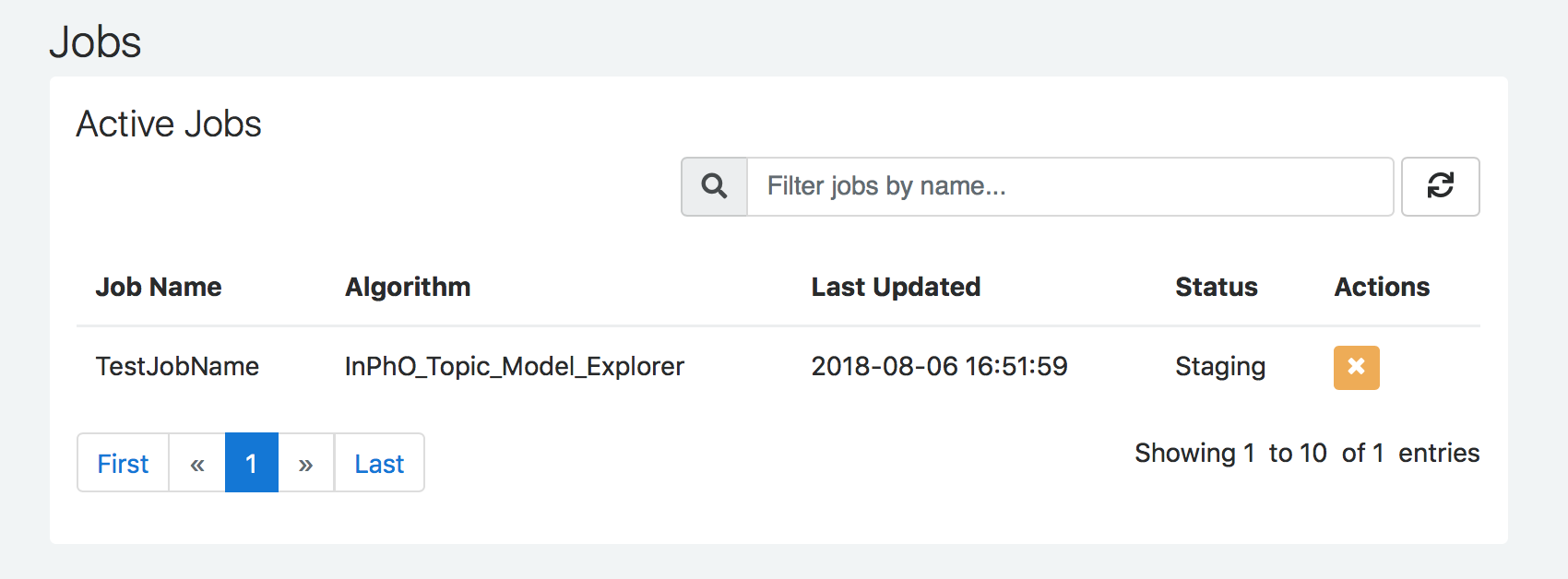
1. From the homepage of HTRC Analytics, click “Algorithms.”
2. Click on the “Execute” button under the name and description of the algorithm you want to run. Select “InPhO Topic Model Explorer (v1.0)” for this activity.



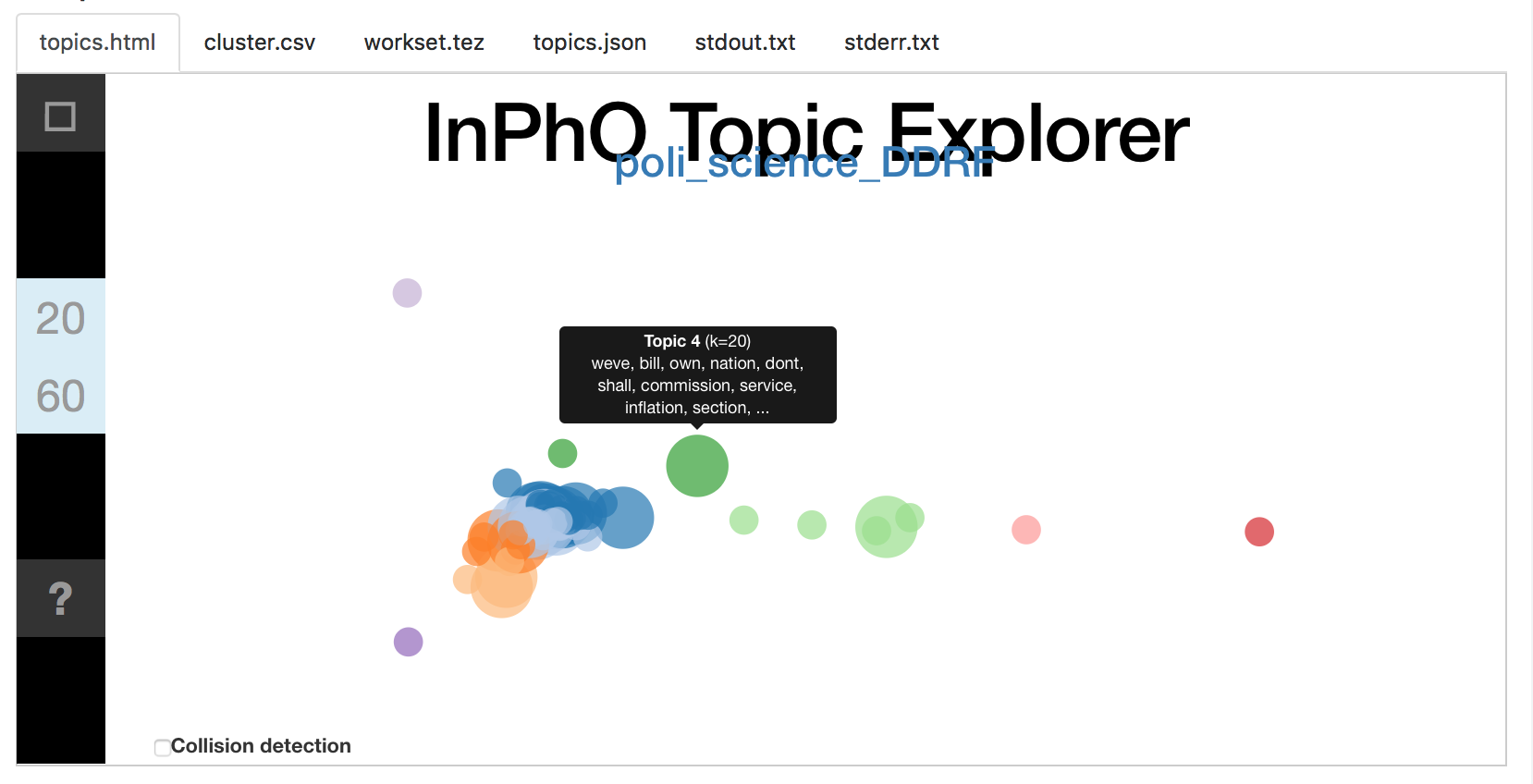
1. Choose a workset from either all worksets or just your private worksets.
2. For this example exercise, check the “Include public worksets” option and select “**poli\_science\_DDRF@eleanordickson**”.



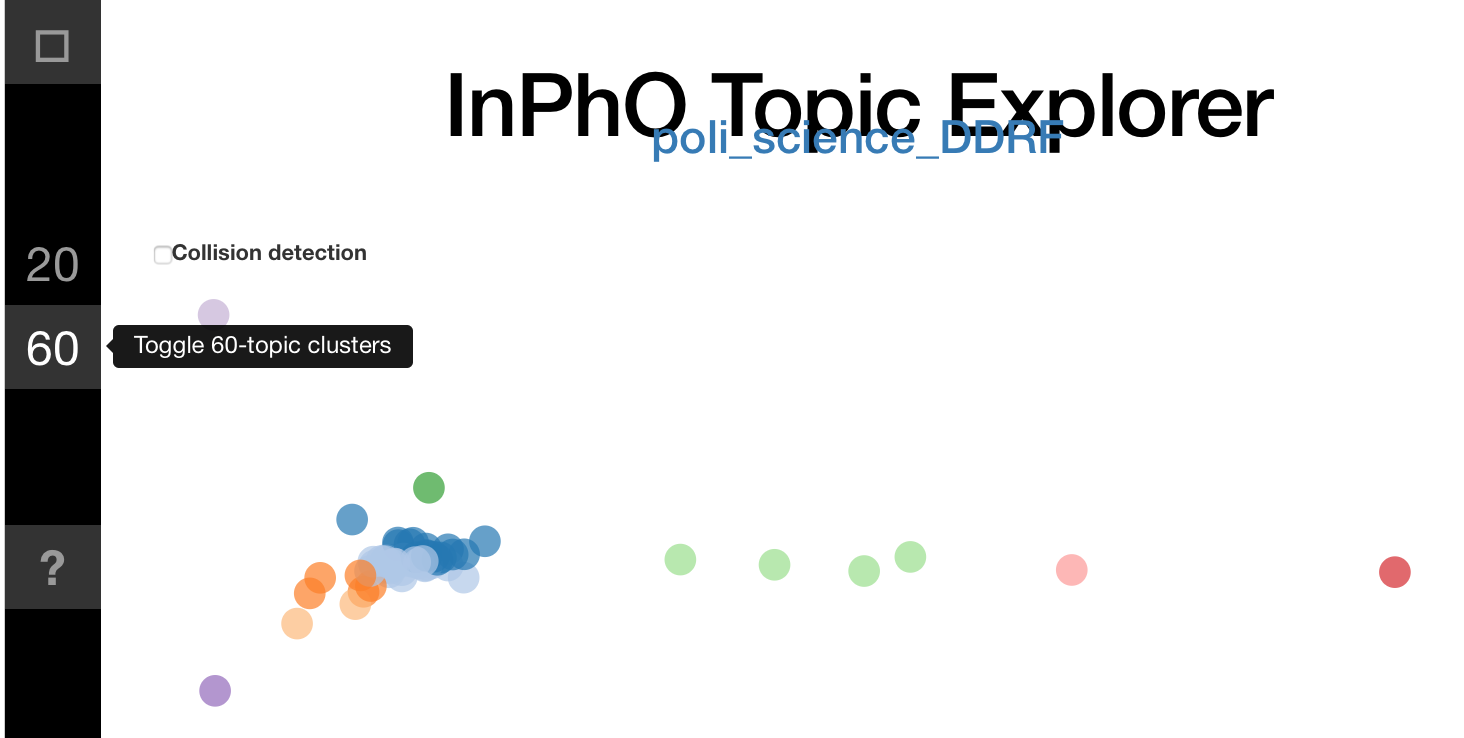
1. To navigate to the workset more quickly, after clicking on the arrow button to expand the list of worksets, type “EF” and the down arrow and the workset that we need will appear at the bottom of the list.
2. Enter a name for your job, type “200” for the number of iterations, and type “20 60” for the number   
   of topics to be created. Click “Submit.”
3. See the current job in “Active Jobs” and refresh your screen to see the status change.



1. You may have to be patient while it finishes, especially if the workset is large.
2. Once the job is done, it will be listed under “Completed Jobs.”



1. 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.



1. 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.
2. 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.

