# Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

## Module 4.1 Performing Text Analysis: Using Off-the-shelf Tools Instructor Guide

Further reading: go.illinois.edu/ddrf-resources

---

### Narrative

The following is a suggested narrative to accompany the slides. Please feel free to modify content as needed when delivering workshops. Additional information and examples that are provided to aid the instructor's understanding of the materials are in *Italic text* – whether to present them to workshop audiences is entirely up to the instructor.

**Slide** 4.1-1

In this lesson, we will be focusing on supporting beginner researchers in performing text analysis with pre-built tools. It will discuss the advantages and constraints of web-based text analysis tools and programming solutions, introduce basic text analysis algorithms available in the HTRC algorithms, and demonstrate how to select, run, and view the results of the topic modeling algorithm.

**Slide** 4.1-2

Here is an outline of this module. We will weigh the benefits and drawbacks of pre-built tools for text analysis, learn how a topic modeling algorithm works, and run the HTRC Topic Modeling algorithm and analyze the results. We will also consider how Sam experimented with HTRC Algorithms to explore his corpus.

**Slide** 4.1-3

By the end of this module, you will have run the HTRC Topic Modeling algorithm on a sample workset of politically-themed texts and analyzed the results.

**Slide** 4.1-4

For performing text analysis, there are both pre-built tools and do-it-yourself tools that you can choose from.

- The benefits of pre-built tools are they are usually easy to use, they don't require too much technical knowledge, and they can be very useful in teaching.
- The main drawback of pre-built tools is because they have predefined functions, they offer less control to the researcher and limit what a user can accomplish.
- Voyant, Lexos, and HTRC algorithms are examples of pre-built tools. The HTRC Topic Modeling algorithm, which identifies "topics" in a set of texts based on words that have a high probability of occurring close together, is a representative HTRC algorithm that can be used for exploratory analysis. As we mentioned before, we will try using this algorithm together later in this module.

**Slide** 4.1-5

You can also use do-it-yourself ("DIY") tools for text analysis as an alternative, and these tools usually involve some degree of programming.

- The advantages of using do-it-yourself tools are you can set your own workflows and parameters, and it allows you to have more control over what you want to do.
- The drawback is they usually require technical knowledge and may be a lot harder to use.
- We'll come back to do-it-yourself options later in the workshop!

**Slide** 4.1-6

Choosing a pre-built tool is based on the goal of the analysis. Each pre-built tool has its own strengths, outputs, and weaknesses.

- For quick analysis and visualizations, Voyant and Lexos are all excellent choices. They don't allow much parameterization, but they are very easy to use.
- To make concordances (seeing key words in context), AntConc or Voyant are tools programmed to do that. Voyant is both web-based and downloadable, so there no installation required unless desired by the researcher, while AntConc is software the user must install on their computer.
- Weka is the tool to use for researchers who want to try machine learning with a tool that has a visual front-end. It also needs to be installed.

**Slide** 4.1-7

For this module, we are going to use the HTRC algorithms as an example of how to use a pre-built tool for text analysis. The HTRC provides some algorithms through the HTRC Analytics, and they can be used to analyze HTRC worksets.

- An **algorithm** is just a way of saying a computer function - text goes in, process happens, and results come out.
- HTRC algorithms can extract, refine, analyze, and visualize worksets. They can basically perform “plug-and-play” text analysis. Because they are built into HTRC Analytics, they are mostly limited in how much they can be tweaked or customized. The algorithms are primarily for users who don’t know how or don’t want to work with custom code. It can be a good tool for learning and just trying things out.

**Slide** 4.1-8

Currently there are four HTRC algorithms available. How do you choose which HTRC algorithm to use? Naturally, it depends on what you want to do.

Most of the algorithms are task oriented. For example, there are ones for generating a list of named entities and for visualizing more frequently used words.

The algorithm that generates topic models is more analytic.

**Slide** 4.1-9

Next, we will focus on a specific text analysis method: topic modeling. One concept closely related to topic modeling is the bag-of-words model.

- “Bag-of-words” is a concept where grammar and word order of the original text are disregarded and frequency is maintained.
- Here is an example of the beginning of *The Gettysburg Address* as a bag of words.

**Slide** 4.1-10

Topic modeling is a method of using statistical models for discovering the abstract "topics" that occur in a collection of documents.
This summarizes what happens in a topic model:

- For this kind of analysis, the text is chunked into “documents”, and stop words (frequently used words such as “the”, “and”, “if”) are removed since they reveal little about the substance of a text.

- The computer treats the documents as bags of words, and guesses which words make up a “topic” based on their proximity to one another in the documents, with the idea the words that frequently co-occur are likely about the same thing.
- The results are groupings of words that the computer has statistically analyzed and determined are likely related to each other about a “topic”.

**Slide** 4.1-11

Here are some tips for topic modeling:

- Treat topic modeling as one part or step within a larger analysis.
- Input affects output: Understand that what you input, including how you set your parameters, will affect the output. Some points to note are:
    - Be careful with how you set the number of texts analyzed, as well as number of topics generated
    - Be familiar with your input data
    - Know that changing your stop word list can have really interesting impacts on your resulting topics, so tread carefully/wisely.
- Make sure to examine your results to see if they make sense: You’ll need to return to your input text at some point, and see if what you’re getting makes sense.
- Also, try to gain some basic understanding of your tool. Reading some relevant documentation is especially important when the tool is within a “black box”.
    - Speaking of understanding your tools, it’s important to note that the HTRC algorithm only has two parameters you can set right now, so it’s not suitable for robust analysis with topic modeling. But for teaching and exploration of HT text specifically, the HTRC topic modeling algorithm can be a good place to start.

**Slide** 4.1-12

Together, let’s read the description of the HTRC topic modeling algorithm to see if we can understand what it does. The description is on the screen, or you can find it on the Algorithms page in HTRC Analytics (you will need to log in first to be able to see the page).

- 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 - What is the stoplist removal protocol for this algorithm? *(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.*
    - What else can you download as the Results of your Job?

If you want to read more documentation of the Topic Explorer, you can follow the provided link to the documentation ( https://inpho.github.io/topic-explorer/).

*What is the stoplist removal protocol for this algorithm?*

*What other items are generated with the visualization as the Results of the Job?*

**Slide** 4.1-13

Let's do a short activity together before moving on to the next section. With a partner, please read the descriptions for the HTRC algorithms in the chart included in the handout.

- Explain to one another what each of these algorithms in the chart does. Then, can you think of the kind of research questions might they help answer? Please discuss with your partner(s) and we'll share our ideas as a group later.
- There are more algorithms with descriptions on the HTRC Algorithms Wiki page (https://wiki.htrc.illinois.edu/x/HoJnAQ ). Take a look at the descriptions of the other algorithms as well if you have time.

*(Instructor organizes the activity. Possible answers:*

*Token Count and Tag Cloud Creator: Questions that involve frequency of words in a thematic set of texts, e.g., What are the most frequent words used in the works of Charles Dickens?;*

*Named Entity Recognizer: Questions that involve identifying people, places, etc. in a workset, e.g., "What are the geographical places named the works of 18$^{th}$ century literature?")*

*Topic Model Explorer: Questions that involve identifying 'topics' or themes in a workset, e.g., topics in presidential speeches)*

*If time runs short, instructor can ask participants to take a few minutes to read the descriptions on their own and skip the part where they explain concepts to each other. Go directly to discussing what the algorithms can do and the research questions suitable for each algorithm.*

**Slide** 4.1-14

Returning to our sample reference question, let's explore how you might use the HTRC algorithms to do this research. One way to approach it is to run the topic modeling algorithm to see what topics are present in a body of political speech texts. We already have a sample workset of the public papers of the presidents to play with.

**Slide** 4.1-15

In this hands-on activity, we will run the topic modeling algorithm in HTRC Analytics to explore the most prevalent topics in our presidential public papers workset. We will be accessing and running the algorithm via HTRC Analytics (https://analytics.hathitrust.org), and will be using a sample workset already prepared and available in Analytics.

**(NOTE: see intro to workset on next slide)**

*(Instructor may choose to demo live/go through the screenshots to introduce the entire process, and then let participants do the activity on their own.)*

**Slide** 4.1-16

Before we actually run the algorithm, here's a quick introduction to the sample political science workset we will be using in this activity.

- This workset consists of a set of volumes from a government-published series, *Public Papers of the Presidents of the United States*.
- It contains the public messages, speeches, and statements of the U.S. presidents. The workset we will use includes 16 volumes of speeches by Jimmy Carter, Gerald Ford, and Richard Nixon.
- We are working with the 1970s, because we were able to find volumes representing every year in that decade in the HathiTrust, so it represents a relatively complete, discrete set.
- The workset is called 'poli_science_DDRF@eleanordickson'.
- In order for us to examine the same results, we will all use the same workset for this activity.

**Slide** 4.1-17

Now, why don't you run the topic modeling algorithm on this workset. The instructions are included in the handout, and you can work alone or with a partner. Raise your hand if you get stuck. I'll be demoing up here if you'd like to follow along.

To start, we will log into HTRC Analytics and then click on "Algorithms" in the top menu.

**Slide** 4.1-18

*(Optional narrative below)*

For this activity, let's find the InPhO Topic Model Explorer (v1.0) in the list of algorithms and then click "Execute" button.

**Slide** 4.1-19

*(Optional narrative below)*

You will be brought to a page with a more detailed description of the functions of the Topic Model Explorer algorithm. When you scroll down, you will see entry boxes where you can set the parameters according to your needs.

**Slide** 4.1-20

*(Optional narrative below)*

First, you will have to choose a workset to run the algorithm on. Click on the drop-down menu to select from your own worksets, or you can select the "Include public worksets" checkbox option on the right to select from public worksets as well as your own worksets.

**Slide** 4.1-21

*(Optional narrative below)*

For this activity, check the "Include public worksets" option and select "poli_science_DDRF@eleanordickson". 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 to get to the right part of the list, and the workset that we need will appear at the bottom of the list.

**Slide** 4.1-22

*(Optional narrative below)*

After selecting this workset for analysis, enter a Job Name of your choice, and this is the name that will show up later as your "Job Name" when looking at results.

**Slide** 4.1-23

*(Optional narrative below)*

Next, we can set the number of training iterations and the number of topics to be created. Let's keep the default setting of 200 iterations, but edit the number of topics to 20 and 60 to give us fewer results to review.

Running few training iterations will process more quickly, while running more iterations will take longer, but produce more precise results. 200 tends to be good for experimenting and 1000 is best when producing something for publication.

Hit the submit button to run the algorithm.

*NOTE: the topic modeling algorithm will return slightly different results each time it is run, since it is probabilistic, unless the* ***exact*** *same parameters are used. If an attendee gets different results than you, ensure them that their outcome is normal and use it as an opportunity to discuss what it means to do probabilistic, computational work.*

**Slide** 4.1-24

*(Optional narrative below)*

Once you click the "Submit" button to start your job, you'll be taken to a screen with your job history. You'll see active jobs at the top. Notice that the status may change when you refresh the screen, and you'll see your completed jobs below.

It can take a while to get results based on the size of the workset, complexity of the algorithm, and the load on the machine at the time.

*(If some users cannot get their jobs processed quickly, the instructor may show the results on the screen.)*

When done, the job name will move down to the Completed Jobs section.

**Slide** 4.1-25

Once the job is completed, click on the job name under the Completed Jobs section. You will be taken to a page with the results of this job.

Scroll to the "output" area, and you will see the bubble visualization of the generated topics, showing how the topics cluster.

If you hover over a bubble, you'll see the top terms in that topic.

The clusters and colors are determined automatically by an algorithm, and provide only a rough guide to groups of topics that have similar themes.

Checking the collision detection checkbox will minimize overlap among the nodes but distort the underlying similarity relationships.

**Slide** 4.1-26

These numbers on the side relate to the number of topics generated, as do the size of the bubbles. You can toggle the display of the n-topic clusters by clicking on the numbers. You'll see that the bubbles in the 20-topic clusters are larger than the bubbles in the 60-topic clusters.

**Slide** 4.1-27

You can download 3 results files:

- topics.json:
    - This file contains the topics for each model, the top 10 terms in each, and the term probabilities within that topic
- cluster.csv:
    - This file contains the information that drives the visualization.
    - In column *K*, you can see which training iteration the rows refer to
    - The column *topic* shows which topic within that training iteration the row refers to
    - The columns orig_x, orig_y, and cluster tell the visualization how to group and display the bubbles
- workset.tez: can be loaded into an instance of the InPho Topic Model Explorer on your own machine
    - The topic modeling tool that underlies this algorithm can also be installed and run locally.
    - You can feed the outputs of your HTRC algorithm into that tool if you have it installed and play with the visualization in more depth, as well as gain access to additional visualization view
    - More information can be found by clicking the link in the algorithm description ( https://inpho.github.io/topic-explorer/.)

*Instructors may click through the tabs and show how results can be downloaded.*

**Slides** 4.1-28 to 4.1-29

Let's look at these results together. As we have mentioned before, topic modeling basically makes guesses about topics in the form of groupings of words, but it cannot directly give you what these groups of words stand for. What exactly are these topics still needs to be identified by the researcher. What would you name these topics? Are you skeptical of any of the results? Did you learn anything new from the topics produced?

*(Leave it open to the group to talk about what they see and if they are useful. There are no right answers.)*

*Some possible names for the topics:*

- *Topic 1: National security*
- *Topic 2: U.S.-Soviet relations*
- *Topic 3: Administration*
- *Topic 4: International affairs*
- *Topic 5: Economic issues*

**Slide** 4.1-30

Now, let's return to our Creativity Boom case study by Sam. In the early stages of Sam's research project, he experimented with HTRC algorithms, using an earlier version of the topic modeling algorithm.

- He had built a workset of volumes containing "creativ" in the title. They were public domain texts from 1950 to the present
- Then he did topic modeling because he wanted to see what topics were prevalent in volumes related to the subject of creativity.

**Slide** 4.1-31

These are the topics that can be generated by the algorithm using Sam's workset of volumes that contain creativ* in the title. Look at the results. Do you think they are illuminating?

We see that the dataset can affect the output in not-so-great ways.

**Slide** 4.1-32

Remember our tips for topic modeling, in particular, "be familiar with input text" and "examine your results to see if they make sense": Sam's workset included a substantial number of government documents. Do you think they would have skewed his results?

**Slide** 4.1-33

Because Sam wanted to explore 20th century (i.e. predominately in-copyright) volumes, he applied for an HTRC Advanced Collaborative Support (ACS) award that allowed him to have assistance in analyzing otherwise restricted text.

Sam's false start with his public domain corpus and topic modeling exploration is a great real-world example of the non-linear research process. Using the web-based algorithm, he was able to quickly discern there was an issue with his workset. This realization allowed him to move on to other options.

**Slide** 4.1-34

Again, remember that there are a host of pre-built tools that you can use for text analysis. Here are some we mentioned before:

- Voyant is very easy to use: it's flexible in the kinds of files that it can examine, and it is excellent for teaching;
- Lexos has good text cleaning capabilities and can also produce nice visualizations like word clouds;
- AntConc is recommended for building concordances (seeing key words in context);
- The WEKA Workbench is a downloadable tool that assists with machine learning;
- HTRC Analytics algorithms are good for when a researcher wants to analyze text from the HathiTrust corpus specifically, and they are suited for quick analysis or teaching.

**Slide** 4.1-35

Before we end this session, let's have a short discussion about digital scholarship in an educational or research setting.

- *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?*

*(This is should be a small group discussion for 5-10 minutes. If time runs short, try to cut this discussion down to 5 minutes at most instead of doing a 5-10-minute discussion.*

**Slide** 4.1-36

That's all we have for this lesson, and we will be happy to take any questions from you.

**Slide** 4.1-37

*(Show references so attendees know where to find them later.)*