

**Digging Deeper,
Reaching Further**

Module 4.1: Analyzing Textual Data

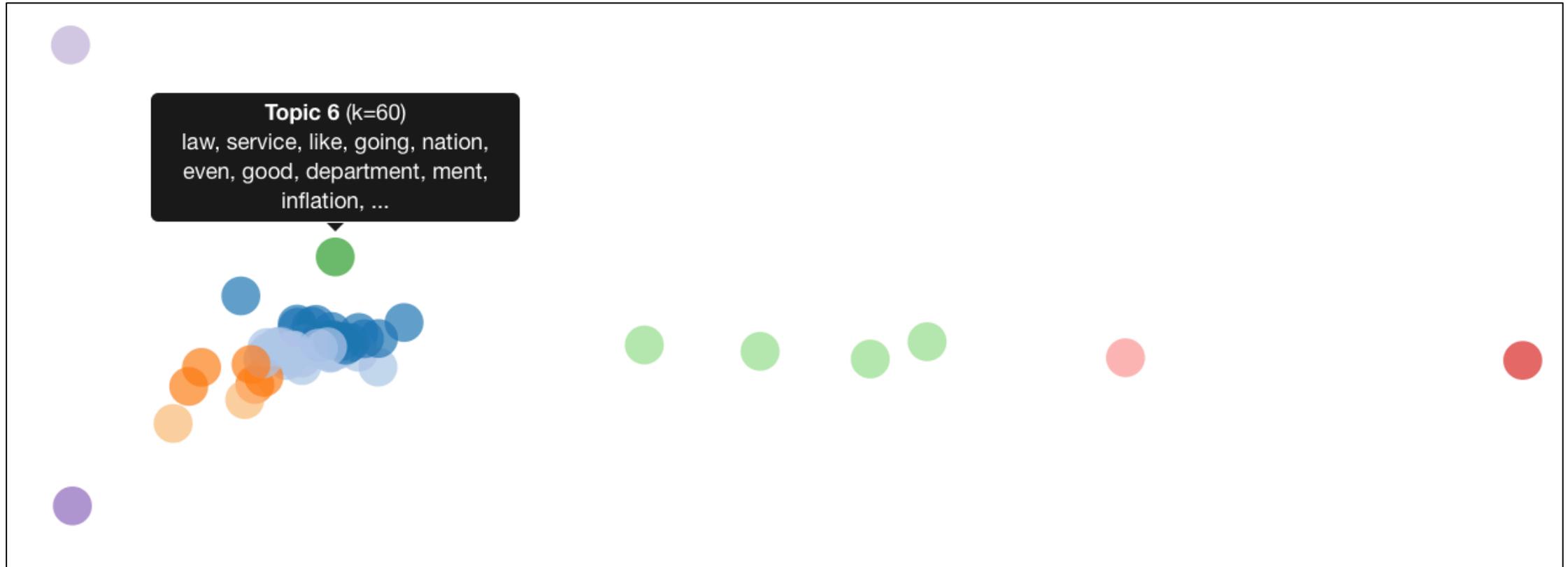
Using Off-the-Shelf Tools

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/>.



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



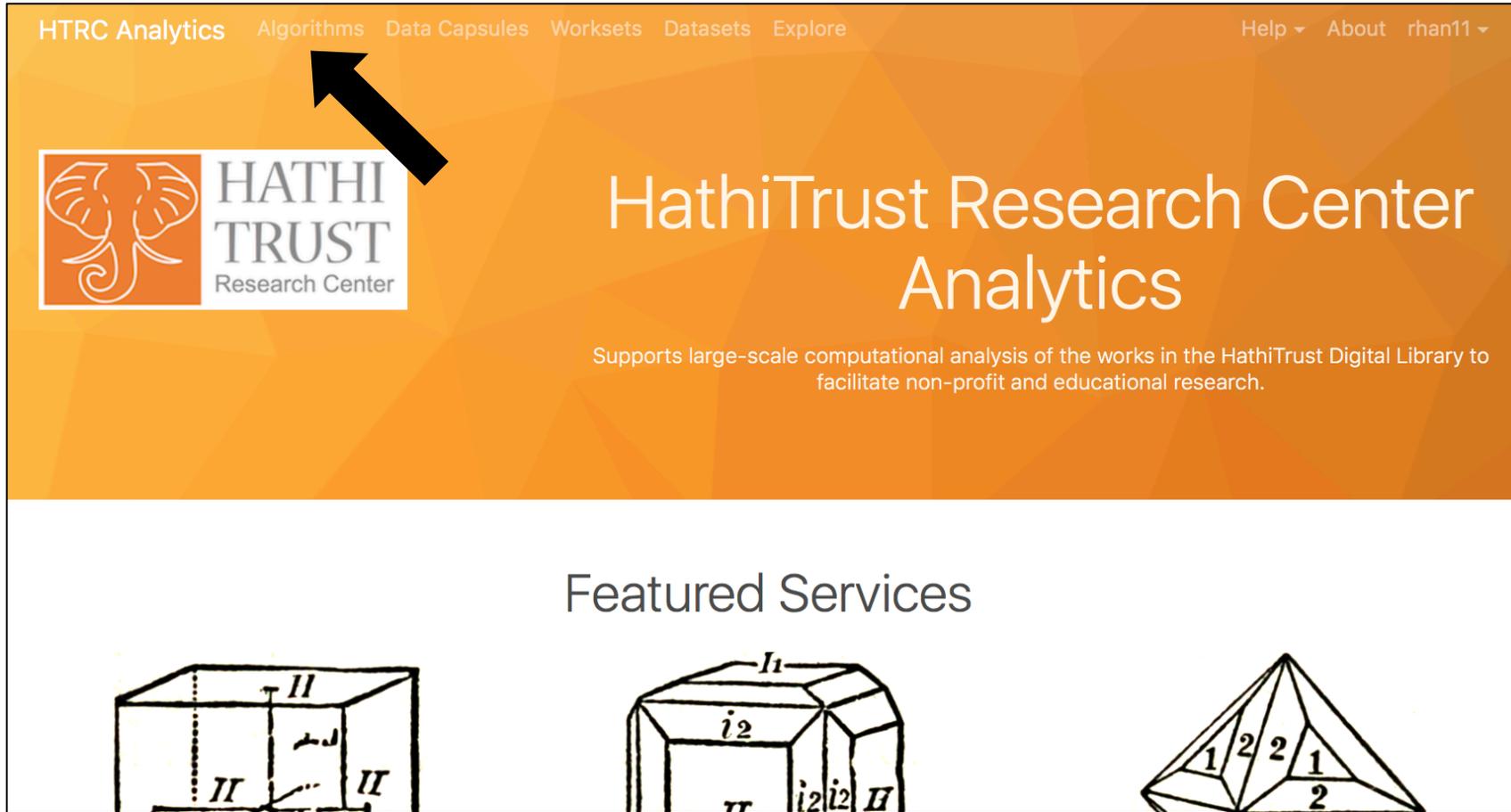
About the political science workset

- 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



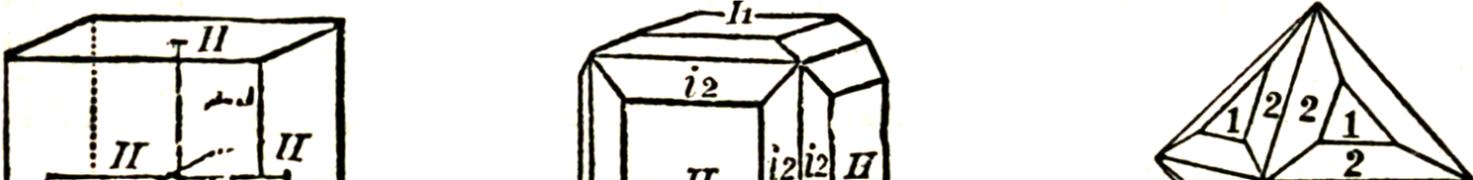
HTRC Analytics Algorithms Data Capsules Worksets Datasets Explore Help About rhan11



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



Analysis in the HTRC

 See Handout p. 2

Algorithms

Extracted Features Download Helper (v3.0.2)

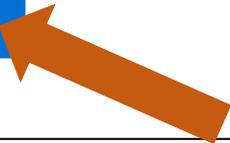
Generate a script that allows you to download extracted features data for your workset of choice. The script is a file containing a list of the rsync commands to access the volumes of the workset. After you download the script from HTRC Analytics, it can be run locally (from your computer), which will then download the extracted features data to your computer via rsync. For more information on the extracted features data see the [documentation](#).

Execute

InPhO Topic Model Explorer (v1.0)

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](#).

Execute



Prepare to run an algorithm

Author(s)

Jaimie Murdock

Job Name (required)

Collection (required)

Select workset 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)

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

Number of topics (required)

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.



Prepare to run an algorithm

Job Name (required)

Collection (required)

Select workset



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.



Choose workset(s) for analysis

DetectiveFicSomewhatSorted@imbeths
Monster@claireystew
TwainSmallSample@tcole3
AncienRegime2e1856-Tocq-LoC@jgoldfield85
Lyrical_Ballads_1800@niamhmcguigan
EEBOmatch@mfall3
disobedience_Venice@gicols
philippines_1900@thomasgpadilla
last@ericleasemorgan
wwwwww@shliyana
THATCampTest@scotfrench
test-workset@jiaazeng
southern-architect-building-news@bdodd
Testing_Blacklight@mpathira
ff@gabrielelazzari
HeineTest1@sayan
UWMadisonLeanne@leannemoble
archimedes@sheilahoover
Middle_East_Egypt_Travel@amaliasl
Public_Papers_LEARNING_20_APRIL_2017@amsticksel
Fishmongers1900@imbeths
Test_test_test@mcech
us_music_hist_crit@fgiannetti
ws-public-newag@sampleuser
InfoLiteracy@lisa librarian
SigourneyTexts1851@eljenns
Folklore-keyword-Tramp-fulltext@openfolklore
jdmillerSet@jdmiller
poli_science_DDRF@eleanordickson

Include public worksets

has less than 3000 v

20 40 60 80". You will be able to toggle

Check box to include public worksets first



Prepare to run an algorithm

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 40 60 80

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.

Submit



Set the number of topics

Job Name (required)

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)

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

Number of topics (required)

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.



Run the analysis

Jobs

Active Jobs

Filter jobs by name...

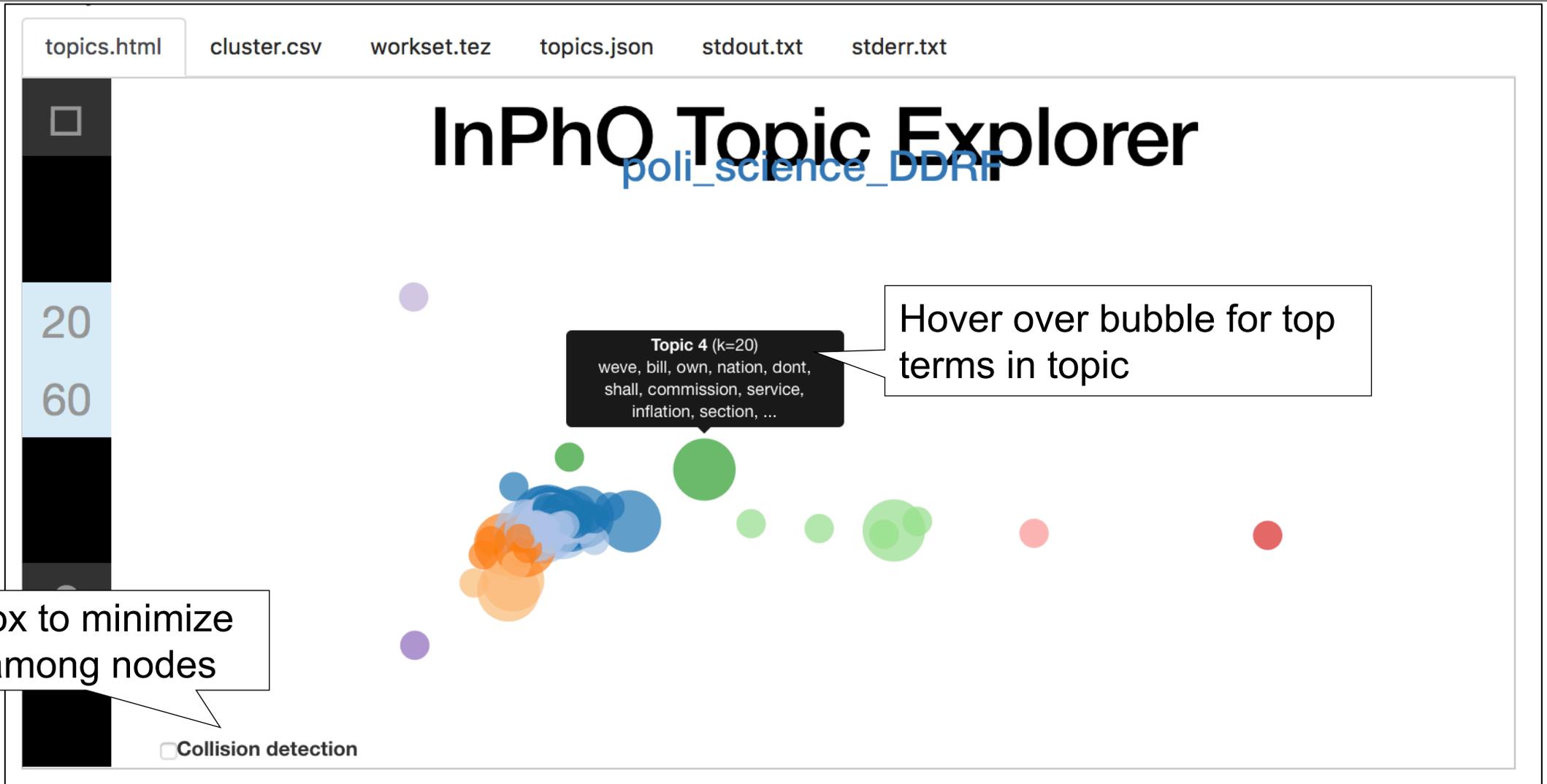
Job Name	Algorithm	Last Updated	Status	Actions
TestJobName	InPhO_Topic_Model_Explorer	2018-08-06 16:51:59	Staging	

Showing 1 to 10 of 1 entries

Navigation: First << 1 >> Last



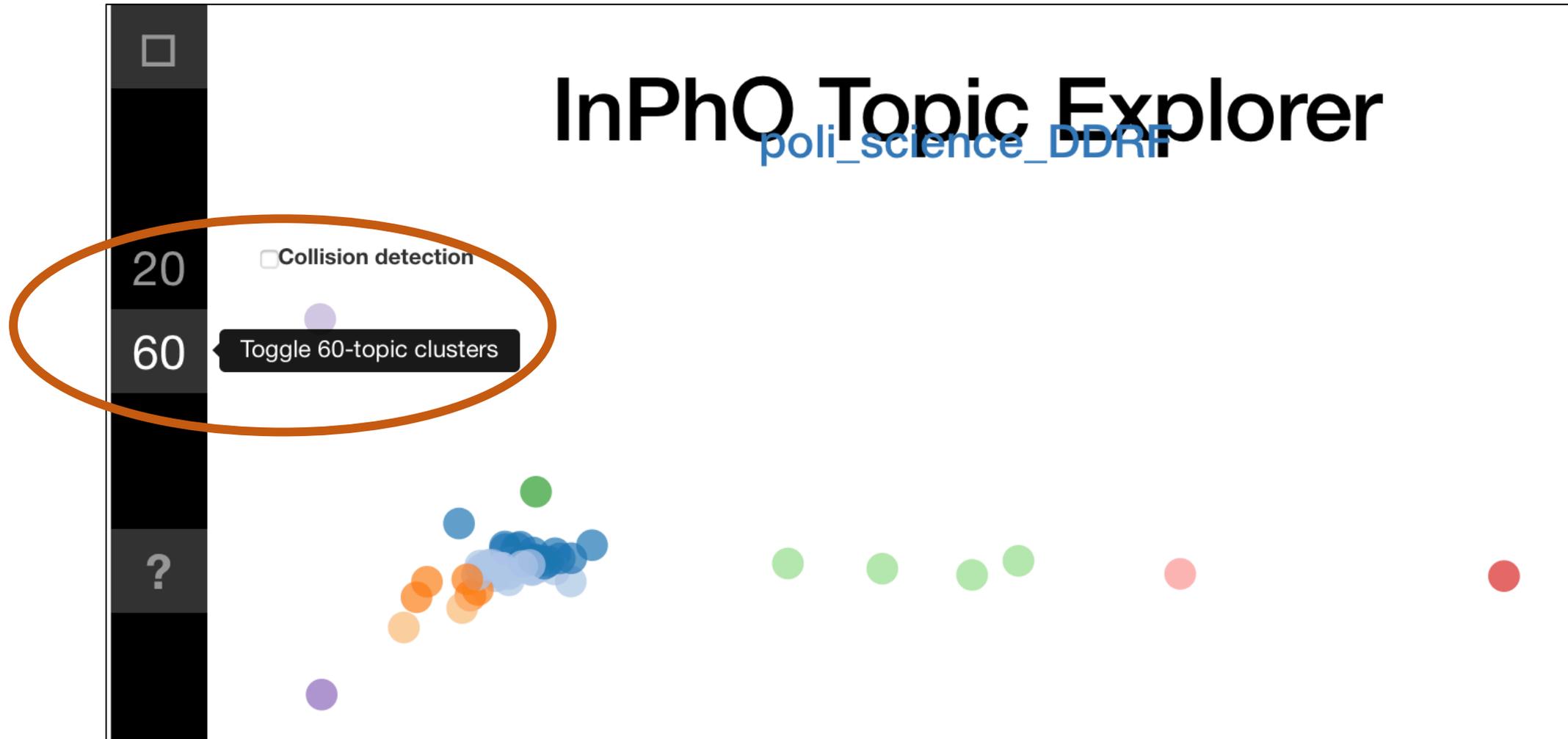
View results



Check box to minimize overlap among nodes



Topics visualized



Results files

Output

topics.html

cluster.csv

workspace.tez

topics.json

stdout.txt

stderr.txt

[Click here to open topics.json in a new tab](#)

```
{
  "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,
      "model": 0.013505781060207026
```



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?



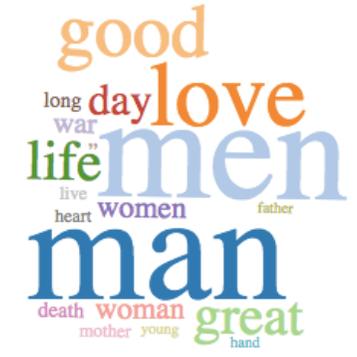
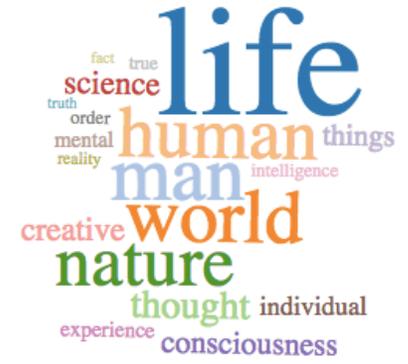
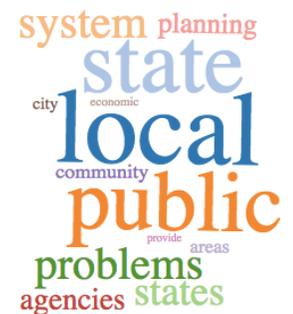
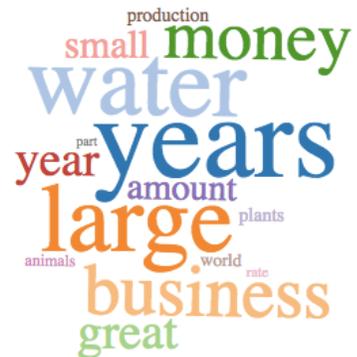
Case Study: *Inside the Creativity Boom*

- 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?



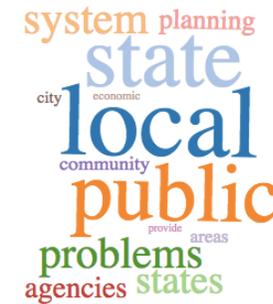
Tips for topic modeling

- 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!



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>.

