Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 5 Visualizing Textual Data: An Introduction 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 M5-1

This lesson is an introduction to data visualization in general, with a focus on textual data analysis. It also introduces the HathiTrust+Bookworm interface that allows the user to visualize word usage over time.

Slide M5-2

In this module, we will first Introduce some common visualization strategies for text data. Next, we will use a web-based visualization tool called HathiTrust+Bookworm to explore lexical trends. Finally, we will see how Sam used HT+Bookworm in his project for visualizing his data.

Slide M5-3

By the end of this module you will have used HathiTrust+Bookworm to create a visualization of word usage trends across the HathiTrust corpus.

Slide M5-4

First, it is important to understand what data visualization is.

- Data visualization is the process of converting data sources into a visual representation. This representation is usually in graphical form. Broadly speaking, anything that displays

data in some visual form can be called a data visualization, including both traditional graphs and charts, as well as more innovative data art. In this lesson, we will be focusing on more common and well-established forms of data visualization.

- Visualizations present particular ways of interpreting data. It is not a transparent, objective projection of what the data is. By selecting different types of visualization and adjusting parameters, the resulting visualization is a researcher's specific way of interpreting and presenting data.
- Data visualization is an entire field of study, so we're barely scratching the surface in this module!

Slide M5-5

Why do researchers visualize data? In general, visualization can help us gain new insights about textual data. Some of the reasons one might want to visualize textual data include:

- By visualizing textual datasets, researchers can understand the general "gist" or broader themes of texts, and they may discover some patterns that cannot be easily extracted by reading texts word-by-word or text-to-text.
- They can also help with tracking any general changes that occur to a certain collection over time and space. For example, we can use HT+Bookworm, a tool we will be introducing later in this lesson, to track lexical trends.
- Visualization tools can also cluster/group texts for the researcher for overview or classification purposes according to different parameters.
- In some cases, a researcher may also want to compare multiple collections of texts, or to correlate patterns in text to those in other data. Visualization can aid in revealing connections and differences between datasets.

Adapted from Jason Chuang's Text Visualization course at Stanford University: http://hci.stanford.edu/courses/cs448b/f11/lectures/CS448B-20111117-Text.pdf

Slide M5-6

Data visualization can be used in two stages during the research process. It can happen in both the earlier exploration stage as well as the later explanation stage.

- In the earlier exploration stage, visualization can be used as a discovery tool. By visualizing data, researchers can explore the full range of the data and extract features and themes/trends in the data. For example, a researcher can use word clouds to visualize the results of topic modelling in order to identify topics more easily, or they can visualize social networks to discover relationships.
- We saw an example of exploratory data visualization in the last module.

- In the later explanation stage, visualizations are commonly used for communicative purposes, often for clearer explanation and presentation of results in publications or other venues. *In many cases, visualizations can also better capture the attention and interest of audiences, and serve as prompts for further questions.*

Slide M5-7

There are many types of data visualizations that can be categorized in multiple ways. In this section, we will present some common types of textual data visualizations.

Most of you have probably seen a **word cloud** before –it's usually a graphical representation of word frequency. We saw word clouds earlier in our topic models from the HTRC algorithms, where size related to prominence or salience within the topic. It's a relatively unsophisticated type of visualization, but often quite effective.

In conventional word clouds, only the font size is controlled to represent one dimension of data (word frequency), but other variables such as text color, font style and tag orientation can also be manipulated to encode additional data dimensions if needed (for example, see research conducted by Waldner et al. in the further reading section of our website).

Slide M5-8

Another common type of textual data visualization is **trees/hierarchies.** Tree-type structures or hierarchies in texts can be analyzed to gain insights on their characteristics. One way to visualize such relationships in texts is a **word tree**.

A word tree is a type of visualization that displays the connection of words in a text in a tree-like structure. It is often used to show the different contexts in which a word or phrase appears and can reveal themes and recurrences.

The example here on the slide is a word tree of Martin Luther King's "I have a dream" speech made by Wattenberg and Viégas. The main "root" of the tree is the word "I", and the "branches" of the tree are all the words or phrases that follow the word "I" in the speech. In a word tree, words that show up more frequently in combination with the selected "root" word or phrase are displayed in larger font size. As we can see, in this word tree, the phrase "have a dream" most frequently follows the root word "I". If we go further down this branch, we can see that multiple sentences in the speech begin with "I have a dream that one day". Therefore, this word tree visually displays King's use of repetition in his speech to build momentum and emotion.

Slide M5-9

Networks are also common among visualizations.

- Often textual data contains web-like relationships among items that cannot be conveniently captured with a tree structure, and in such cases networks may be used to visualize these complex relationships.
- A **node-link diagram** is a commonly-used type of visualization for representing networks. It captures named entities (such as people, places, and topics) as nodes (also called “vertices”) and relationships as links (also called “edges”), with a circle or dot representing a node, and a line representing a link.
- The figure on the screen is a node-link diagram of topics in English books from 1850-1899 by Ted Underwood.
- In it, each node (dot) represents a topic he uncovered using topic modeling on his dataset. The color of the dots represents genres (poetry, drama, fiction, nonfiction, and biography). He correlated the topics to one another to see what words the topics had in common. The visualization shows those connections between correlated topics.

Slide M5-10

Another common category is **temporal/spatial-based visualizations**, and they are used when textual data is tied to temporal or spatial elements that researchers want to highlight.

- A common temporal visualization is a timeline (like those we often see in history books), but here’s another example. The visualization on the slide shows the percentage representation of female characters in English-language fiction over time.

Slide M5-11

Additionally, maps are a kind of temporal/spatial-based visualization especially popular in digital humanities.

- Spatial-based data can usually be visualized with various kinds of thematic **maps**.
- The map on the screen shows the percent of newspapers in the US containing the term “hoosier.” Notice the concentration in Indiana!

Slide M5-12

Textual data can also be visualized in many other ways to display other kinds of “multi-dimensional” relationships between the attributes associated with the data, and we will call them **other “multi-dimensional” visualizations** here for the ease of our discussion. Familiar

visualizations such as pie charts and bar charts are all under this category. For textual data, **bubble charts** and **heat maps** are common types of visualizations that can be used.

- **Bubble charts** can display three dimensions of data, with one variable plotted on the x-axis, another one on the y-axis, and the third one represented by the size of the bubbles. The example on the slide visualizes the readability of U.S. presidential speeches by the Guardian.
- *Each speech is represented by a colored bubble – the larger the bubble, the longer the speech. Clusters of same-colored bubbles are speeches made by the same president (for example, the small cluster of dark blue bubbles on the far right of the visualization are all speeches made by Obama). The vertical axis represents the reading level of the speeches – the higher a bubble is placed on the graph vertically, the more difficult the speech is to read. The horizontal axis represents the time the speeches were made – the later the speech was made, the bubble representing the speech is placed more near the right side of the graph.*

Slide M5-13

Heat maps can also visualize three dimensions of data, this time with the third dimension being the shade of color.

- The example on the slide shows a heat map of MARC cataloging at the Library of Congress by book year and cataloging year, Ben Schmidt. *The year the item was published is on the vertical axis, the year the MARC record was created is on horizontal axis, and the color relates to the number of books cataloged. As we can see, many MARC catalog records were created in the early days of MARC as backlogs were cleared. Over time, only more recently published items were cataloged with MARC.*

Slide M5-14

Now let's do a short activity. Think about the kinds of visualizations we have looked at so far. Working alone or with your partner, can you match the kinds of visualizations to the various cases when they might be used? There isn't necessarily a 1-to-1 match here. We'll discuss when you're finished.

If you have time, see if you can think about the kinds of variables (data points like location, data, etc) you would need for each kind of visualization.

Slide M5-15

There are a number of easy-to-use web-based tools for creating visualizations of textual data. The following tools are suitable for beginners to explore some basic types of visualizations.

- To create word clouds, you can try Voyant and Wordle. **Voyant** can create many types of visualizations including word clouds, bubble charts, networks, and word trees, and has a user-friendly interface that works great as a learning tool. **Wordle** only creates word clouds and provides less opportunities for fine-tuning your results, but can still be helpful in getting a general idea of the text.
- To visualize word usage trends, **Google Books Ngram Viewer** and **HathiTrust+Bookworm** both enables users to search for words in corpora of texts and visualize their usage over time.
- For tabular data visualization, **Tableau** is a great choice.
- If you are interested in mapping, **ArcGIS Online/StoryMaps** can be used to incorporate GIS information and maps into interactive timelines and stories. Tableau also has similar functions.
- For making network graphs, we recommend using **Gephi, NodeXL,** or **DH Press**.

Slide M5-16

The advanced researcher may choose from a number of do-it-yourself visualization strategies, which generally involve specific programming packages that will create charts and tables from a dataset.

- In Python, the package Pandas—which as you'll remember is good for working with data—has a plot() function that will make quick charts. Another Python package for visualizations is called ggplot. If you want to use ggplot, you'll have to make sure you have it installed and then call it in a script or program.
- One reason some researchers like to use R versus Python is because of its visualization capabilities. Some find it easier to visualize with R, and the package many use is called ggplot2. *There are wars in the data science and programming communities about R versus Python and ggplot versus ggplot2. It's not important to understand the details of this preference-battle, but it might be useful to know that opinions about which is better are strongly held.*

- Finally, there is a JavaScript library for visualizations called D3.js or just D3. It is a relatively straightforward program for making visualizations that can be displayed on the web.

Slide M5-17

In the next section, we will be introducing you to HathiTrust+Bookworm. But first, let's return briefly to a term we learned about earlier in the text pre-processing module.

- N-grams are a contiguous chain of n-items (i.e. words) where n is the number of items in the chain. Notice how in the bigram example on the screen the window of the gram slides across the text, so in bigrams, a word will occur with both the word preceding and following it.
- HathiTrust+Bookworm makes use of unigrams – or single words.

Slide M5-18

Bookworm is a tool that visualizes language usage trends in repositories of digitized texts in a simple and powerful way.

- It is a tool that facilitates the observation of chronological trends for words and phrases in large digitized collections of textual documents with metadata facets.
- **HathiTrust + Bookworm** can visualize word frequencies over time in texts from the HathiTrust Digital Library. Currently, only unigrams (single words) can be searched and visualized with HT+Bookworm.

Slide M5-19

Normal search engines are very good at finding individual texts: Bookworm is good at finding and understanding **categories** in a library.

- The simplest use case is returning **usage of a word across years**: the goal here is to understand something about the years through the words they use.
- Bookworm is a framework that can be applied to other corpora. The HT+BW version is just one implementation!

Slide M5-20

Here is an example of how Bookworm can be used to compared word usage trends. In this example, we can see some overall trends in the usage of the word "lady" when compared to that of the word "woman" over the years.

Slide M5-21

- To quantify the dynamics of language evolution, Aiden et al. studied the regularization of English verbs over hundreds of years. His group originally conducted a manual study first and published the results in Nature in 2007, and then proceeded to study this phenomenon further using computational methods in 2007.
- They called this approach Culturomics, defined as “the application of high-throughput data collection and analysis to the study of human culture”.
- One of their examples was “burned” vs “burnt.” Let’s try!

(Instructor should demo this segment live if possible and use the screenshots in the slides as backup)

- Search “burned” and “burnt” in Bookworm. They both work as the past tense and past participle of the verb “burn”, but has the usage of the two changed over time, respectively? After visualizing the usage of the two words in Bookworm, we can see that before the 1860s, “burnt” was used more than “burned”, but afterwards it became the opposite.
- You can try visualizing other sets of verbs in Bookworm. Do you see any trends?

Slide M5-22

Now let’s take a closer look at the Bookworm interface.

Link to HT+BW page: https://bookworm.htrc.illinois.edu/develop

- Begin a search by typing in a word in the search column. You can click on the plus and minus signs next to the search columns to add or reduce the number of columns to search for multiple words at the same time for comparison. By clicking on the funnel icon next to “All texts”, you can also limit your search with facets, for example “Language” or “Publication Country”.

Slide M5-23

By clicking on the “Dates”, “Metric”, and “Case” tabs on the top right corner of the page, you can also fine-tune your results by controlling the time period, changing metrics, switching between case-sensitive and case-insensitive, and smoothing with different parameter settings. When you’re all set, hit the “Search” button to generate the visualization.

Slide M5-24

If everything goes well, your visualization will appear on the screen. When you point your cursor to a specific spot on a curve, a window will display the option “click for texts”. After a single click, another window that shows the top results of your searched word in that year will appear. You can then click on the name of each entry to be directed to its corresponding volume in the HathiTrust Digital Library.

Slide M5-25

Let’s think about our sample reference question again. One possible approach is to explore usage trends of political concepts using HathiTrust+Bookworm.

Slide M5-26

For our hands-on activity, you will use HT+BW to visualize lexical trends related to your own research interests.

Access HT+BW from the link on the slide and try searching some terms. Be creative and feel free to experiment! Also, try using faceting to get more interesting results.

If time runs short, while introducing the activity, ask participants to share their results in pairs very briefly as soon as they finish exploring on their own. In this way, the discussion can be folded into the hands-on activity and take up less time.

Slide M5-27

(Optional example)

If you are interested in political concepts, here is one example to try. Use HT+BW to visualize “socialism” and “fascism”. Do you see any trends?

Slide M5-28

(Optional example)

Another example is graphing “nationalism” and “internationalism”. Do you see any trends?

Slide M5-29

(After 5-15 minutes of independent hands-on time.)

What trends did you discover? Would anyone like to share?

Slide M5-30

Let's check in on Sam. HT+BW was relatively important to Sam's project. He was able to track to usage of words like "creative" and "creativity" over time. Notice how steeply it rose! Sam's inclination about the rise of "creativity" matches the literature.

Slide M5-31

Sam used a few beta visualizations with HT+BW, to create different kinds of visualizations.

Slide M5-32

One he made was this heatmap view showing the relative prominence of "creativity" in the literature over time.

Slide M5-33

He also made this alternate heat map, which divides the corpus by Library of Congress subject heading. These views are examples of how Bookworm can display more than just word-frequency line graphs.

Slide M5-34

Finally, let's wrap-up with a discussion. Where does visual literacy fit into data literacy overall?

What would it mean to be visually literate, particularly with regard to text analysis?

If time runs short, skip this discussion and wrap up the workshop.

Slide M5-35

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

Slides M5-36 to M5-37
(Show references so attendees know where to find them later.)