Digging Deeper,
Reaching Further

Module 4.1: Analyzing Textual Data

Using Off-the-Shelf Tools

In this module we'll…

- Weigh the benefits and drawbacks of pre-built tools for text analysis

  → *Evaluate researcher questions and requests, and match tool to request*

- Learn how a web-based topic modeling algorithm works

  → *Gain experience with off-the-shelf solutions text mining*

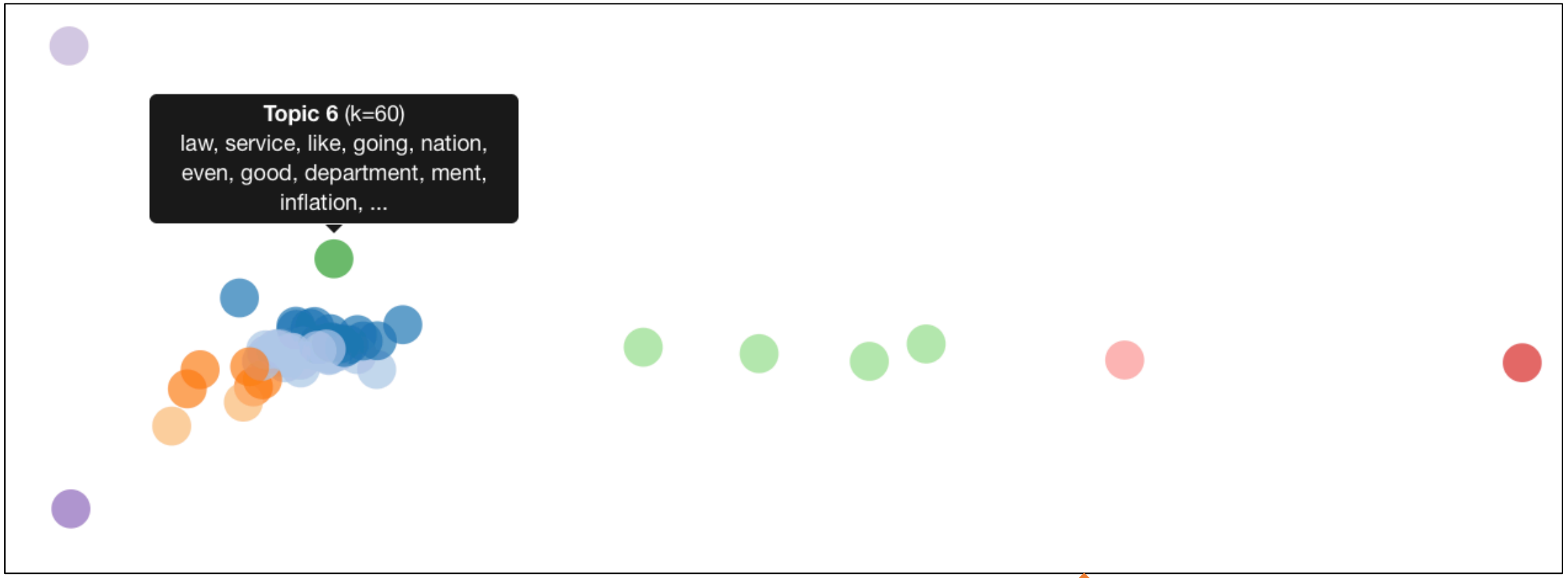
- Run the HTRC Topic Modeling algorithm and analyze the results

  → *Build confidence with the outcomes of data-intensive research*

- See how Sam explored HTRC Algorithms for his research

  → *Understand how a researcher evaluated an off-the-shelf tool*

Where we'll end up



Bubble visualization of topics created with HTRC algorithm

Pre-built tools

- Benefits
  - Easy to use, good for teaching
- Drawbacks
  - Less control, limited capabilities
- Examples:
  - Voyant, Lexos
  - HTRC algorithms: e.g. Topic Modeling algorithm

Do-it-yourself tools

- Alternative to pre-built, off-the-shelf tools
- Involve programming
- Benefits:
  - Run on your own, allow for more parameterization and control
- Drawback:
  - Require technical knowledge

- *We'll return to this later…*

Choosing a pre-built tool

Depends on goal of researcher:

- Quick analysis and visualizations:
  - Voyant
  - Lexos
- Concordances:
  - AntConc
  - Voyant
- Machine learning
  - WEKA Workbench aids machine learning

HTRC algorithms

- Plug-and-play text analysis
- Built into the HTRC interface
  - Mostly “as-is”
  - Limited parameterization
  - Analyze HTRC worksets
- Good when you want to use HT text specifically

Choosing an HTRC algorithm

- Task-oriented algorithms:
  - Produce list of named entities
  - Visualize most frequently used words
  - Generate script for downloading Extracted Features files
- Analytic algorithms:
  - Generate topic models

Key terms in text analysis

Bag-of-words

Concept where grammar and word order of the original text are disregarded and frequency is maintained.

created the four in new are ago Liberty fathers that forth continent a nation seven and conceived equal score dedicated on to years this all our men brought and proposition

Key terms in context

Topic Modeling

- **Chunk** text into documents
- Documents = **bags of words**
- **Stop words** are removed
- Each word in each document is compared
- Words that tend to occur together in documents are likely to be about the same thing
- Topics are predictions of words co-occurrence

Tips for topic modeling

- Treat topic modeling as step in analysis
- Input affects output
  - Number of texts analyzed, number of topics generated
  - Be familiar with your input data
  - Know that stop words can shape results
- Examine results to see if they make sense
- Understand the tool

HTRC topic modeling description

InPhO Topic Model Explorer

Volumes in a workset may be inaccessible to HTRC Analytics algorithms. You can validate your workset before submitting your job using the Workset Validation page to check which volumes are accessible.

Description

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

How it works:

- Downloads each HathiTrust volume from the Data API.
- Tokenizes each volume using the topicexplorer init command.
- Apply stoplists based on the frequency of terms in the corpus, removing the most frequent words accounting for 50% of the collection and the least frequent words accounting for 10% of the collection.
- Create a new topic model for each number of topics specified. For example, "20 40 60 80" would train separate models with 20 topics, 40 topics, 60 topics and 80 topics.
- Display a visualization of how topics across models cluster together. This enables a user to see the granularity of the different models and how terms may be grouped together into "larger" topics.

More documentation of the Topic Explorer is available at https://inpho.github.io/topic-explorer/.

https://analytics.hathitrust.org/algorithms/InPhO_Topic_Model_Explorer

Hands-on activity

☞ *See Handout p. 1*

- With a partner, each read the algorithm descriptions.
- Take turns explaining what they do.
- *Bonus:* Do you have experience with a research question well-suited to each algorithm? Describe it to your group.

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

https://wiki.htrc.illinois.edu/x/HoJnAQ

Sample Reference Question

I'm a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I'd like to examine how concepts such as liberty change over time.

Approach: Run topic modeling algorithm to get a feel for the topics present in your workset.

Hands-on activity

☞ *See Handout pp. 2-3*

In this activity you will run the topic modeling algorithm in HTRC Analytics to explore the most prevalent topics in our president public papers workset.

What You Need:

Website: https://analytics.hathitrust.org

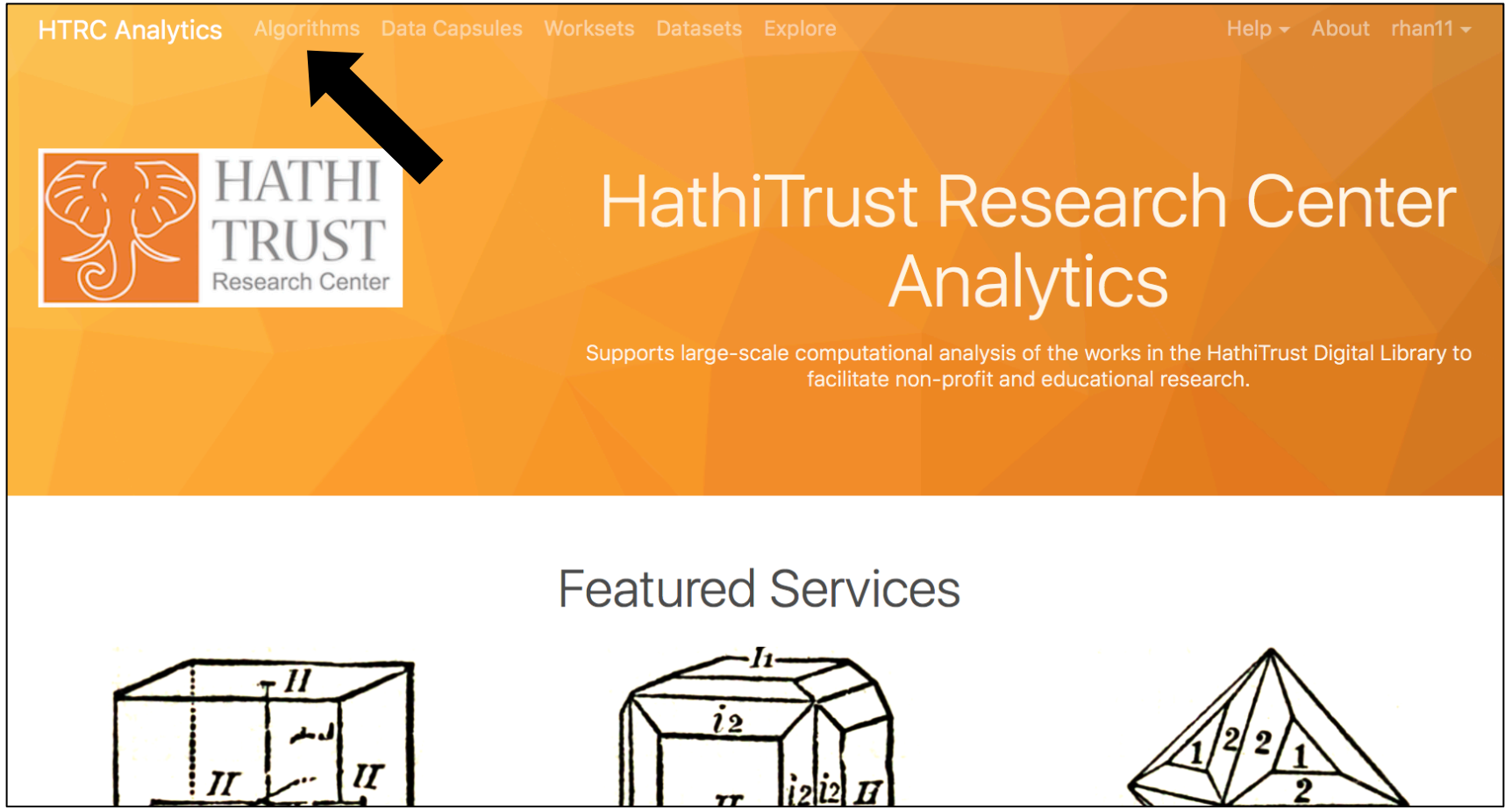
Workset: poli_science_DDRF

About the political science workset

- Government-published series: *Public papers of the presidents of the United States*
  - “Public Messages, Speeches, and Statements of the President”
- 16 volumes from U.S. presidents during the 1970s:
  - Jimmy Carter
  - Gerald Ford
  - Richard Nixon
- We’ll use the same workset (‘poli_science_DDRF@eleanordickson’) so that we can all examine the same results!

Using the HTRC Algorithms

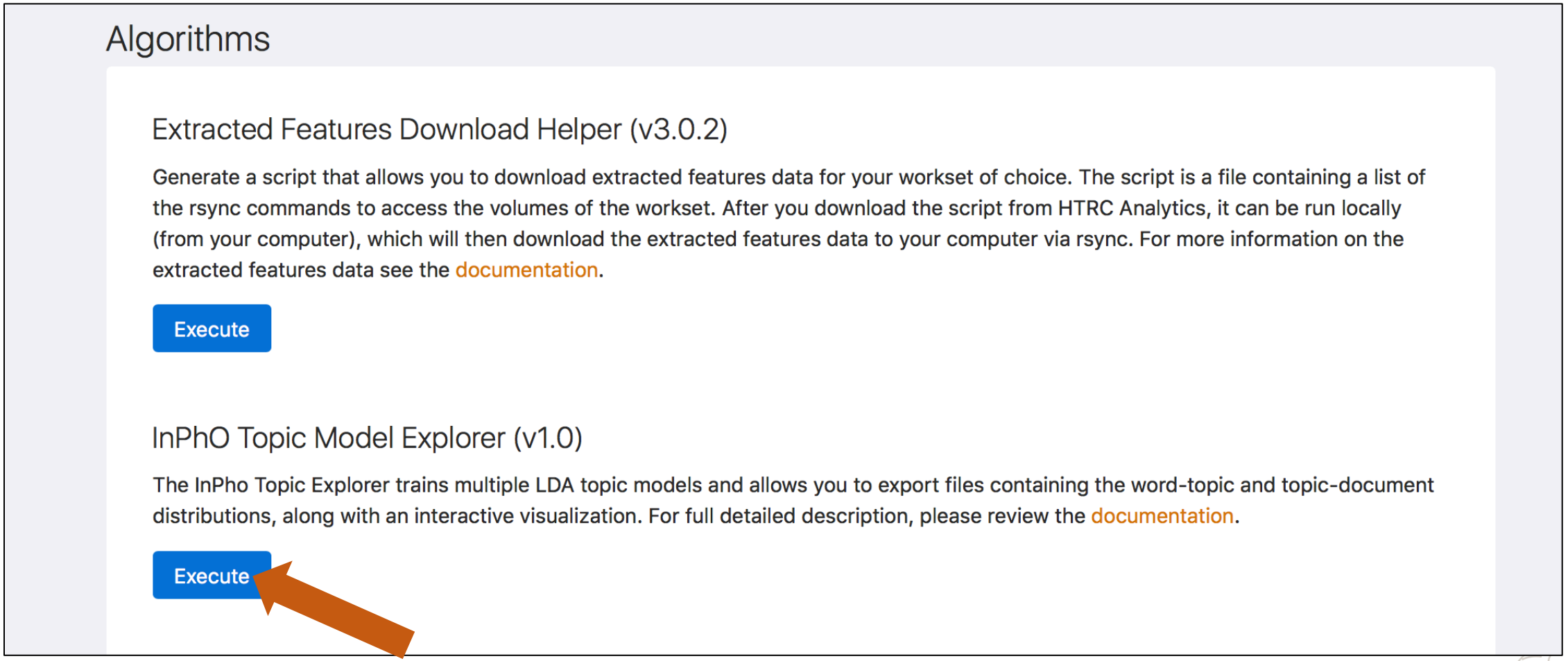
☞ *See Handout p. 2*



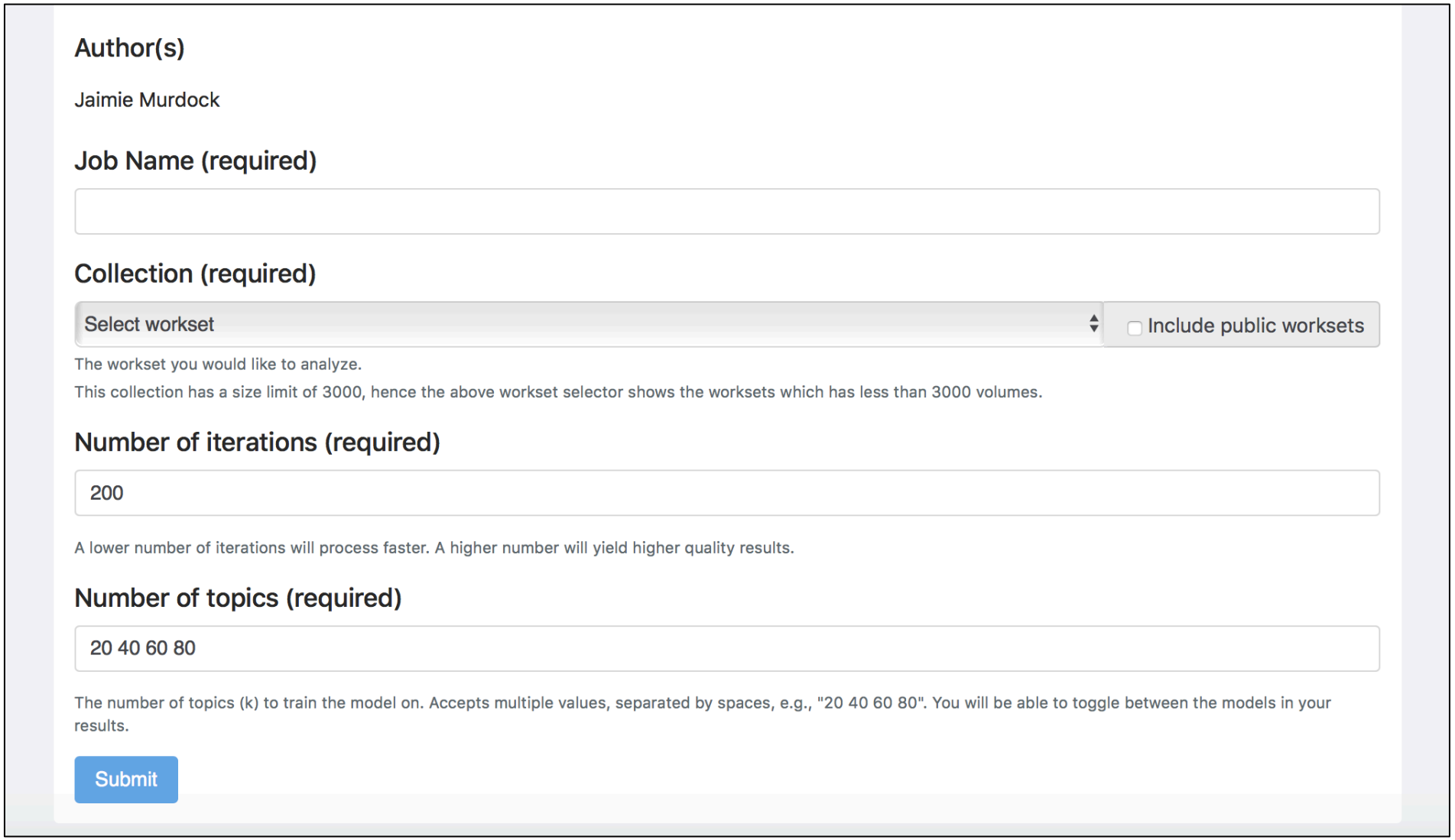
https://analytics.hathitrust.org

Analysis in the HTRC

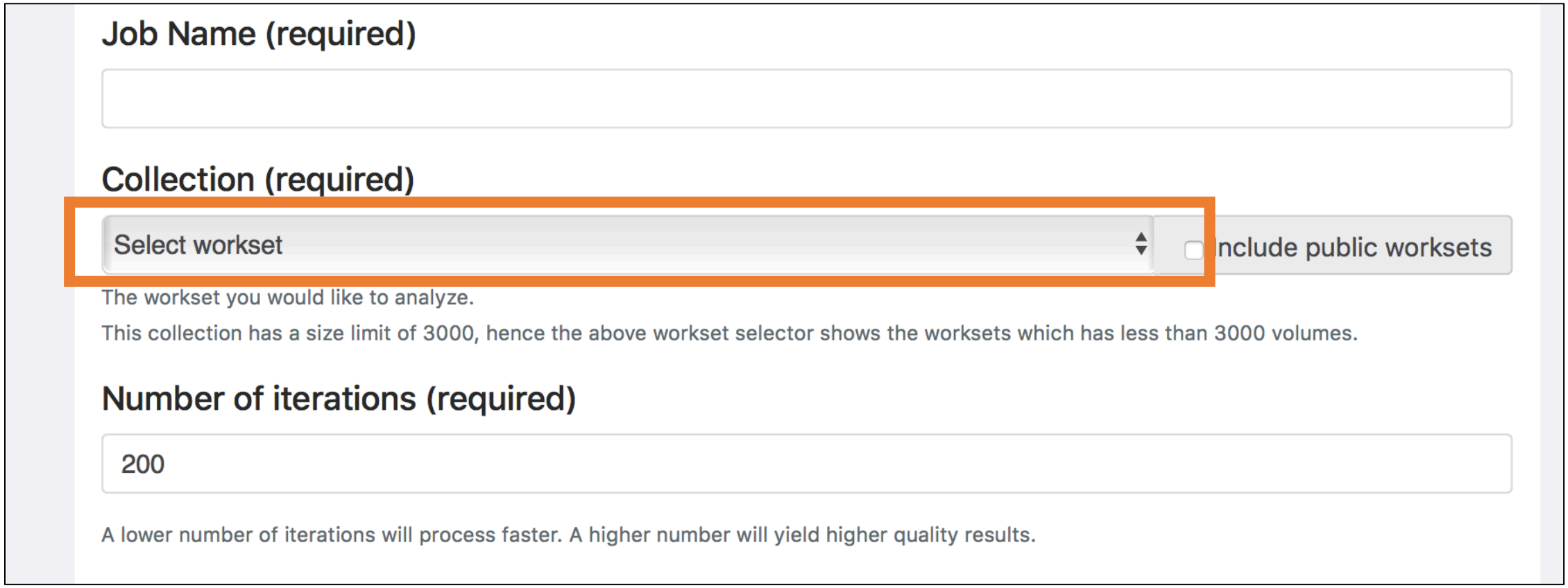
☞ *See Handout p. 2*



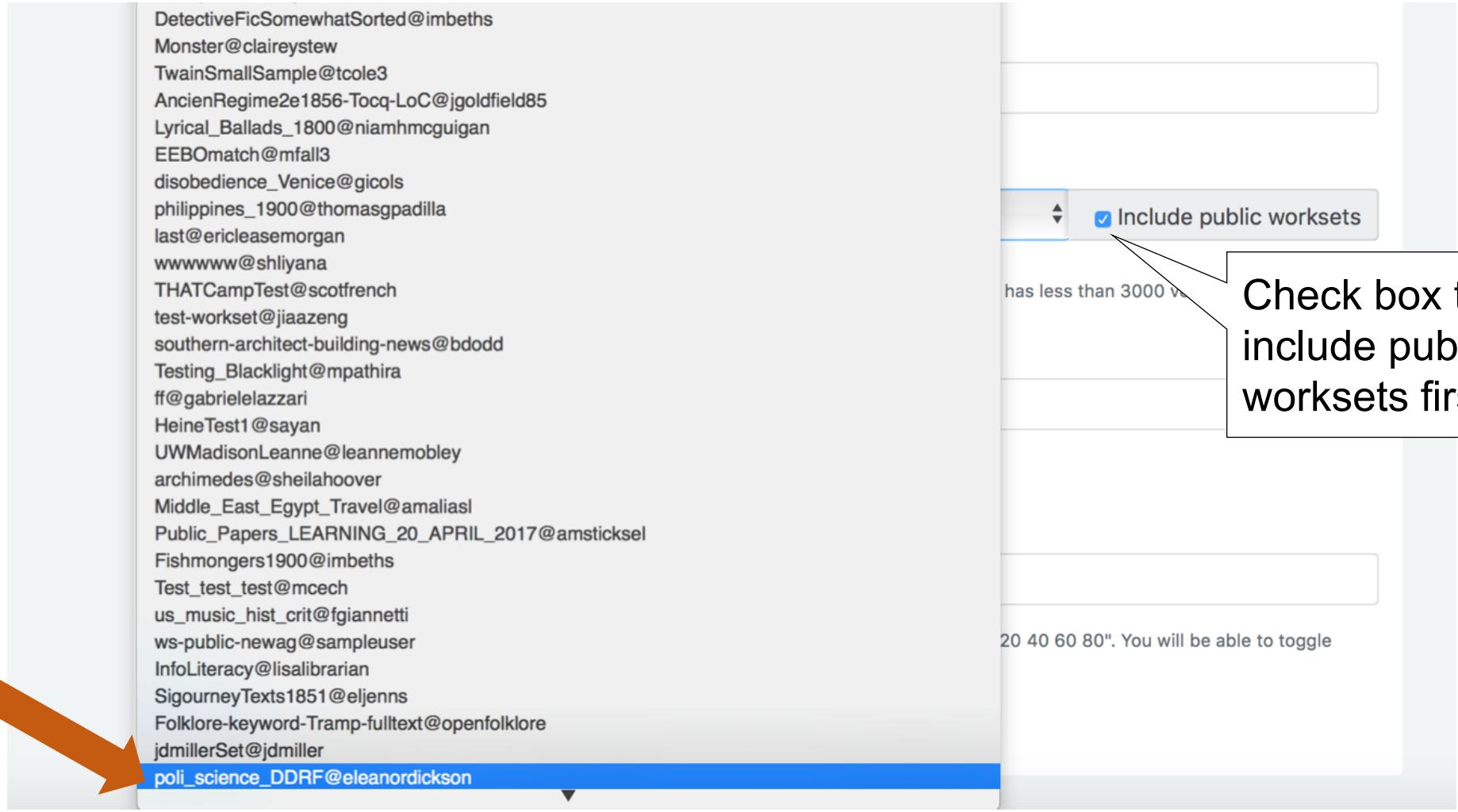
Prepare to run an algorithm



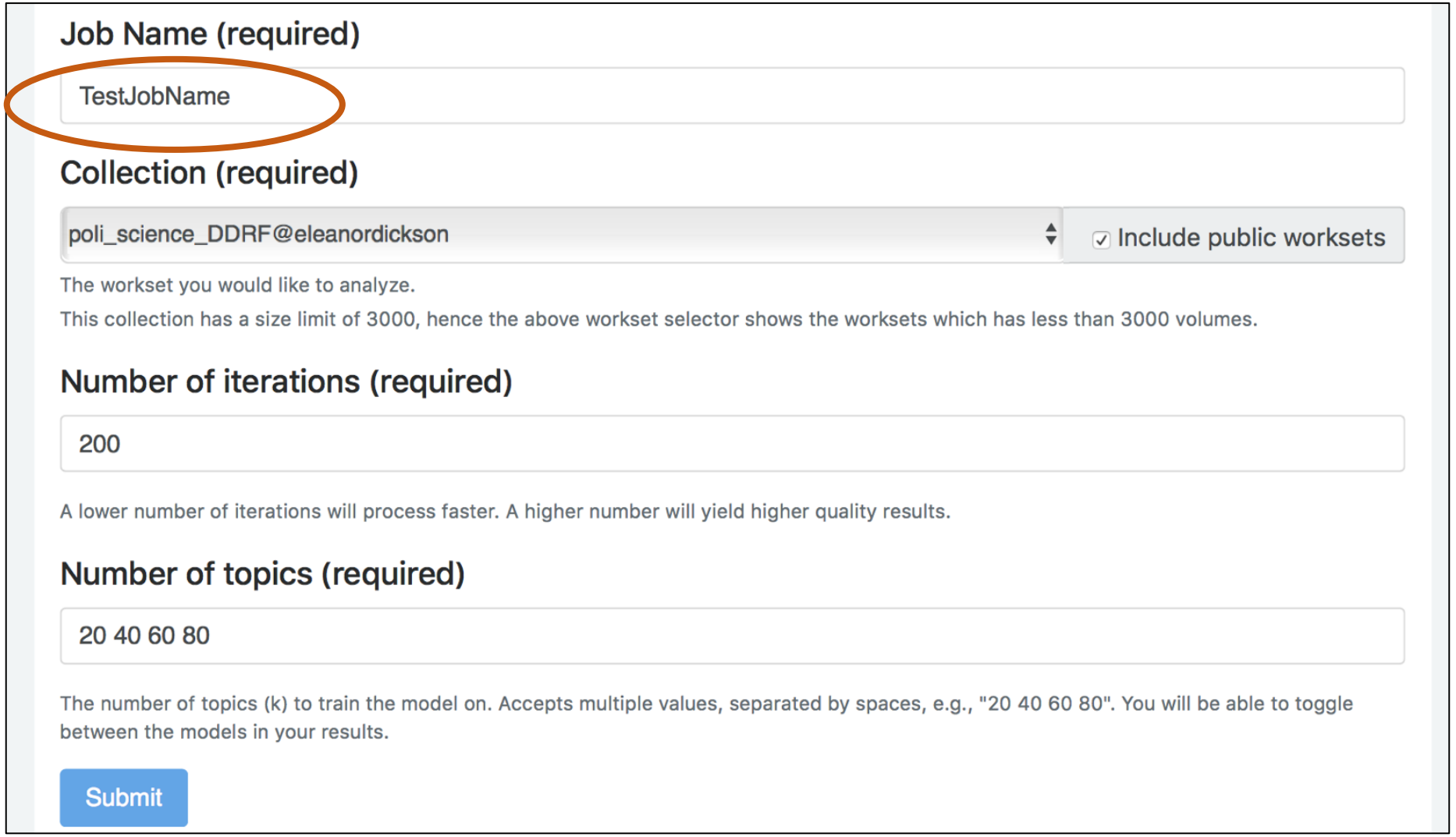
Prepare to run an algorithm



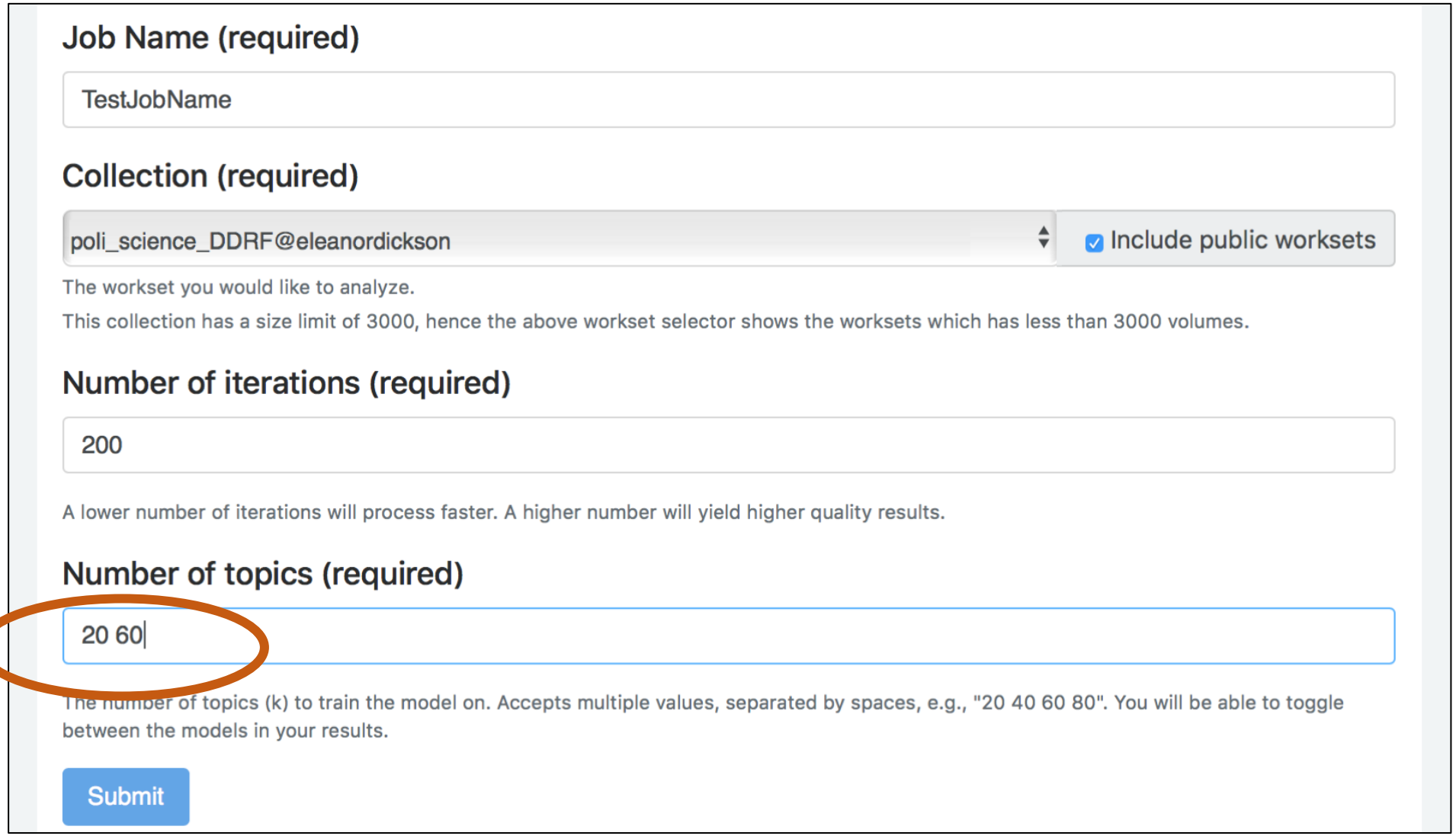
Choose workset(s) for analysis



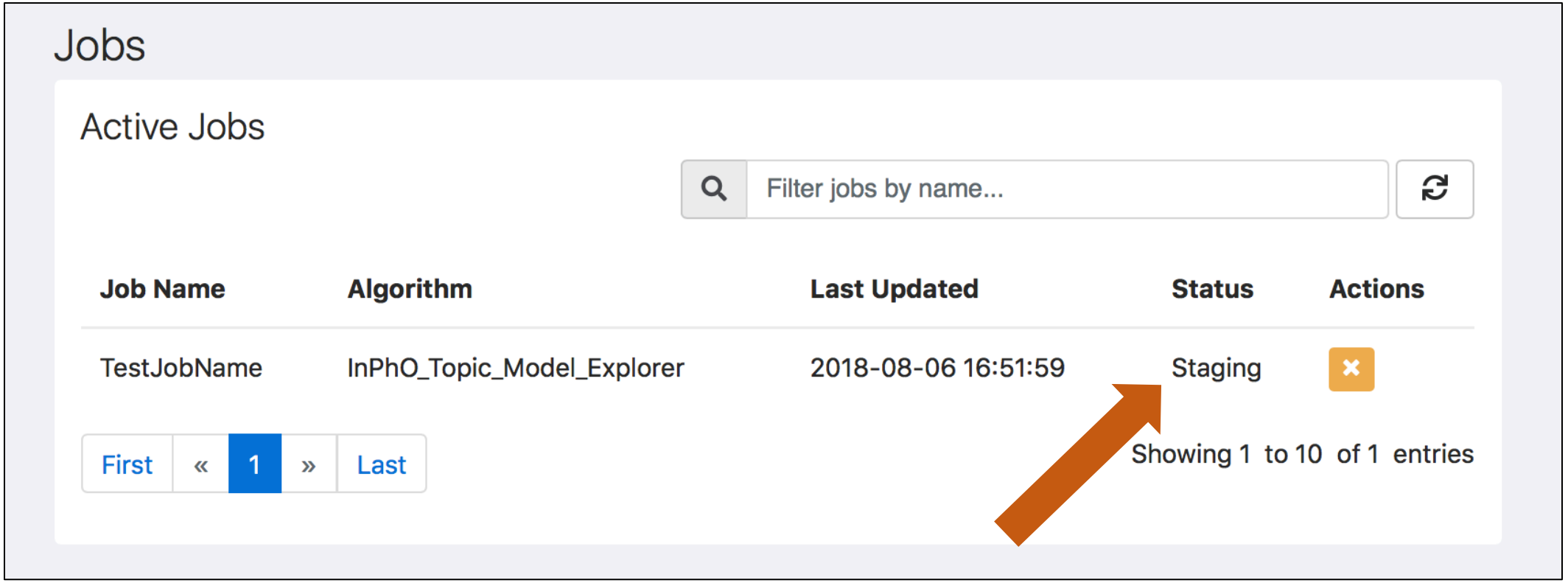
Prepare to run an algorithm



Set the number of topics

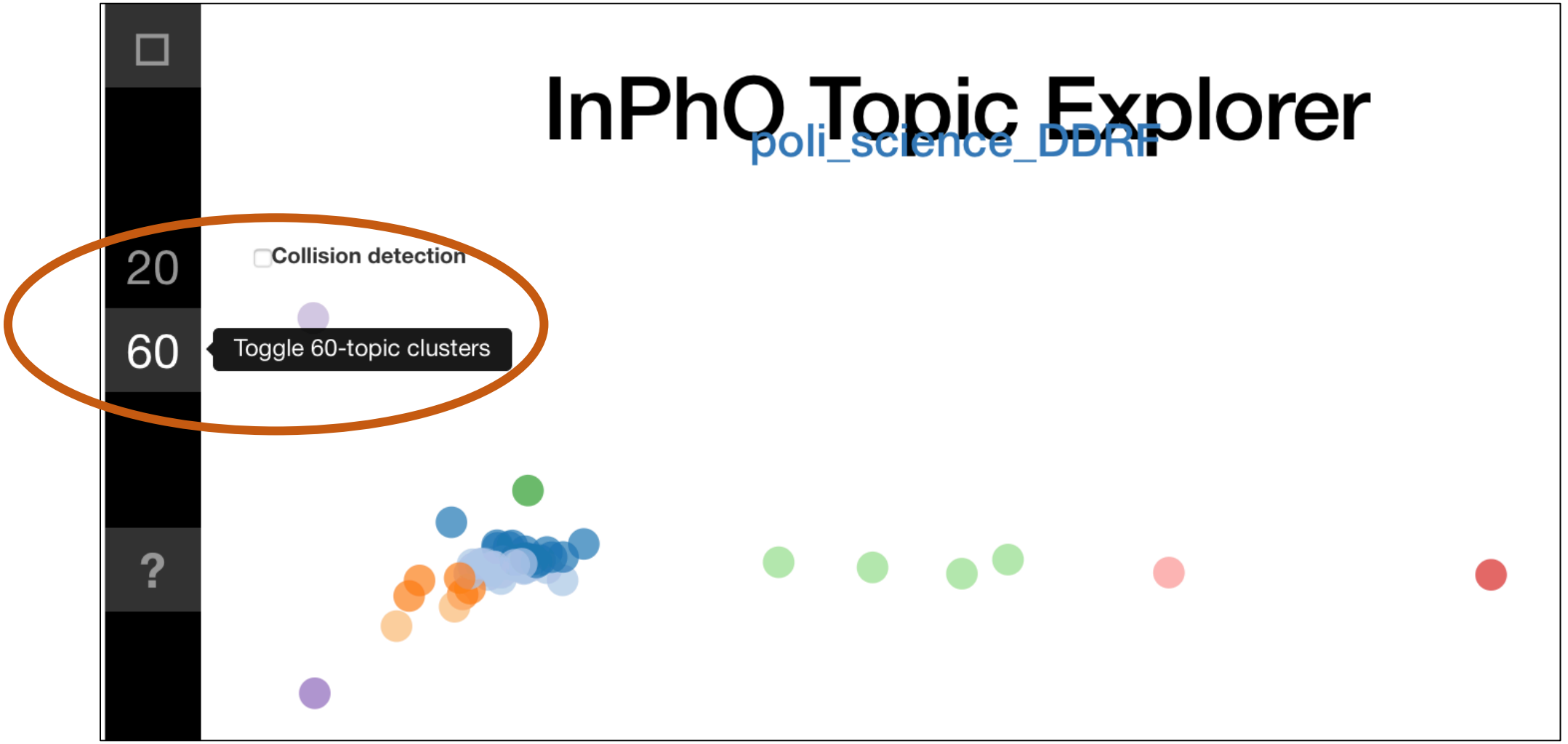


Run the analysis

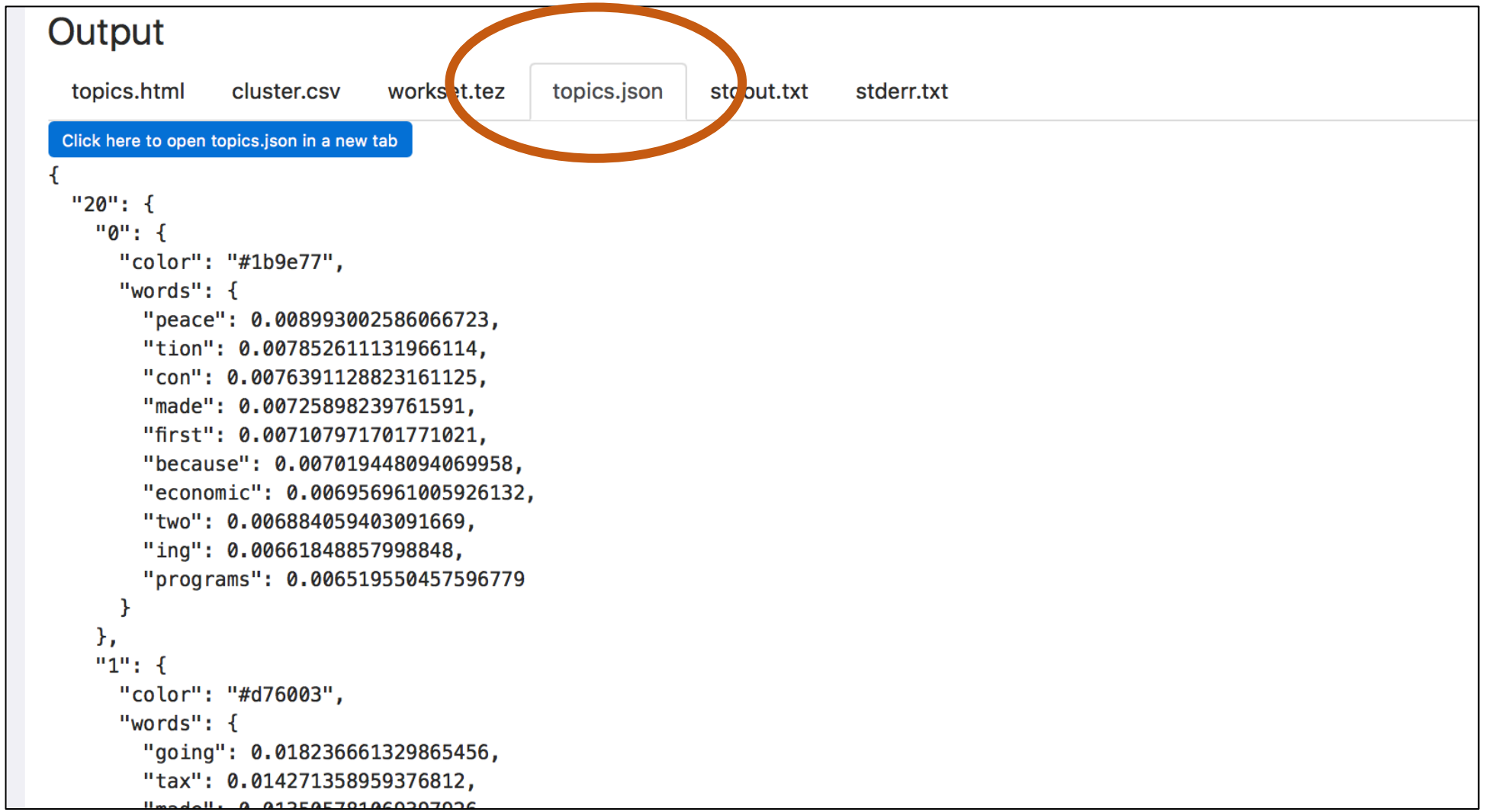


View results

Topics visualized



Results files



Topics listed

Examples from 20-topics cluster:

Topic 1

nation, because, problems, under, america, security, nations, programs, con, much

Topic 2

may, such, peace, war, between, america, last, must, after, soviet

Examples from 60-topics cluster:

Topic 3

like, department, percent, said, things, office, get, assistance, programs, every

Topic 4

oil, programs, presidents, nations, cooperation, york, billion, council, kind, visit

Topic 5

problems, much, system, economy, proposed, must, each, end, case, effective

Analyzing results

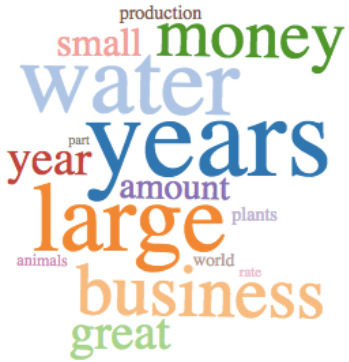
- What would you name these topics?
- Are you skeptical of any of the results?
- Did you learn anything new from the topics produced?

Case Study: *Inside the Creativity Boom*

- Before making his Creativity Corpus, Sam experimented with an older version of the HTRC topic modeling algorithm
- His practice HTRC workset included public domain texts from 1950 to present
  - Creativ* in the title

Case Study: *Inside the Creativity Boom*

Are these good topics?



Tips for topic modeling

- Treat topic modeling as step in analysis
- **Be familiar with input text**
- **Examine results to see if they make sense**
- Know that stop words can shape results
- Understand the tool

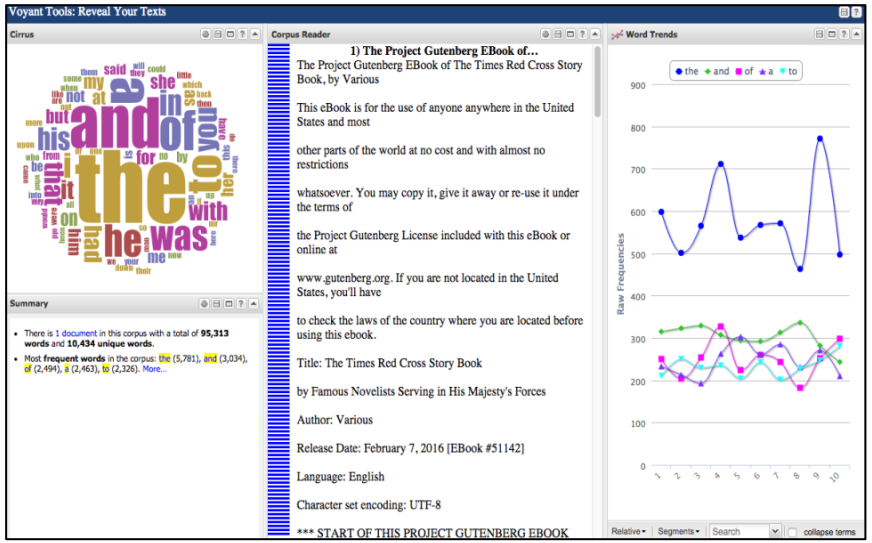
Case Study: *Inside the Creativity Boom*

- Sam ended up using HTRC Extracted Features to get the data needed to analyze contemporary material
- The fits and starts of his project are a great real-world example!

Other pre-built tools

Depends on your goal for analysis:

- **Voyant:** easy to use, good for instruction
- **Lexos**: text cleaning, visualizations
- **AntConc**: good for concordances
- **WEKA Workbench:** aids machine learning
- **HTRC algorithms:** good for using HT text specifically



Voyant dashboard

Discussion

- To what kinds of researchers on your campus would you recommend pre-built text analysis tools?
- Do you have any techniques for introducing these tools that have worked well in the past?
- If you have not taught digital scholarship tools, what techniques appeal most to you at this point?

Questions?

References

- Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM* 55, 4 (April 2012), 77-84. http://dx.doi.org/10.1145/2133806.2133826.