

Welcome!

When you arrive:

- Sign in
- Pick up full handout
- Connect to Wi-Fi

Materials for Today

- **Workshop materials and resources:**
<http://go.illinois.edu/ddrf-curriculum>
- **HTRC and HTDL accounts:**
<https://analytics.hathitrust.org>
<https://www.hathitrust.org>
- **PythonAnywhere account:**
<https://pythonanywhere.com>

 See Handout pp. 1-2, Getting Set Up





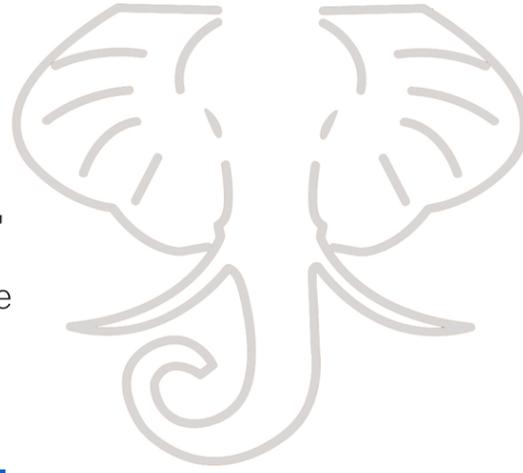
**Digging Deeper,
Reaching Further**

Text analysis with the HathiTrust Research Center



Digging Deeper, Reaching Further

Libraries Empowering Users to Mine the
HathiTrust Digital Library Resources



#RE-00-15-0112-15

<http://teach.htrc.illinois.edu>

GOALS:

- Arm librarians with instructional content and tool skills in digital scholarship and digital humanities;
- Empower librarians to become research partners on digital projects;
- Enable librarians to build foundations for digital scholarship centers and services

Thanks to all our team members!

- **PI**

Harriett Green, [University of Illinois at Urbana-Champaign Library](#)

- **Co-PIs**

Angela Courtney, [Indiana University-Bloomington Libraries](#)

J. Stephen Downie, [School of Information Sciences, University of Illinois](#)

Terese Heidenwolf, [Lafayette College Library](#)

Geoffrey Morse, [Northwestern University Libraries](#)

Amanda Henley, [University of North Carolina at Chapel Hill Library](#)

Project Members (Past and Present)

*University of Illinois at
Urbana-Champaign*

Eleanor Dickson Koehl

Ruohua Han

*Rachel Blomer

*Gabriella Stuardi

*Beth Sandore Namachchivaya

*Ryan Dubnicek

*Sayan Bhattacharyya

Indiana University

Leanne Nay

Robert McDonald

Nicholae Cline

*Ewa Zegler-Poleska

Lafayette College

Lijuan Xu

Ben Jahre

*Neil McElroy

*University of North Carolina at
Chapel Hill*

Sarah Morris

*Stewart Varner

*Jacob Hill

Northwestern University

*Geoffrey Swindells

* = former members



Getting started

☞ *See Handout pp. 1-2 for instructions*

- **Workshop materials and resources:**

<http://go.illinois.edu/ddrf-curriculum>

- **HTRC Analytics and the HTDL:**

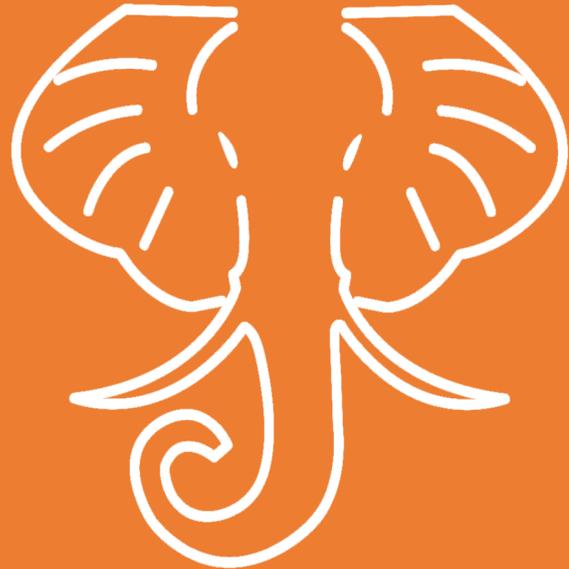
<https://analytics.hathitrust.org> | <https://www.hathitrust.org>

- **PythonAnywhere:**

<https://pythonanywhere.com>

- *Friendly reminder: Please **do not use Internet Explorer** for our activities today!*





**Digging Deeper,
Reaching Further**

Module 1: Getting Started

In this module we'll...

- Introduce text analysis and broad text analysis workflows
 - *Make sense of digital scholarly research practices*
- Introduce HathiTrust and the HathiTrust Research Center
 - *Understand the context for one text analysis tool provider*
- Introduce our hands-on example and case study
 - *Recognize research questions text analysis can answer*



What is text analysis?

- Using computers to reveal information in and about text (Hearst, 2003)
 - Algorithms discern patterns
 - Text may be “unstructured”
 - More than just search
- What is it used for?
 - Seeking out patterns in scientific literature
 - Identifying spam e-mail



How does it work?

- Break textual data into smaller pieces
- Abstract (reduce) text so that a computer can crunch it
- Counting!
 - Words, phrases, parts of speech, etc.
- Computational statistics
 - Develop hypotheses based on counts of textual features



How does it impact research?

- Shift in perspective, leads to shift in research questions
 - Scale-up to “distant reading” (Moretti, 2013)
- One step in the research process
 - Can be combined with close reading
- Opens up:
 - Questions not provable by human reading alone
 - Larger corpora for analysis
 - Studies that cover longer time spans



Discussion

- *What examples have you seen of text analysis?*
- *In what contexts do you see yourself using text analysis? What about the researchers you support?*



Text analysis research questions

- May involve:
 - Change over time
 - Pattern recognition
 - Comparative analysis



Hands-on activity

☞ *See Handout p. 3*

In pairs or small groups, review the summarized research projects available at <http://go.illinois.edu/ddrf-research-examples>. Then discuss the following questions:

- How do the projects involve change over time, pattern recognition, or comparative analysis?
- What kind of text data do they use (time period, source, etc.)?
- What are their findings?



Example: *Rowling and “Galbraith”*: an authorial analysis

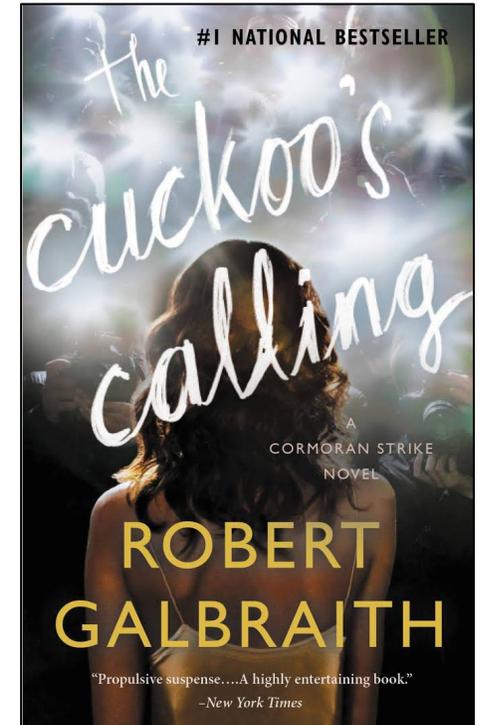
Question:

Did JK Rowling write The Cuckoo’s Calling under the pen name Robert Galbraith?

Would be impossible to prove through human reading alone!

comparative | patterns

Read more: Rowling and “Galbraith”: an authorial analysis (Juola, 2013)



Book cover for The Cuckoo’s Calling



Example: *Rowling and “Galbraith”*: an authorial analysis

Approach:

- Reading led to hunch about authorship
- Computational comparison of diction between this book and others written by Rowling
- Statistical ‘proof’ of authorial fingerprint

Read more: Rowling and “Galbraith”: an authorial analysis (Juola, 2013)



Example: *Significant Themes in 19th Century Literature*

Question:

What themes are common in 19th century literature?

Answering this question requires a very large corpus and an impossible amount of human reading!

patterns | comparative



Example: *Significant Themes in 19th Century Literature*

Approach:

- Run large quantities of text through a statistical algorithm
- Words that co-occur are likely to be about the same thing
- Co-occurring words are represented as topics

Read more: Significant Themes in 19th Century Literature (Jockers and Mimno, 2012)



Example: *The Emergence of Literary Diction*

Question:

What textual characteristics constitute “literary language”?

This question covers a very large time span!

change over time | patterns

Read more: *The Emergence of Literary Diction* (Underwood and Sellers, 2012)



Example: *The Emergence of Literary Diction*

Approach:

- Train a computational model to identify literary genres
- Compare which words are most frequently used over time in non-fiction prose versus “literary” genres
- Demonstrated tendency for poetry, drama, and fiction to use older English words

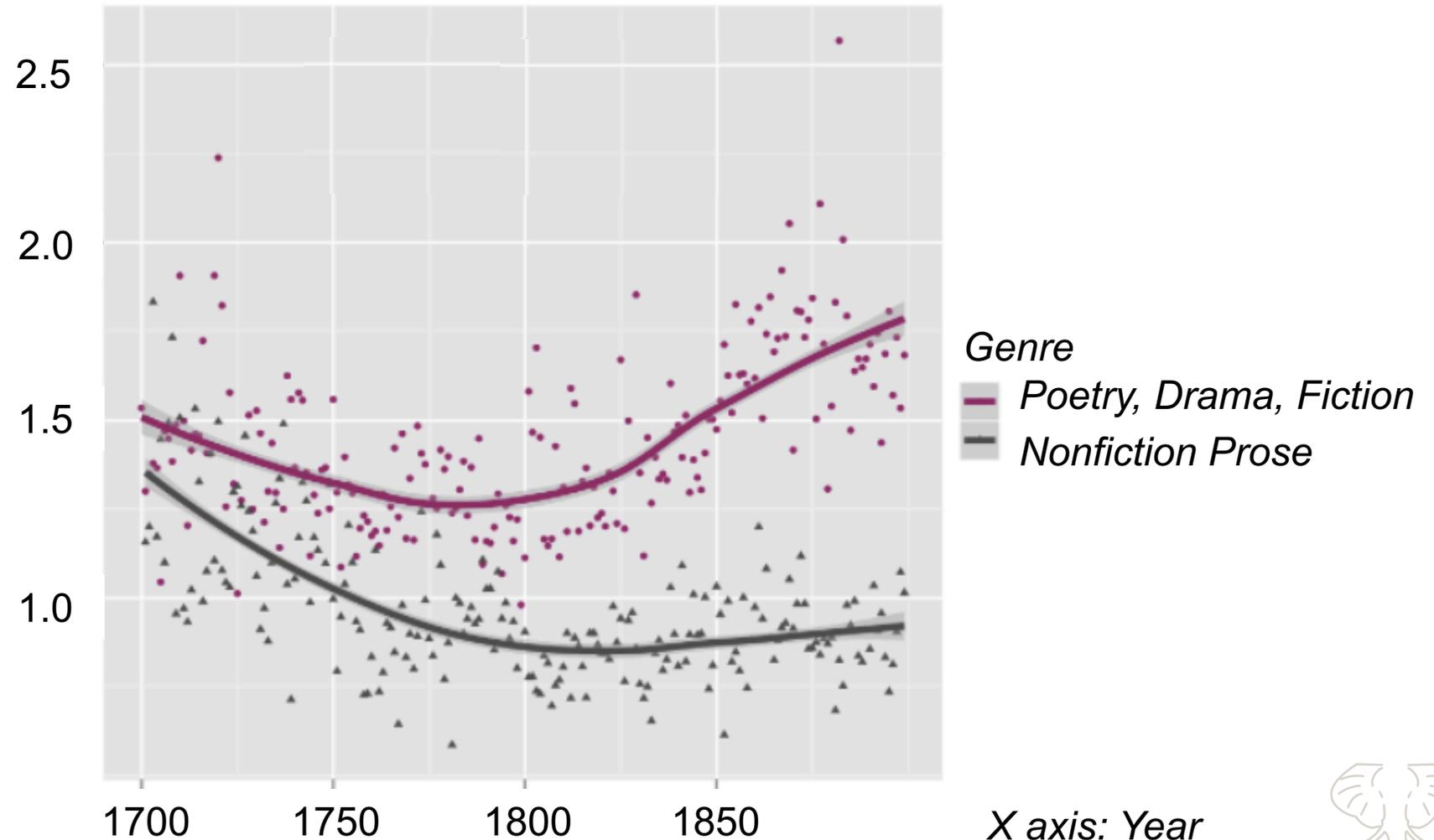
Read more: *The Emergence of Literary Diction* (Underwood and Sellers, 2012)



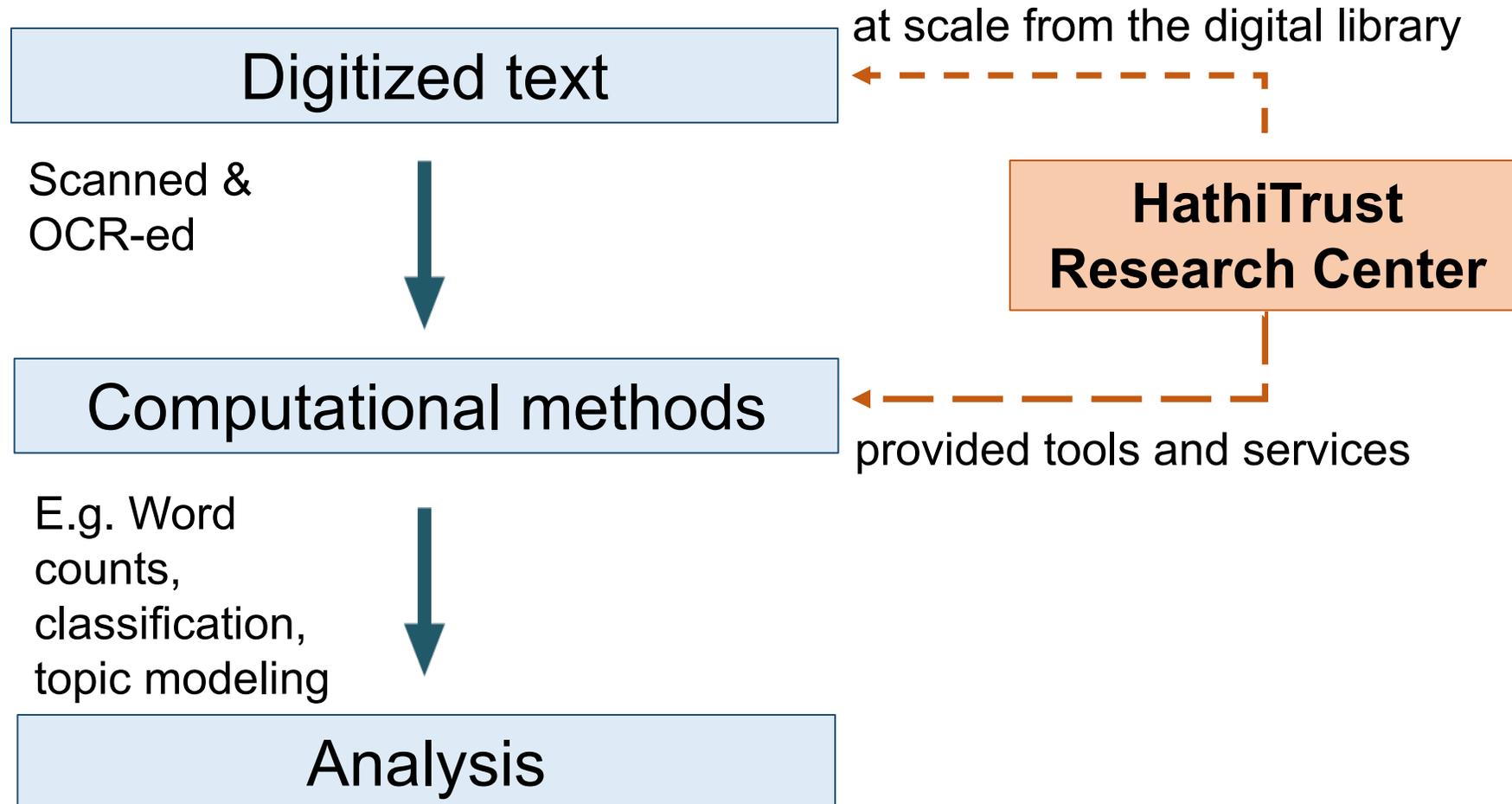
Example: *The Emergence of Literary Diction*

Y axis: Yearly ratio of words that entered English before 1150 / words that entered from 1150-1699

From paper: graph of diction patterns between genres, using frequency counts



HTRC for text analysis



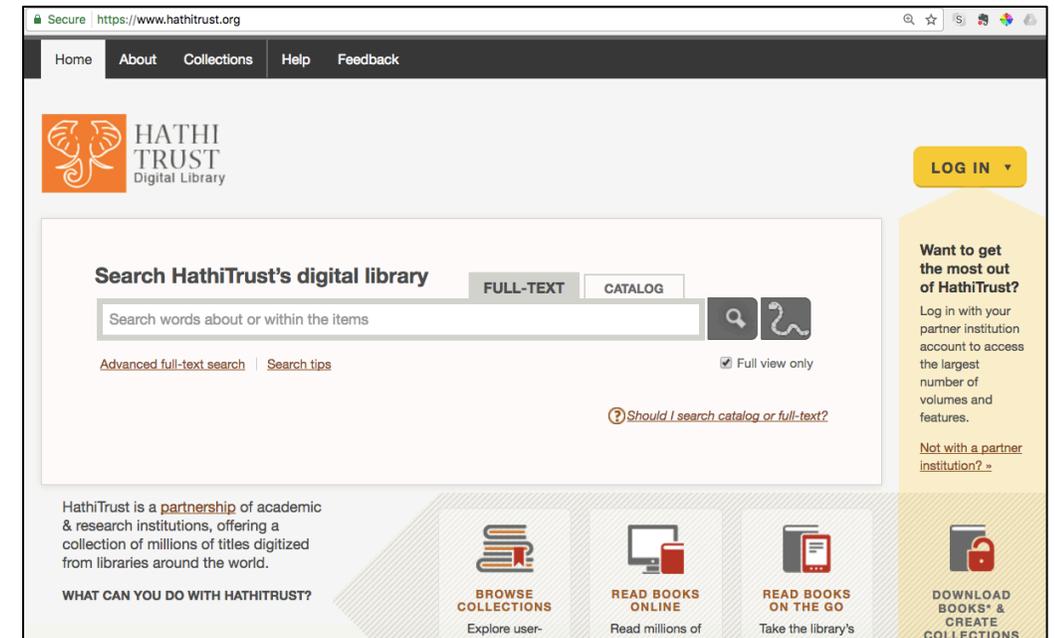
HathiTrust

- Founded in 2008
- Grew out of large-scale digitization initiative at academic research libraries
 - With roots in Google Books project
- Over 120 partner institutions continue to contribute



HathiTrust Digital Library

- Contains over 16 million volumes
 - ~ 50% English
 - From the 15th to 21st century, 20th century concentration
 - ~ 63% in copyright or of undetermined status
- Search and read books in the public domain



HathiTrust Research Center

- Facilitates text analysis of HTDL content
- Research & Development
- Located at Indiana University and the University of Illinois



Non-consumptive research

Research in which computational analysis is performed on text, but not research in which a researcher reads or displays substantial portions of the text to understand the expressive content presented within it.

- Complies with copyright law
- Foundation of HTRC work
- Other terms: non-expressive use



Discussion

Are you (or your colleagues) currently offering research support for text analysis?

- How so?
- Why or why not?
- What kinds of questions and/or projects does your library handle?



Workshop outline

- Follow the research process:
 - Gathering textual data: 2 modules
 - Working with textual data: 1 module
 - Analyzing textual data: 2 modules
 - Visualizing textual data: 1 module
- Hands-on activities around a central research question & case study example at each step
 - Using both HTRC and non-HTRC tools



Workshop outline

- Build skills to engage with text analysis research
- Covers programming concepts
 - But won't teach you to code!
- Introduces computational methods
 - But won't delve into all nuances



Sample Reference Question

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:

- We'll practice approaches for answer this question throughout the workshop



Case Study

Inside the Creativity Boom | Researcher: Samuel Franklin

Question:

How do the use and meaning of creative and creativity change over the 20th century?

Approach:

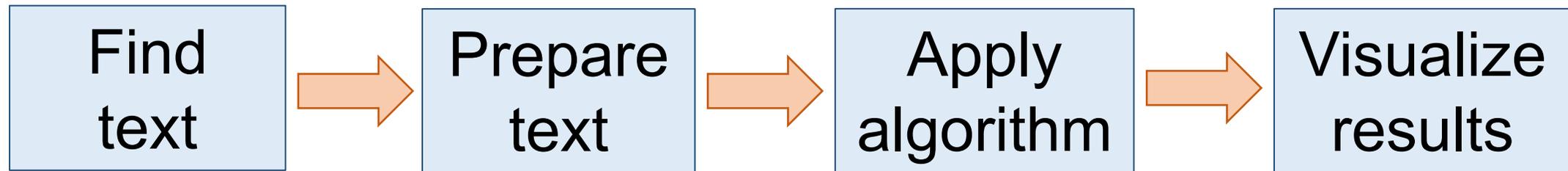
- We'll discuss how this researcher approached his question throughout the workshop

Learn more: <https://wiki.htrc.illinois.edu/x/CADiAQ>



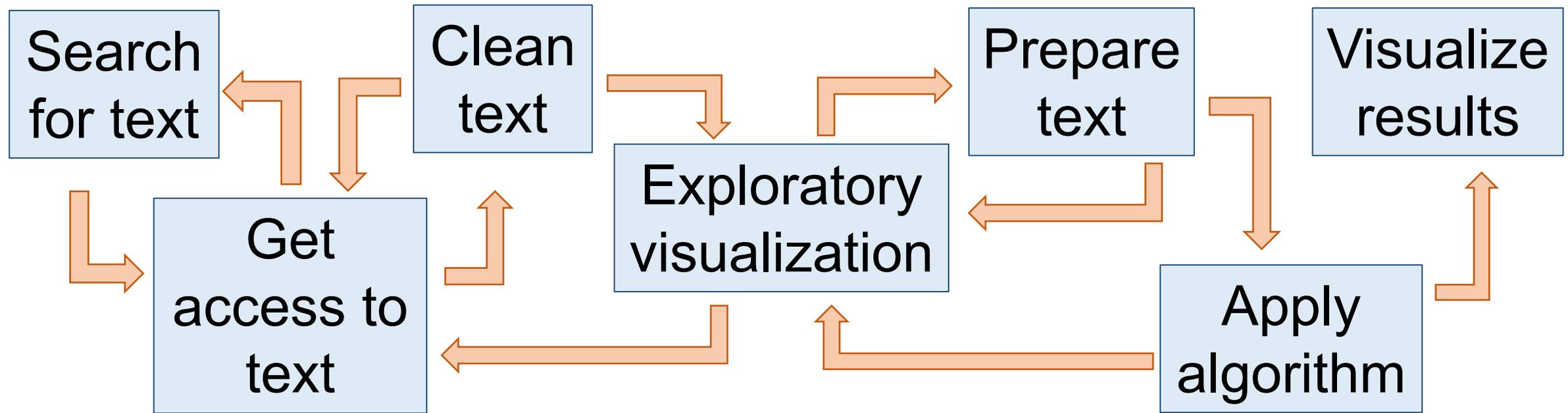
A word of caution...

Workshop outline suggests research workflow like:



A word of caution...

Actual research workflow like:



Discussion

- *What are some of the characteristics of a good candidate research question/project for using text analysis methods?*



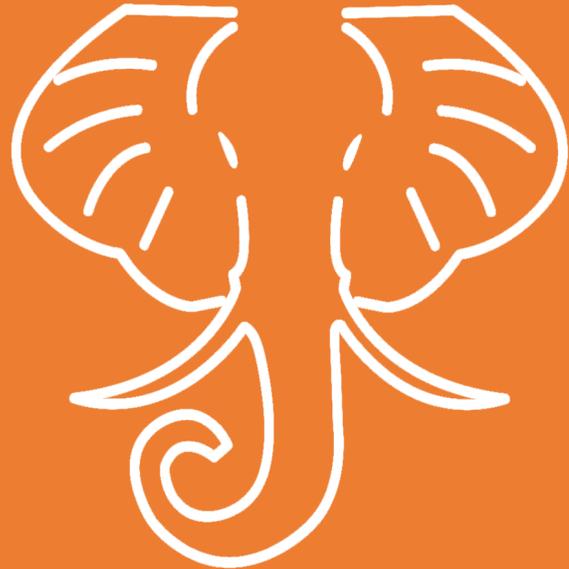
Questions?



References

- Hearst, M. (2003). What is text mining. SIMS, UC Berkeley.
<http://people.ischool.berkeley.edu/~hearst/text-mining.html>
- Jockers, M. L., & Mimno, D. (2012). Significant themes in 19th-century literature. [pre-print]
<http://digitalcommons.unl.edu/englishfacpubs/105/> .
- Juola, P. Language Log » Rowling and “Galbraith”: an authorial analysis. July 16, 2013.
Retrieved January 25, 2017, from <http://languagelog.ldc.upenn.edu/nll/?p=5315>
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Underwood, T., & Sellers, J. (2012). The emergence of literary diction. *Journal of Digital Humanities*, 1(2), 1-2. <http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/> .





**Digging Deeper,
Reaching Further**

Module 2.1: Gathering Textual Data

Finding and Curating Text Data

In this module we'll...

- Explore the concept of a text data and where to find it
 - *Provide data reference for researchers*
- Build a HathiTrust workset
 - *Gain experience in building a textual dataset*
- Learn how Sam built a *Creativity Corpus* of HathiTrust volumes
 - *Understand real-world data collection strategies*



Where we'll end up

poli_science_DDRF

[Download](#)

Description : Political science collection for DDRF workshop

Owner	Last Modified Time	Number of Volumes	Tags
rhan11	2017-10-05T18:21:35Z	16	

Filter volume by title...

Volume ID	Title	Authors	Year	Language
mdp.49015002203223	Public papers of the presidents of the United States.	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972; Hoover, Herbert 1874-1964; United States Federal Register Division; United States Office of the Federal Register	1978	eng
mdp.49015002203272	Public papers of the presidents of the	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972;	1979	eng

Create a collection of volumes from the HathiTrust Digital Library and prepare it for analysis in HTRC Analytics as a workset



Kludging access

“Text analysis projects share in common 3 challenges. **First**, data of interest must be found. **Second**, data must be gettable. **Third**, if it’s not already formed according to wildest dreams, ways must be known of getting data into a state that they are readily usable with desired methods and tools.”

Kludging: Web to TXT (Padilla, 2015)

<http://www.thomaspadilla.org/2015/08/03/kludge/>



Finding text

- Not always easy
 - copyright restrictions
 - licensing restrictions
 - format limitations
 - hard-to-navigate systems

** issues more pronounced at scale**



Vendor databases

- Be aware of licensing restrictions
- Strategies
 - Addendums to libraries' contracts
 - Vendor-provided services
 - Asking for special permission case-by-case
- Example: JSTOR Data for Research



Library/archives digital collections

- Wealth of material, but:
 - Often siloed
 - Access not formulated for research at scale
- Things to look for:
 - Plain text
 - Bulk download
- Example: UNC's DocSouth Data



Social media

- Popular with social science researchers
- To access:
 - Some provide systems to access text
 - Or there are 3rd-party tools on the market
- Example: Twitter API (Application Programming Interface)



Hands-on activity

 See Handout p. 4

Building a corpus for political history, what are the strengths and weaknesses of each of these broad sources for textual data?

	Strengths	Weaknesses
Vendor database		
Library/archives digital collections		
Social media		



Evaluating sources of text data

Does the researcher already have a data source in mind?

Is the text they want to use already digitized?

Are there copyright and licensing concerns?

How technically experienced is the researcher?

What is the period, place, person of interest?

How much flexibility is needed for working with the data?

Does the researcher have funding?

What format does the researcher expect the data in?



Building corpora

- Identify texts through full text search
 - Use a key term or phrase
- Identify texts through metadata
 - Search by certain author(s)
 - Search within a date range
 - Search for a specific genre
- Or some combination of the two!



Building corpora

- Process usually involves deduplication
- What to keep/discard is project dependent
- Examples of deduplication:
 - OCR quality
 - Earliest edition
 - Editions without forewords or afterwords



HTRC Worksets

- User-created collections of text from the HathiTrust Digital Library
 - think of them as textual datasets
- Can be shared and cited
- Suited for non-consumptive access



HTRC Worksets

poli_science_DDRF

[Download](#)

Description : Political science collection for DDRF workshop

Owner	Last Modified Time	Number of Volumes	Tags
rhan11	2017-10-05T18:21:35Z	16	

Filter volume by title...

Volume ID	Title	Authors	Year	Language
mdp.49015002203223	Public papers of the presidents of the United States.	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972; Hoover, Herbert 1874-1964; United States Federal Register Division; United States Office of the Federal Register	1978	eng
mdp.49015002203272	Public papers of the presidents of the	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972;	1979	eng

Workset viewed on the web

mdp.49015002221845
mdp.49015002221837
mdp.49015002221829
mdp.49015002221787
mdp.49015002221811
mdp.49015002221761
mdp.49015002221779
mdp.49015002203140
mdp.49015002203157
mdp.49015002203033
mdp.49015002203231
mdp.49015002203249
mdp.49015002203223
mdp.49015002203405
mdp.49015002203272
mdp.49015002203215

Workset manifest



Building worksets

- Stored in HTRC
 - Require account with university email address
- Ways to build:
 - Import from HT Collection Builder
 - Compile volume IDs elsewhere



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:

- Create a textual dataset of volumes related to political speech in America with the HT Collection Builder, and upload it to HTRC Analytics as a workset for analysis



Hands-on activity

 *See Handout pp. 5-6*

In this activity, you will log in to HTDL and create a collection containing volumes of the public papers of the presidents of the United States, and import it into HTRC Analytics as a workset. Follow the instructions on the handout to build your workset.

Websites:

- HTDL: <https://www.hathitrust.org>
- HTRC Analytics: <https://analytics.hathitrust.org>



Go to HTDL interface

The screenshot displays the HathiTrust Digital Library homepage. At the top, a dark navigation bar contains links for Home, About, Collections, Help, and Feedback. Below this is the HathiTrust logo, which consists of an orange elephant head icon and the text 'HATHI TRUST Digital Library'. To the right of the logo is a yellow 'LOG IN' button with a dropdown arrow.

The main content area features a search bar with the text 'Search HathiTrust's digital library'. The search bar has two tabs: 'FULL-TEXT' and 'CATALOG', with 'CATALOG' currently selected. The search input field contains the placeholder text 'Search words about the items'. To the right of the input field is a dropdown menu set to 'All Fields' and a 'Search' button with a magnifying glass icon. Below the search bar are links for 'Advanced catalog search' and 'Search tips', and a checkbox for 'Full view only'. A help icon and the text '? Should I search catalog or full-text?' are also present.

Below the search bar, there is a paragraph: 'HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.'

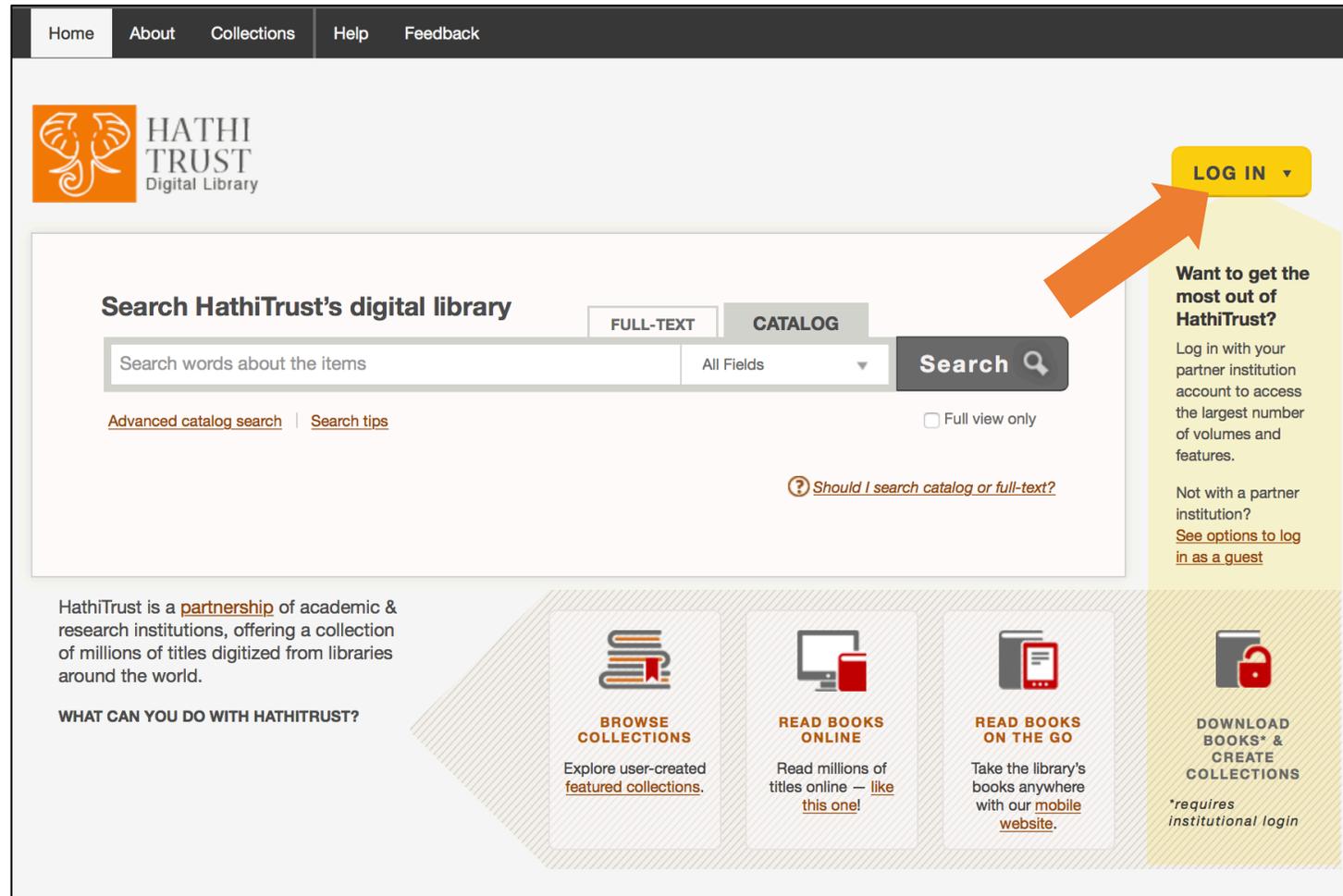
Underneath this paragraph is the heading 'WHAT CAN YOU DO WITH HATHITRUST?' followed by four featured collection cards:

- BROWSE COLLECTIONS**: Explore user-created featured collections.
- READ BOOKS ONLINE**: Read millions of titles online — like this one!
- READ BOOKS ON THE GO**: Take the library's books anywhere with our mobile website.
- DOWNLOAD BOOKS* & CREATE COLLECTIONS**: *requires institutional login

On the right side of the page, there is a yellow callout box with the heading 'Want to get the most out of HathiTrust?'. It contains the text: 'Log in with your partner institution account to access the largest number of volumes and features. Not with a partner institution? See options to log in as a guest'.



Log in



The screenshot shows the HathiTrust Digital Library homepage. At the top, there is a navigation menu with links for Home, About, Collections, Help, and Feedback. Below the menu is the HathiTrust logo, which consists of an orange elephant head icon and the text 'HATHI TRUST Digital Library'. The main content area features a search bar with the text 'Search HathiTrust's digital library'. The search bar has two tabs: 'FULL-TEXT' and 'CATALOG'. Below the search bar is a search input field with the placeholder text 'Search words about the items', a dropdown menu for 'All Fields', and a 'Search' button with a magnifying glass icon. There are also links for 'Advanced catalog search' and 'Search tips', and a checkbox for 'Full view only'. A question mark icon with the text 'Should I search catalog or full-text?' is also present. On the right side, there is a yellow callout box with the text 'Want to get the most out of HathiTrust?' and a 'LOG IN' button with a dropdown arrow. The callout box contains the text: 'Log in with your partner institution account to access the largest number of volumes and features. Not with a partner institution? See options to log in as a guest'. At the bottom, there is a section titled 'WHAT CAN YOU DO WITH HATHITRUST?' with four columns of content: 'BROWSE COLLECTIONS' (Explore user-created featured collections), 'READ BOOKS ONLINE' (Read millions of titles online — like this one!), 'READ BOOKS ON THE GO' (Take the library's books anywhere with our mobile website), and 'DOWNLOAD BOOKS* & CREATE COLLECTIONS' (*requires institutional login).



Log in

LOG IN ▾

Trust's digital library

Find your partner institution:

University of Illinois at Urbana-Champaign ▾

CONTINUE →

[Why isn't my institution listed?](#)

Not with a partner institution?
[See options to log in as a guest](#)

BROWSE COLLECTIONS
Explore user-created [featured collections](#).

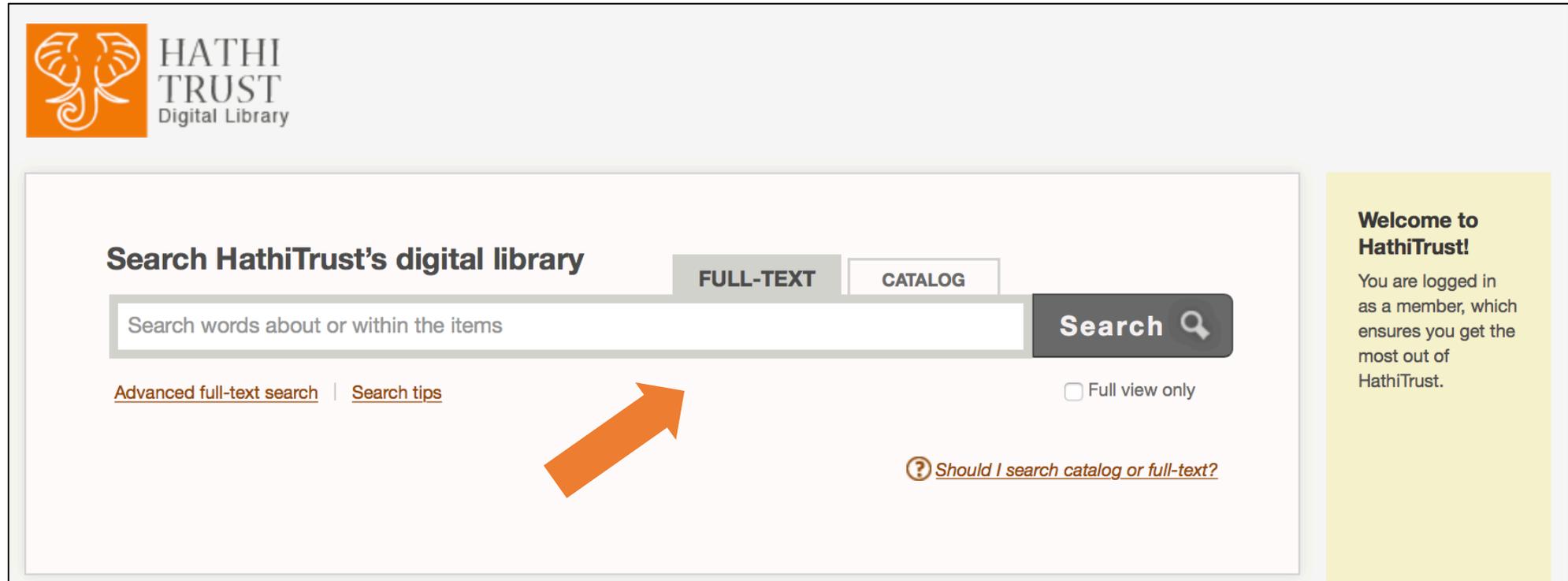
READ BOOKS ONLINE
Read millions of titles online — [like this one!](#)

READ BOOKS ON THE GO
Take the library's books anywhere with our [mobile](#)

DOWNLOAD BOOKS* & CREATE COLLECTIONS
*requires institutional login



Search for volumes



HATHI TRUST
Digital Library

Search HathiTrust's digital library

FULL-TEXT **CATALOG**

Search words about or within the items

Search 🔍

[Advanced full-text search](#) | [Search tips](#)

Full view only

🔗 [Should I search catalog or full-text?](#)

Welcome to HathiTrust!

You are logged in as a member, which ensures you get the most out of HathiTrust.

- Click on “Advance full-text search”



Search for volumes

Advanced Full-text Search :

Search information *within or about* an item

[Search Tips](#)

Prefer to search items in an [Advanced Search?](#)

in

in

[+ Add a pair of search fields](#)

Limit to:

Full view only Year of publication:

Language Limit to Original Format



Filter results and select volumes

Filter results on the left sidebar

Select all or some of the returned search items for your collection.

The screenshot shows a search results interface. On the left is a 'Refine Results' sidebar with filters for Subject, Language, Place of Publication, and Date of Publication. The main area shows 'Search Results: 1,729 items found' with an advanced search query: 'this exact phrase: United States in Title' AND 'this exact phrase: public papers in Title'. Below the search bar are options to 'Revise this advanced search', view counts for 'All Items (1,729)' and 'Full View (1,583)', and pagination controls. A list of search results is displayed, each with a checkbox for selection. The first result is 'The public papers and addresses of Franklin D. Roosevelt. 1943 volume...' and the second is 'Public papers of the presidents of the United States. 1987 pt.2'. Each result includes a 'Catalog Record' link and a 'Limited (search-only)' or 'Full view' link.

An advanced search for volumes that contain all the words/phrases below in the title field: “public papers” and “United States”



Add volumes to collection

The screenshot shows a digital library interface. At the top, there are two tabs: "All Items (1,607)" and "Full View (1,549)". Below the tabs is a pagination control showing "25 per page" and a list of page numbers from 1 to 65, with a "Next" button. A grey bar contains a "Select all on page" checkbox, a "[CREATE NEW COLLECTION]" dropdown menu, and an "Add Selected" button. An orange arrow points from the "Select all on page" checkbox to the "[CREATE NEW COLLECTION]" dropdown. Below this bar is a list of items. The first item is "Public papers of the presidents of the United States. 1989 pt. 1" by United States. President, published 1989, with a checked checkbox and links for "Catalog Record" and "Full view". The second item is "Herbert Hoover proclamations and executive orders, March 4, 1929 to March 4, 1933. proc.001" by Hoover, Herbert, 1874-1964, published 1974, with an unchecked checkbox and links for "Catalog Record" and "Full view". The third item is "Public papers of the presidents of the United States. 1988 pt.1" by United States. President, with a checked checkbox.

Once texts are selected, click “Select Collection” → choose “[CREATE NEW COLLECTION]” → click “Add Selected”



Add collection metadata

Start Over this exact phrase: public papers in Title

AND this exact phrase: United States in Title all of these words: president in Title

Collection Name 85

Description 174

Private Public

Cancel Save Changes

[Catalog Record](#) [Full view](#)

Herbert Hoover proclamations and executive orders, March 4, 1929 to March 4, 1933. proc.001
by Hoover, Herbert, 1874-1964.
Published 1974



View your collection

The screenshot shows the HathiTrust Digital Library interface. At the top, there is a navigation bar with links for Home, About, Collections, Help, Feedback, Member (University of Illinois at Urbana-Champaign), My Collections, and Logout. An orange arrow points to the 'My Collections' link. Below the navigation bar is the HathiTrust logo and a search bar containing the text 'public papers'. To the right of the search bar are buttons for 'FULL-TEXT' and 'CATALOG', and a search icon. Below the search bar are links for 'Advanced full-text search' and 'Search tips', and a checkbox for 'Full view only'. A green notification bar at the top of the results area states '2 items were added to PoliticalSpeech'. The main content area displays 'Search Results: 1,607 items found'. Below this are several filters: 'Start Over', 'this exact phrase: public papers in Title', 'AND this exact phrase: United States in Title', and 'all of these words: president in Title'. A 'Revise this advanced search' button is also present. At the bottom of the results area, there are buttons for 'All Items (1,607)' and 'Full View (1,549)'. Below these are pagination controls showing '25 per page' and a list of page numbers from 1 to 65, with a 'Next' button. At the very bottom, there are controls for 'Select all on page', a dropdown menu for '[CREATE NEW COLLECTION]', and an 'Add Selected' button.



View your collection

Showing 5 of your collections [Reset](#)

[1930s political speeches DDRF](#)

1930s political speeches collection for DDRF workshop 5 items
last updated: 08/31/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Public : Make Private](#) [Delete Collection](#)

[1970s political speeches DDRF](#)

1970s political speeches collection for DDRF workshop 16 items
last updated: 08/31/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Public : Make Private](#) [Delete Collection](#)

[PoliticalSpeech](#)

A collection of volumes of public speeches by the presidents of the United States 2 items
last updated: 10/05/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Private : Make Public](#) [Delete Collection](#)



A collection of mostly 19th-20th-century musical score women composers held at University of Michigan Mus

UM Press



The Univ. of Michigan Press available in HathiTrust

University Press of Flori



Selected publications of the



Grab the collection URL



https://babel.hathitrust.org/cgi/mb?a=listis;c=1848985365

Home About Collections Help Feedback

HATHI TRUST Digital Library

FULL-TEXT CATALOG

Search words about or within the items

Advanced full-text search | Search tips

Full view only

Share

Link to this collection

<https://babel.hathitrust.org/cgi/mb?a=l>

Download Metadata | Convert to HTRC Workset

About this collection

Owner
Ruohua Han

Status
public

poli_science_DDRF

Political science collection for DDRF workshop

Search in this collection Find

All Items (16)

Sort by: Title A-Z | 25 per page | 1

Select all on page Select Collection Add Selected

Public papers of the presidents of the United States. 1971
by United States. President.
Published 1971

In my collections: ---

[Catalog Record](#) [Full view](#)
[Download Extracted Features](#)



Go to HTRC Analytics

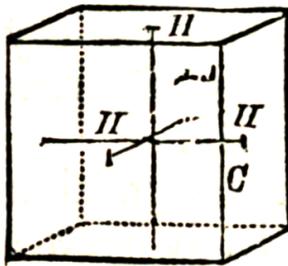
HTRC Analytics Algorithms Data Capsules Worksets Datasets Explore Help About Sign In Sign Up



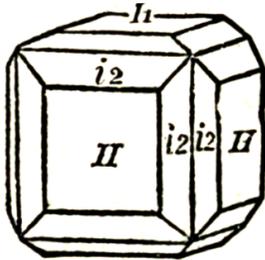
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

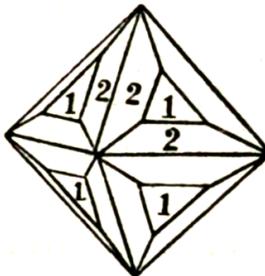
Featured Services



Extracted Features



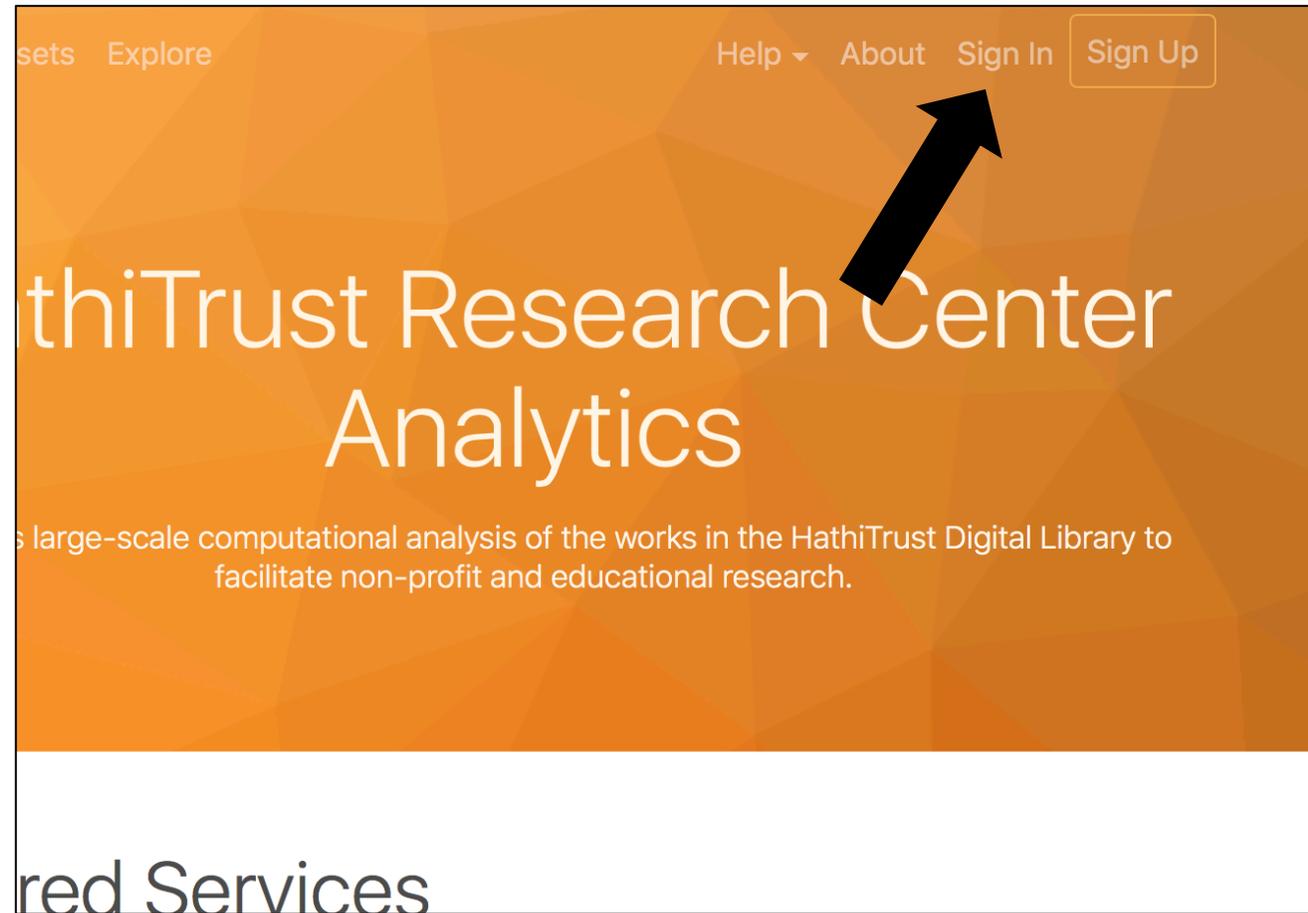
Text Analysis Algorithms



Data Capsules



Sign in / Sign up



Sign in

Welcome! Returning users signing into the new HTRC Analytics interface for the first time must reset their password using the "Forgot Password" link below.

Sign In to HathiTrust Research Center

Username

Password

Remember me on this computer

SIGN IN

[Forgot Password?](#) | [Forgot Username?](#) | [Create Account](#)

HathiTrust Research Center | © 2017



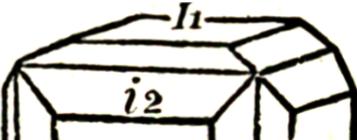
Sign in

es Worksets Datasets Explore Help About rhan11

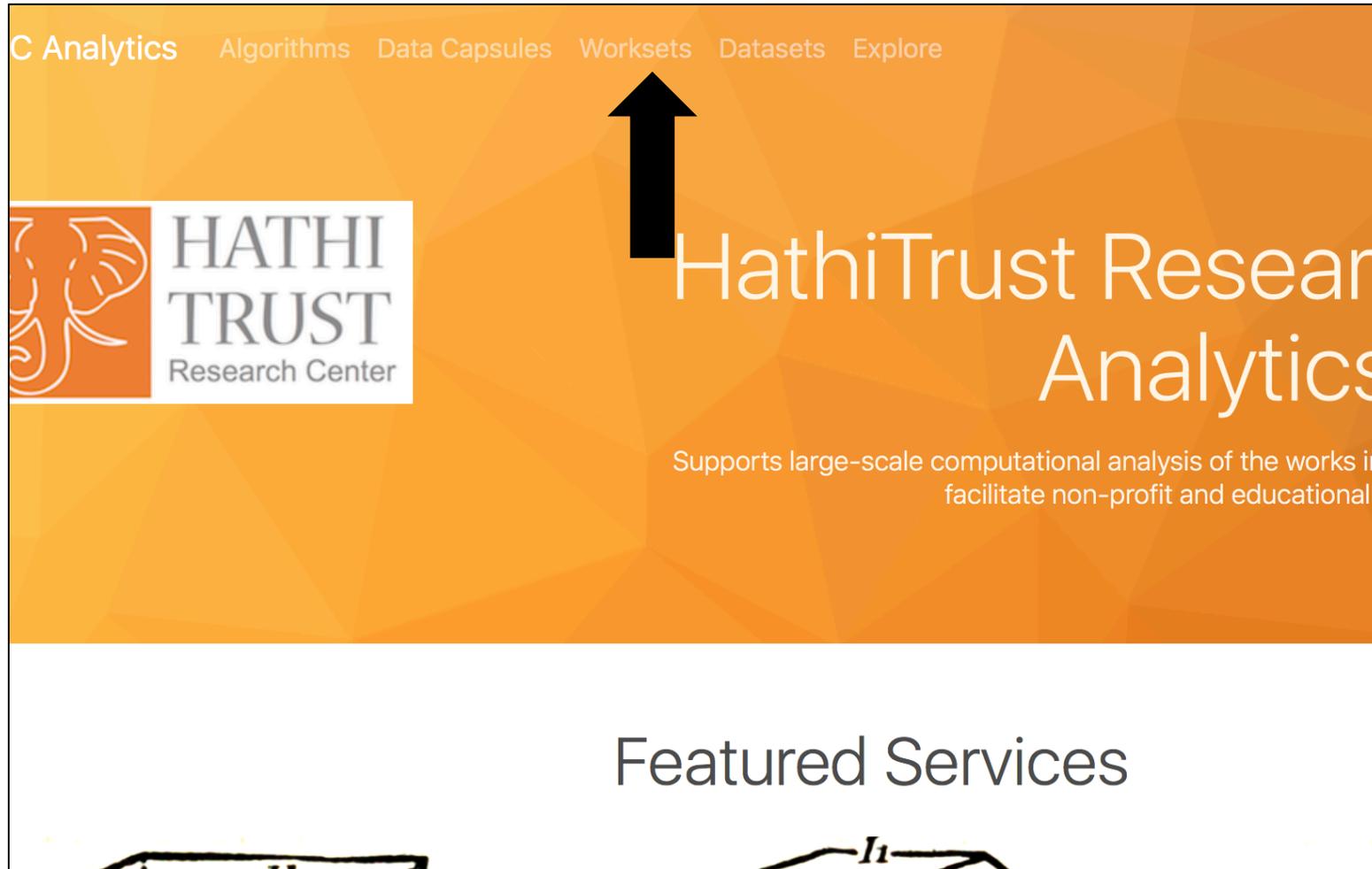
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

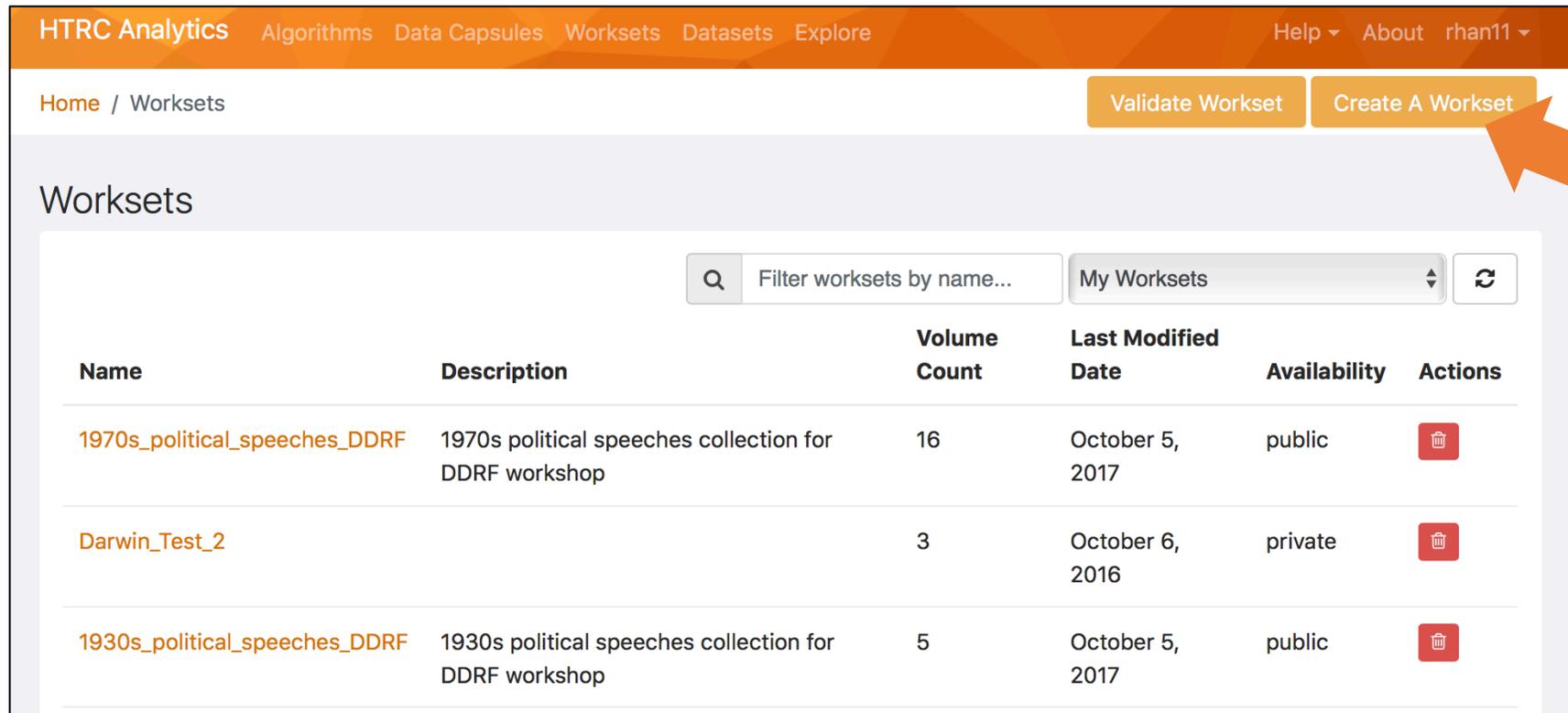
Featured Services



Go to Worksets page



Choose to create a workset



The screenshot shows the HTRC Analytics interface. The top navigation bar includes 'HTRC Analytics', 'Algorithms', 'Data Capsules', 'Worksets', 'Datasets', and 'Explore'. On the right, there are links for 'Help', 'About', and the user 'rhan11'. Below the navigation bar, the breadcrumb 'Home / Worksets' is visible. Two buttons, 'Validate Workset' and 'Create A Workset', are located in the top right of the main content area. An orange arrow points to the 'Create A Workset' button. The main content area is titled 'Worksets' and contains a search bar with the text 'Filter worksets by name...', a dropdown menu set to 'My Worksets', and a refresh icon. Below this is a table with the following data:

Name	Description	Volume Count	Last Modified Date	Availability	Actions
1970s_political_speeches_DDRF	1970s political speeches collection for DDRF workshop	16	October 5, 2017	public	
Darwin_Test_2		3	October 6, 2016	private	
1930s_political_speeches_DDRF	1930s political speeches collection for DDRF workshop	5	October 5, 2017	public	



Choose creation method

How would you like to create your workset?

<p>Upload File</p>	<p>Import From HathiTrust</p>
<p>Create a workset from a file of HathiTrust volume IDs</p>	<p>Create a workset from an existing, public HathiTrust collection</p>



Input workset information

Create A Workset

Import a collection from HathiTrust using the collection's URL. While HathiTrust grows daily, HTRC syncs data periodically from the HathiTrust Digital Library. Some volumes you would like to include in your workset may not be available. Any volumes in your workset not available through HTRC will be skipped by the algorithm.

Find collection URL

When viewing your [collection on HathiTrust](#), simply copy the URL from your browser, or copy the "Link to this collection" found on the left sidebar, and paste the URL below.

Import

Hit "Fetch Collection" and your collection will be transformed into an HTRC workset. You may need to edit the default name in order to meet HTRC requirements.

HathiTrust Collection URL

Name

Disallowed characters: ~ ! @ # ; % ^ * + = [] | < > , ' " \ /

Description

Private Workset
If checked, your workset will be accessible to only you.

Add collection URL here



View created workset

Worksets

Filter worksets by name... My Worksets

Name	Description	Volume Count	Last Modified Date	Availability	Actions
1970s_political_speeches_DDRF	1970s political speeches collection for DDRF workshop	16	October 5, 2017	public	
Darwin_Test_2		3	October 6, 2016	private	
1930s_political_speeches_DDRF	1930s political speeches collection for DDRF workshop	5	October 5, 2017	public	
poli_science_DDRF	Political science collection for DDRF workshop	16	October 5, 2017	public	
Charles_Darwin_Test	testing purposes	32	August 24, 2016	private	

First « 1 » Last

Showing 1 to 10 of 5 entries



Workset review

- *How did it go?*
- *What kind of search criteria did you use?*
- *Did you find any challenges?*



Case study: *Inside the Creativity Boom*

Building a creativity corpus

- Searched across full text of HTDL for creativ*
- Made initial list of over million volumes
- Deduplicated
 - Kept different editions of same work; discard multiple copies of same edition
- Ended up with refined list (workset) of volumes



Discussion

- *What expertise do librarians already have to help with building a corpus for textual analysis?*



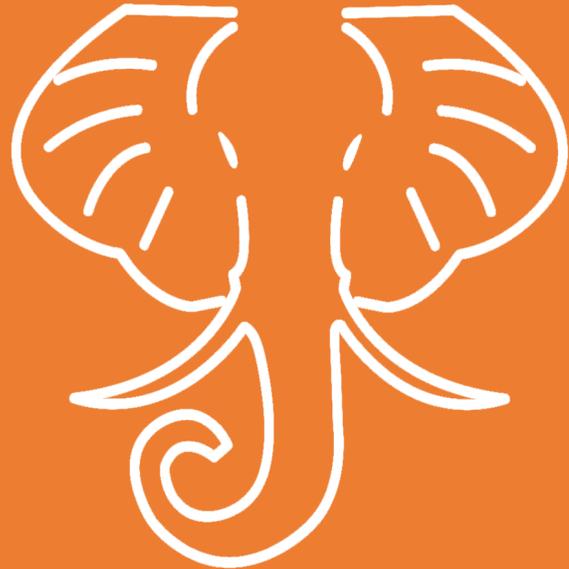
Questions?



References

- Padilla, T. (2015). Kludging: Web to TXT. Retrieved August 16, 2017, from <http://www.thomaspadilla.org/2015/08/03/kludge/> .





**Digging Deeper,
Reaching Further**

Module 2.2: Gathering Textual Data

Bulk retrieval

In this module we'll...

- Introduce options for automated data retrieval
 - *Facilitate researcher needs for data access*
- Practice web scraping with `wget` on the command line
 - *Gain confidence automating data collection*
- See how Sam gathered data for his Creativity Corpus
 - *Learn how one scholar used automated data collection*



Bulk retrieval

- Most researchers will need more than 1 or 10 texts
 - Hundreds, thousands, or millions of texts
- Getting lots of data could take lots of time!
 - Point-and-click is inefficient
 - Automate when possible



Automating retrieval

Transferring files

- FTP or SFTP: (Secure) File Transfer Protocol
 - moves files from one place to another on the internet
- rsync
 - Efficient: sends only the differences
 - Run from command line
 - Used by HathiTrust, can be used to download Extracted Features data



Automating retrieval

Web scraping (grabbing text on the web)

- Avoids tedious copying-and-pasting
- Some ways to scrape text from the web:
 - Run commands such as wget or curl in the command line
 - Write and run a script (a file of programming statements)
 - Use software such as webscraper.io or Kimono

We'll do an activity with this one later!



Web scraping for the wise

- Web scraping puts a large workload on targeted server
 - This can upset the data holder
- Some data providers are more than willing to share
 - Ask for access
 - Check for an API
- Otherwise, time your requests to add a delay between server hits
 - It's polite
 - Also signifies you are not a malicious attacker



Automating retrieval

APIs (Application Programming Interfaces)

- Digital pathways to or from content
 - Sometimes need a “key” for access
- Can be used to gain programmatic access
 - Usually need to write code to retrieve content
 - Sometimes have graphical user interface (GUI)
- Examples: A number of digital content providers have APIs
 - Twitter API: display tweets on a non-Twitter website
 - Chronicling America API: <https://chroniclingamerica.loc.gov/about/api/>



HathiTrust Bibliographic API

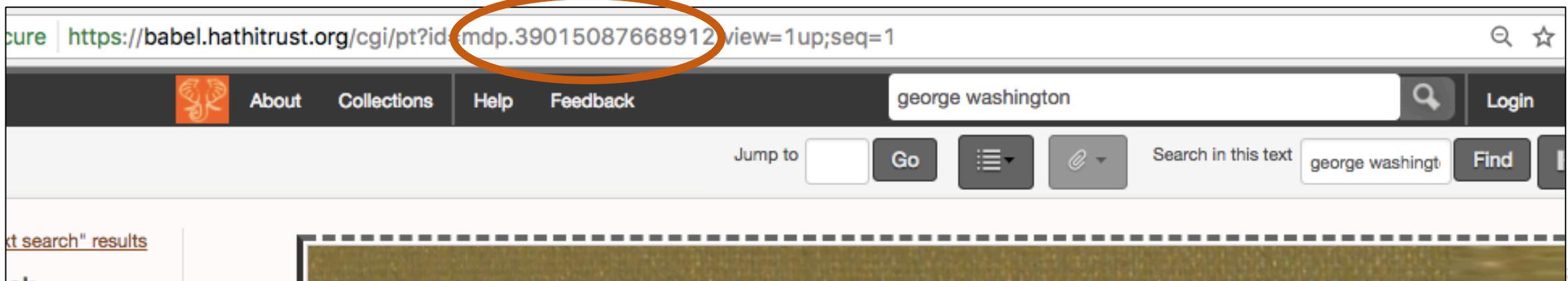
- Provides programmatic access to bibliographic metadata for volumes in the HathiTrust
- Metadata retrieved using a specially formatted URL and the volume ID number
- More information: https://www.hathitrust.org/bib_api



Hands-on activity

 See Handout p. 8

- Use the HathiTrust Bibliographic API to retrieve the metadata for a volume of your choice
 - Search HTDL for a volume (hathitrust.org)
 - Click on the “Full view” or “Limited (search-only)” option of the volume
 - Find the volume ID in the URL (all characters between “id=” and “;”, including all numbers and letters)



Anatomy of an API call

- Types of volume ID: htid, OCLC, record number, lccn, isbn, issn
- A structured URL

https://catalog.hathitrust.org/api/volumes

/brief **OR** /full

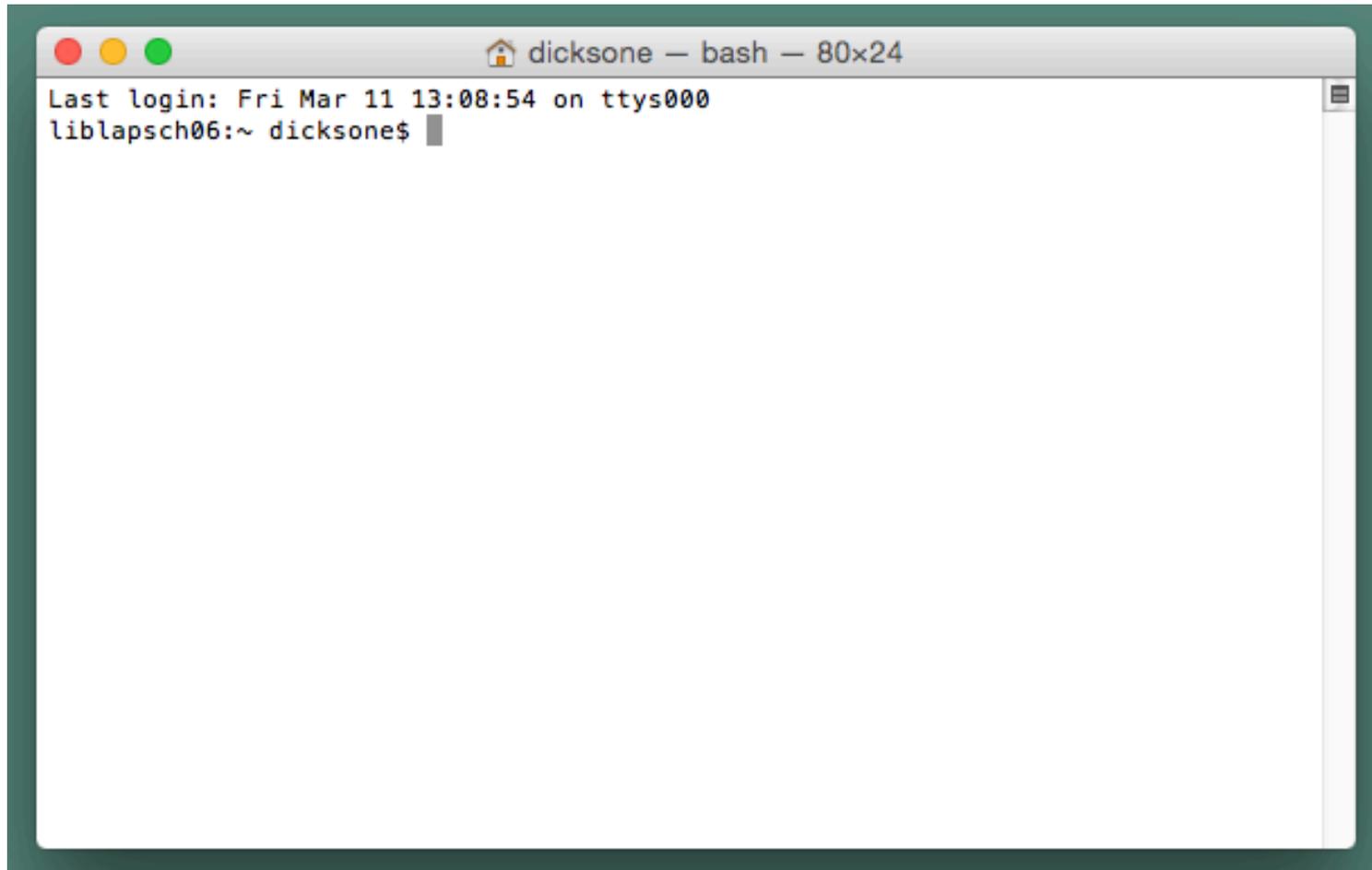
/<ID type>

/<ID.json>

https://catalog.hathitrust.org/api/volumes/full/htid/mdp.39015005337046.json



Intro to the command line



```
dicksonsone — bash — 80x24
Last login: Fri Mar 11 13:08:54 on ttys000
liblapsch06:~ dicksonsone$
```



Background

- Command line interfaces
 - On Mac: Terminal
 - On Windows: Command Prompt
 - They do not speak the same language!
- Bash shell = Unix shell
 - Default on Mac
 - To use on Windows, install GitBash or Cygwin
 - *Bash on Ubuntu on Windows* comes installed on Windows 10

We're going to use the built-in Bash shell environment in PythonAnywhere



What's PythonAnywhere?

- PythonAnywhere is a browser-based programming environment
 - Write code
 - Save files
- Comes with a built-in Bash console
 - Run commands
- Does not interact with your local file system



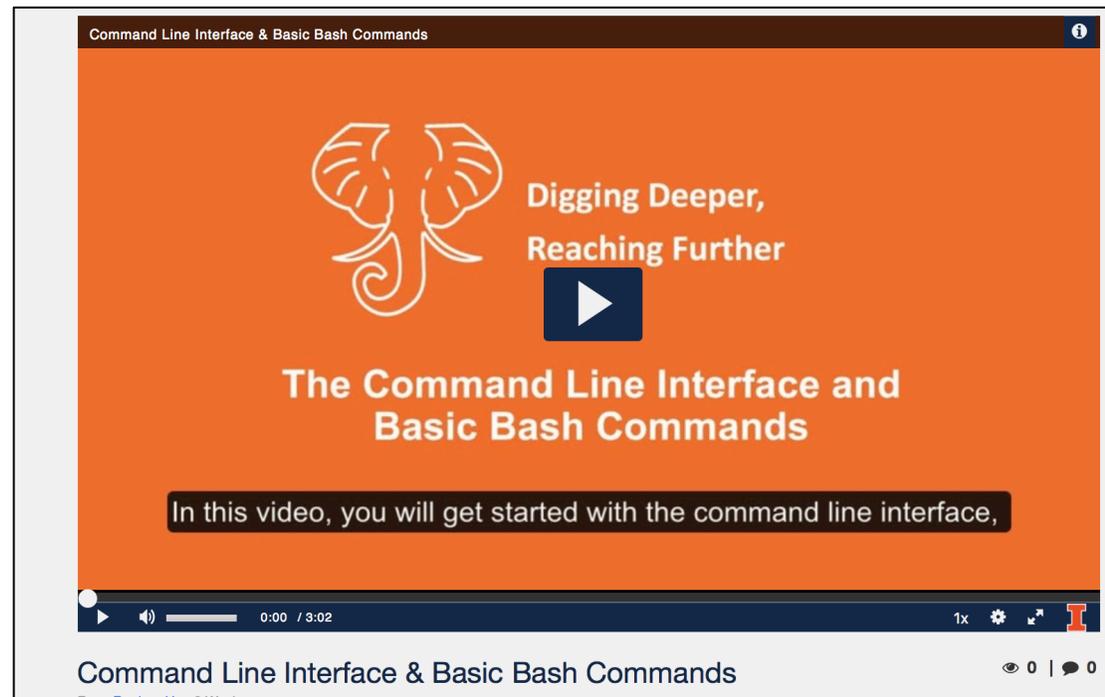
Tips for working in a shell

- Directory = folder
- Case, spaces, and punctuation matter
- Tab to autocomplete a line
- Hit up/down arrow to see last commands entered
- Use “q” to quit viewing a file



Basic Bash commands

- Introductory video:



<http://go.illinois.edu/ddrf-bash-video>



Hands-on activity

 *See Handout p. 9*

Open a Bash console in PythonAnywhere, and practice the basic commands to:

- See your current directory

```
pwd
```

- Are there any files in your directory? List them:

```
ls
```



Hands-on activity

 See Handout p. 9

In your Bash console, practice the following:

- Unzip activity files (activity_files.zip) in your user directory:

```
unzip activity_files.zip
```

- Files were unzipped to a new directory called activity_files.

- Move the files from activity_files to your user directory. Substituting your Python Anywhere user name for *[your_username]*:

```
mv activity_files/* /home/[your_username]/
```



Hands-on activity

 See Handout p. 9

Use the ls command to check if files were successfully unzipped and moved:

- List files and directories under your current directory: `ls`
- List should include:

```
1930 README.txt activity_files.zip remove_tag.py  
top_adjectives.py 1970 activity_files  
remove_stopwords.py stopwords.txt word_count.py
```



Hands-on activity

 *See Handout p. 9*

More practice:

- Make a directory called 'test'

```
mkdir test
```

- Change into the directory you just made, then back out of it

```
cd test
```

```
cd ..
```

Note: Please make sure you are back in your main directory after finishing the activity!



Command line tool: wget

- wget is a powerful way to transfer files from a server
 - i.e. web scraping
 - Follows links on the page and grabs that content, too
- You can specify **options**
 - Options tailor the command to do what you want
 - They are often written as single “flagged” letters, such as `-d` or `-r`
- Options for wget include:
 - `-l`: how far in the hierarchy you want to go
 - `--limit-rate=___`: sets the file transfer speed



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:

Download text from an open resource, WikiSource, to start building a corpus.



Hands-on activity

 *See Handout p. 10*

In this activity, you will run a command to scrape text from a webpage version of George Washington's Fourth State of the Union Address.

Website Source: https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address



Web scraping

- What You Need:

- PythonAnywhere: www.pythonanywhere.com

- Website URL:

- https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address

- Command:

```
wget -l 1 --limit-rate=20k
```

```
https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
```



Web Source

Page [Discussion](#) [Read](#) [Edit](#) [View history](#) [More](#)

George Washington's Fourth State of the Union Address

[← George Washington's Third State of the Union Address](#) **Fourth State of the Union Address** (1792) [George Washington's Fifth State of the Union Address →](#)
by George Washington

Delivered in Washington, D.C. on 8 November 1792.

 [related portals: State of the Union Speeches by United States Presidents.](#)
 [sister projects: data item.](#)

Fellow-Citizens of the Senate and of the House of Representatives:

It is some abatement of the satisfaction with which I meet you on the present occasion that, in felicitating you on a continuance of the national prosperity generally, I am not able to add to it information that the Indian hostilities which have for some time past distressed our Northwestern frontier have terminated.

You will, I am persuaded, learn with no less concern than I communicate it that reiterated endeavors toward effecting a pacification have hitherto issued only in new and outrageous proofs of persevering hostility on the part of the tribes with whom we are in contest. An earnest desire to procure tranquillity to the frontier, to stop the further effusion of blood, to arrest the progress of expense, to forward the prevalent wish of the nation for peace has led to strenuous efforts through various channels to accomplish these desirable purposes; in making which efforts I consulted less my own anticipations of the event, or the scruples which some considerations were calculated to inspire, than the wish to find the object attainable, or if not attainable, to ascertain unequivocally that such is the case.

A detail of the measures which have been pursued and of their consequences, which will be laid before you, while it will confirm to you the want of success thus far, will, I trust, evince that means as proper and as efficacious as could have been devised have been employed. The issue of some of them, indeed, is still depending, but a favorable one, though not to be despaired of, is not promised by anything that has yet happened.

In the course of the attempts which have been made some valuable citizens have fallen victims to their zeal for the public service. A sanction commonly respected even among savages has been found in this instance insufficient to protect from massacre the emissaries of peace. It will, I presume, be duly considered whether the occasion does not call for an exercise of liberality toward the families of the deceased.

It must add to your concern to be informed that, besides the continuation of hostile appearances among the tribes north of the Ohio, some threatening symptoms have of late been revived among some of those south of it.

https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address



Open shell and enter command

```
19:00 ~ $ wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
```

Enter:

```
wget -l 1 --limit-rate=20k  
https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
```

Tip: Copy and paste the URL!



Run command successfully

```
19:00 ~ $ wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
--2017-07-06 19:00:46-- https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address
Resolving proxy.server (proxy.server)... 10.0.0.82
Connecting to proxy.server (proxy.server)|10.0.0.82|:3128... connected.
Proxy request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: 'washington_4.txt'

[          <=>          ] 42,088      20.0KB/s   in 2.1s

2017-07-06 19:00:48 (20.0 KB/s) - 'washington_4.txt' saved [42088]

19:00 ~ $ █
```



View new file

The screenshot shows the PythonAnywhere dashboard for user 'rhan11'. The current directory is '/home/'. The dashboard is divided into 'Directories' and 'Files' sections. The 'Files' section contains a table of files with their names, sizes, and timestamps. An orange arrow points to the file 'word_count.py'.

pythonanywhere [Dashboard](#) [Consoles](#) **Files** [Web](#) [Tasks](#)

/home/ rhan11 [Open Bash console here](#) **7% full** – 36.3 MB of your 512 MB

Directories

Enter new directory name: [New directory](#)

- [.cache/](#)
- [.config/](#)
- [.local/](#)
- [.virtualenvs/](#)
- [1930/](#)
- [1970/](#)
- [__MACOSX/](#)
- [activity_files/](#)

Files

Enter new file name, eg hello.py

	.bash_history				2017-08-17 12:34	297 bytes
	.bashrc				2016-08-24 22:18	559 bytes
	.gitconfig				2016-08-24 22:18	266 bytes
	.profile				2016-08-24 22:18	79 bytes
	.pythonstartup.py				2016-08-24 22:18	77 bytes
	.vimrc				2016-08-24 22:18	4.6 KB
	README.txt				2017-09-19 05:28	235 bytes
	activity_files.zip				2018-01-08 16:51	15.0 MB
	remove_stopwords.py				2017-10-09 18:20	743 bytes
	remove_tag.py				2017-10-09 18:19	800 bytes
	stopwords.txt				2017-10-09 18:20	621 bytes
	tagless_file.txt				2018-01-10 21:35	13.2 KB
	top_adjectives.py				2017-10-09 18:20	1.0 KB
	washington_4.txt				2017-10-13 15:30	40.8 KB
	word_count.py				2017-10-09 18:20	516 bytes

[Upload a file](#)



View new file

```
1 <!DOCTYPE html>
2 <html class="client-nojs" lang="en" dir="ltr">
3 <head>
4 <meta charset="UTF-8"/>
5 <title>George Washington's Fourth State of the Union Address - Wikisource, the free online library</title>
6 <script>document.documentElement.className = document.documentElement.className.replace( /(^\s)client-nojs(\s|$)/, "$1client-js
7 <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,
8 mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});/*@nomin*/
9
10 });mw.loader.load(["site","mediawiki.page.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","mediawiki.searchSu
11 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=ext.uls.interlanguage%7Cext.visualEditor.desktopArt
12 <script async="" src="/w/load.php?debug=false&lang=en&modules=startup&only=scripts&skin=vector"></script>
13 <meta name="ResourceLoaderDynamicStyles" content=""/>
14 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=ext.gadget.Site&only=styles&skin=vector"/>
15 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=site.styles&only=styles&skin=vector"/>
16 <meta name="generator" content="MediaWiki 1.31.0-wmf.15"/>
17 <meta name="referrer" content="origin-when-cross-origin"/>
18 <link rel="alternate" type="application/x-wiki" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_
19 <link rel="edit" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&action=edit"/>
20 <link rel="shortcut icon" href="/static/favicon/wikisource.ico"/>
21 <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikisource (en)"/>
22 <link rel="EditURI" type="application/rsd+xml" href="//en.wikisource.org/w/api.php?action=rsd"/>
23 <link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/">
24 <link rel="alternate" type="application/atom+xml" title="Wikisource Atom feed" href="/w/index.php?title=Special:RecentChanges&am
25 <link rel="canonical" href="https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address"/>
26 <link rel="dns-prefetch" href="//login.wikimedia.org"/>
27 <link rel="dns-prefetch" href="//meta.wikimedia.org" />
28 <!--[if lt IE 9]><script src="/w/load.php?debug=false&lang=en&modules=html5shiv&only=scripts&skin=vector&syn
29 </head>
30 <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-George_Washington_s_Fourth_State_of_the_Union_Addr
31 <div id="mw-head-base" class="noprint"></div>
32 <div id="content" class="mw-body" role="main">
33 <a id="top"></a>
```



What happened?

- You executed a command to scrape the text from the URL
- It grabbed everything on the page
- What's next?
 - Prepare the data for analysis

What might need to be removed from the files before analysis?



On your own

👉 See Handout p. 11

- *Edit the command to scrape George Washington's second State of the Union speech and save to a new text file.*
- *Can you view your scraped files using the `less` command?*



Other options to explore

- *Beautiful Soup*

- <https://www.crummy.com/software/BeautifulSoup/>
- Python-based web scraping
- Lots of options, very powerful

- Write web scraping into a **script**

- A script is a list of directions for your computer to follow
- Direct computer to grab many files to retrieve content in bulk
- Remember to time your requests!



Bulk HathiTrust data access

HT and HTRC datasets

Dataset	Kind of data	Description	Volumes available
HT Custom data request	Full text	Download page images and plain text OCR	Public domain
HTRC Extracted Features	Abstracted text and metadata	JSON files for each of 15.7 million volumes in HathiTrust	All
HTRC Data API	Full text	Plain text OCR	All for HT members; public domain for others



Case Study: *Inside the Creativity Boom*

After creating final list of volumes:

- Used rsync to retrieve HTRC Extracted Features for the volumes
- Remember rsync is a command line utility, like wget, that transfers files between computers



Case Study: *Inside the Creativity Boom*

- What exactly is the HTRC Extracted Features dataset?

JSON files for each volume in the HathiTrust that researchers can download

```
1  {
2    "id": "uc1.b3419888",
3    "metadata": {
4      "schemaVersion": "1.2",
5      "dateCreated": "2015-02-12T13:30",
6      "title": "Zoonomia = or The laws of organic life / by Erasmus Darwin.",
7      "pubDate": "1809",
8      "language": "eng",
9      "htBibUrl": "http://catalog.hathitrust.org/api/volumes/full/htid/uc1.b3419888.json",
10     "handleUrl": "http://hdl.handle.net/2027/uc1.b3419888",
11     "oclc": "3679915",
12     "imprint": "Thomas and Andrews, 1809."
13   },
14   "features": {
15     "schemaVersion": "2.0",
16     "dateCreated": "2015-02-20T23:58",
17     "pageCount": 616,
18     "pages": [
```

Metadata: bibliographic;
inferred
Data: words and word counts

Read and reflect

 *See Handout p. 11*

- Santa Barbara Statement on Collections as Data (Collections as Data National Forum, 2017)
<https://collectionsasdata.github.io/statement/>
- Provides a set of high level principles to guide collections as data work



Read and reflect

☞ *See Handout p. 11*

- “With a few exceptions, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that.”
- “Any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs.”
- “Ethical concerns are integral to collections as data.”
- **Principle 2 for collections as data:** “Collections as data development aims to encourage computational use of digitized and born digital collections.”



Read and reflect

 *See Handout p. 11*

- *Does your library provide access to digital collections as data?*
- *How so? Why not? How could it?*



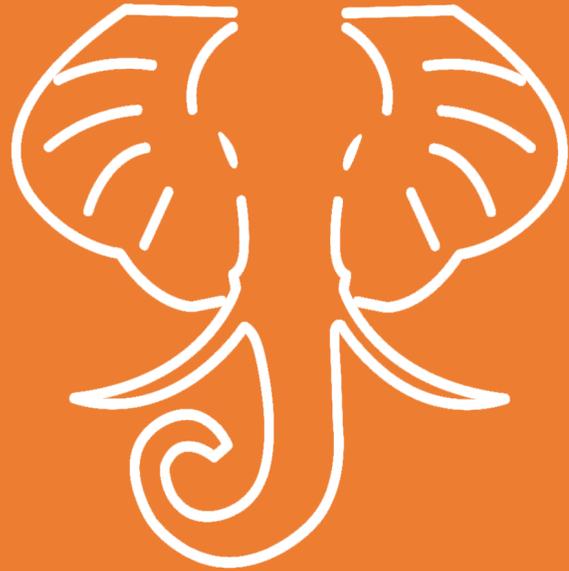
Questions?



References

- Collections as Data National Forum. (2017, March 3). The Santa Barbara Statement on Collections as Data. Retrieved August 16, 2017, from <https://collectionsasdata.github.io/statement/> .





**Digging Deeper,
Reaching Further**

Module 3: Working with Textual Data

In this module we'll...

- Think about what happens when text is data
 - *Understand best practice in the field*
- Consider common steps to cleaning and preparing text data
 - *Make recommendations to researchers*
- Practice data cleaning on the speeches we scraped
 - *Gain experience with Python for working with data*
- Learn how Sam prepared his *Creativity Corpus* for analysis
 - *See how one scholar data prepared data*



Where we'll end up

```
1 <!DOCTYPE html>
2 <html class="client-nojs" lang="en" dir="ltr">
3 <head>
4 <meta charset="UTF-8"/>
5 <title>George Washington's Fourth State of the Union Address - Wikisource, the free online library</title>
6 <script>document.documentElement.className = document.documentElement.className.replace( /(\s)client-nojs/, "$1client-js$2");</script>
7 <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespacesNumber":0,"wgPageName":"George_Washington's_Fourth_Sta
8 mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});/*@nomin*/;
9
10 });mw.loader.load(["site","mediawiki.page.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","mediawiki.searchSuggest","ext.gadget.charinsert","ext.gadget.Easy_LST","ext.gadget.d
11 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=ext.uls.interlanguage&7Cext.visualEditor.desktopArticleTarget.noscript&7Cext.wikimediaBadges&7Cmediawiki.legacy.commo
12 <script async="" src="/w/load.php?debug=false&lang=en&modules=startup&only=scripts&skin=vector"></script>
13 <meta name="ResourceLoaderDynamicStyles" content="" />
14 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=ext.gadget.Site&only=styles&skin=vector"/>
15 <link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=site.styles&only=styles&skin=vector"/>
16 <meta name="generator" content="MediaWiki 1.30.0-wmf.7"/>
17 <meta name="referrer" content="origin-when-cross-origin"/>
18 <link rel="alternate" type="application/x-wiki" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&action=edit"/>
19 <link rel="edit" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&action=edit"/>
20 <link rel="shortcut icon" href="/static/favicon/wikisource.ico"/>
21 <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="
22 <link rel="EditURI" type="application/rsd+xml" href="//en.wikisource.org/w/api.php?action=rsd"/>
23 <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
24 <link rel="alternate" type="application/atom+xml" title="Wikisource Atom feed" href="/w/index.php?tit
25 <link rel="canonical" href="https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the
26 <link rel="dns-prefetch" href="//login.wikimedia.org/" />
27 <link rel="dns-prefetch" href="//meta.wikimedia.org/" />
28 </head>
29 <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-George_Washington_s_F
30 <div id="mw-head-base" class="noprint"></div>
31 <div id="content" class="mw-body" role="main">
32 <a id="top"></a>
33
34 <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div>
35 <div class="mw-indicators mw-body-content">
36 </div>
37 <h1 id="firstHeading" class="firstHeading" lang="en">George Washington's Fourth State of the
38 <div id="bodyContent" class="mw-body-content">
39 <div id="siteSub">From Wikisource</div>
40 <div id="contentSub"></div>
41 <div id="jump-to-nav" class="mw-jump">
```



Go from HTML-formatted text...



...to cleaner text ready for analysis.

```
1 [' Fellow-Citizens of the Senate and of the House of Representatives: ',
2 'It is some abatement of the satisfaction with which I meet you on the present occasion that, in felicitating you on a
3 continuance of the national prosperity generally, I am not able to add to it information that the Indian hostilities which
4 have for some time past distressed our Northwestern frontier have terminated.', 'You will, I am persuaded, learn with no less
5 concern than I communicate it that reiterated endeavors toward effecting a pacification have hitherto issued only in new and
6 outrageous proofs of persevering hostility on the part of the tribes with whom we are in contest. An earnest desire to procure
7 tranquillity to the frontier, to stop the further effusion of blood, to arrest the progress of expense, to forward the prevalent
8 wish of the nation for peace has led to strenuous efforts through various channels to accomplish these desirable purposes; in
9 making which efforts I consulted less my own anticipations of the event, or the scruples which some considerations were
10 calculated to inspire, than the wish to find the object attainable, or if not attainable, to ascertain unequivocally that
11 such is the case.', 'A detail of the measures which have been pursued and of their consequences, which will be laid before
12 you, while it will confirm to you the want of success thus far, will, I trust, evince that means as proper and as efficacious
13 as could have been devised have been employed. The issue of some of them, indeed, is still depending, but a favorable one,
14 though not to be despaired of, is not promised by anything that has yet happened.', 'In the course of the attempts which have
15 been made some valuable citizens have fallen victims to their zeal for the public service. A sanction commonly respected even
16 among savages has been found in this instance insufficient to protect from massacre the emissaries of peace. It will, I presume,
17 be duly considered whether the occasion does not call for an exercise of liberality toward the families of the deceased.',
18 'It must add to your concern to be informed that, besides the continuation of hostile appearances among the tribes north of
19 the Ohio, some threatening symptoms have of late been revived among some of those south of it.', 'A part of the Cherokees,
20 known by the name of Chickamaugas, inhabiting five villages on the Tennessee River, have long been in the practice of
21 committing depredations on the neighboring settlements.', 'It was hoped that the treaty of Holston, made with the Cherokee
22 Nation in July, 1791, would have prevented a repetition of such depredations; but the event has not answered this hope. The
23 Chickamaugas, aided by some banditti of another tribe in their vicinity, have recently perpetrated wanton and unprovoked
24 hostilities upon the citizens of the United States in that quarter. The information which has been received on this subject
25 will be laid before you. Hitherto defensive precautions only have been strictly enjoined and observed.', 'It is not understood
26 that any breach of treaty or aggression whatsoever on the part of the United States or their citizens is even alleged as a
27 pretext for the spirit of hostility in this quarter.', 'I have reason to believe that every practicable exertion has been made
28 (pursuant to the provision by law for that purpose) to be prepared for the alternative of a prosecution of the war in the event
29 of a failure of pacific overtures. A large proportion of the troops authorized to be raised have been recruited, though the
30 number is still incomplete, and pains have been taken to discipline and put them in condition for the particular kind of
```



Humanities data

- Data is material generated or collected while conducting research
- Examples of humanities data:
 - Citations
 - Code/Algorithms
 - Databases
 - Geospatial coordinates

Can you think of others?



Text as data

- Data quality
 - Clean vs. dirty OCR
 - HathiTrust OCR is dirty (uncorrected)
- Analyzed by corpus/corpora
 - Text corpus: a digital collection OR an individual's research text dataset
 - Text corpora: “bodies” of text
- Text decomposition/recomposition (Rockwell, 2003)
 - Cleaning data involves discarding data
 - Prepared text may be illegible to the human reader



Preparing data

A researcher may:

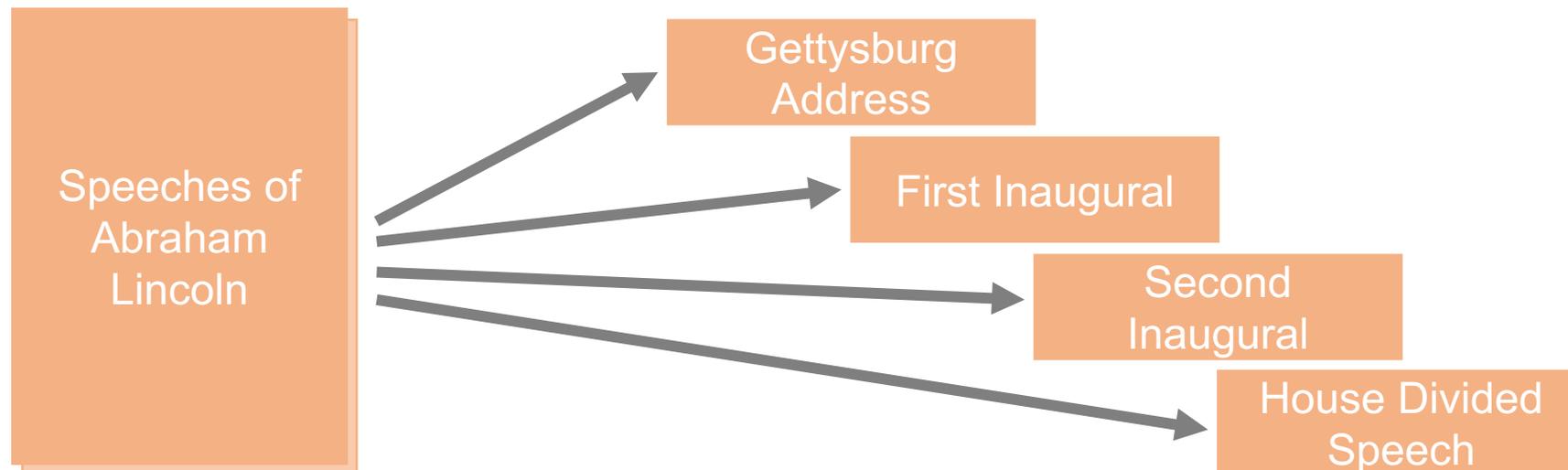
- Correct OCR errors
- Remove title, header information
- Remove html or xml tags
- Split or combine files
- Remove certain words, punctuation marks
- Lowercase text
- Tokenize the words



Key concepts

Chunking text

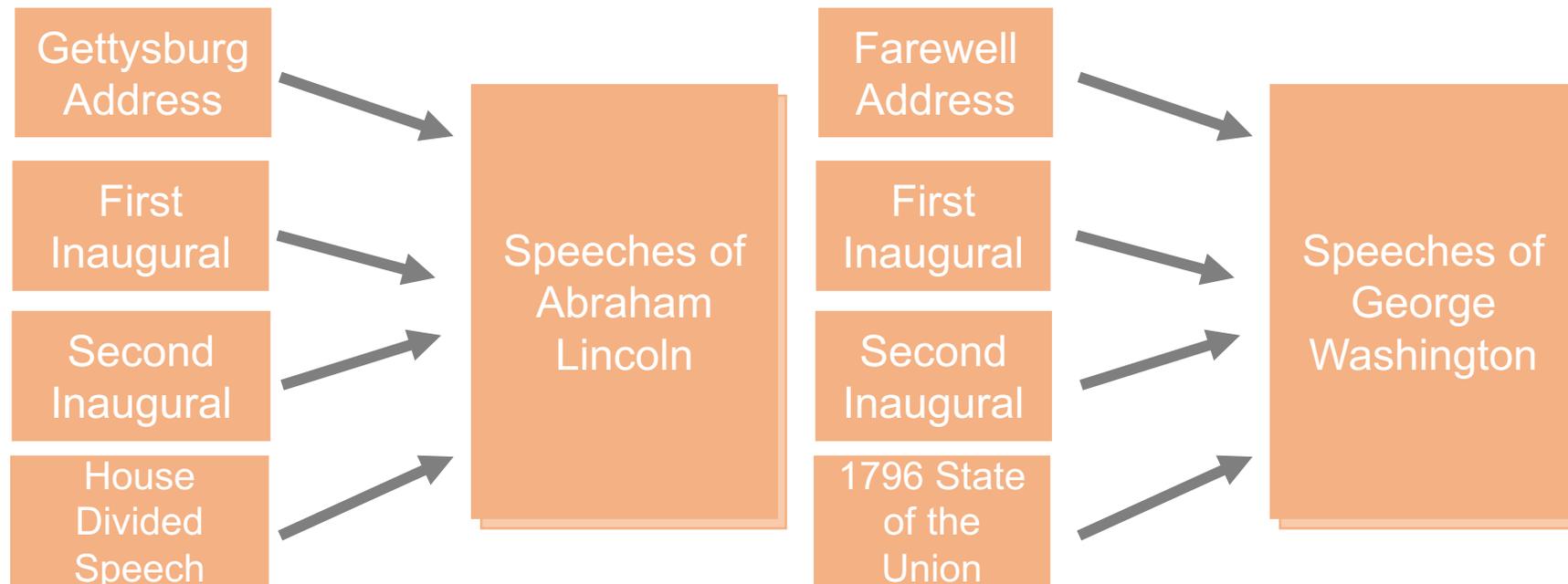
Splitting text into smaller pieces before analysis. May be divided by paragraph, chapter, or a chosen number of words (e.g. 1000 word chunks).



Key concepts

Grouping text

Combining text into larger pieces before analysis.



Key concepts

Tokenization

Breaking text into pieces called tokens. Often certain characters, such as punctuation marks, are discarded in the process

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



Preparing data

- Preparation affects results
 - Amount of text and size of chunks
 - Which stop words removed; which characters are included
 - Whether to lowercase and normalize words
- Preparation for analysis takes time, effort
 - This is where scripting becomes useful!



Hands-on Activity

 See Handout pp. 12-13

- In groups of 2 or 3, assign each person several of the text preparation actions seen in the table to the right (Denny and Spirling, 2017).
- Read the descriptions. Then take turns explaining each to your group.

Term
Punctuation
Numbers
Lowercasing
Stemming
Stopword Removal
n-gram Inclusion
Infrequently Used Terms



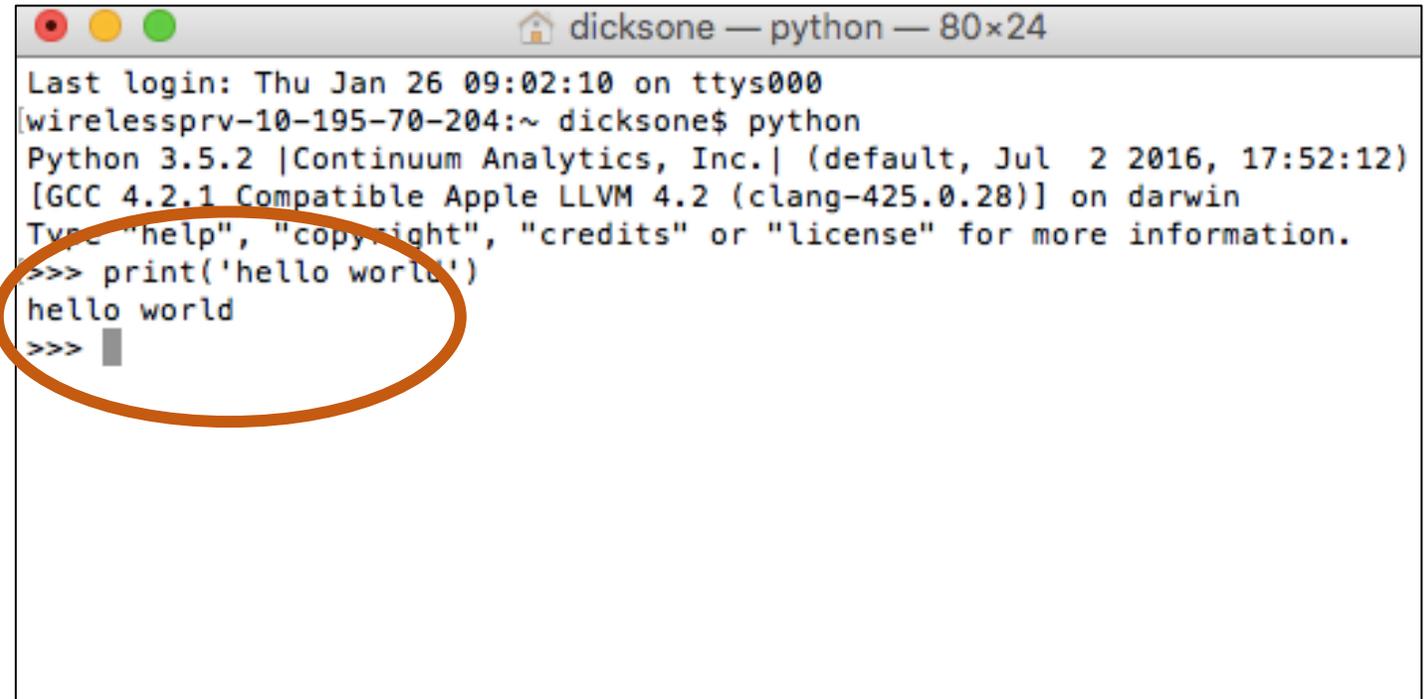
Scripting with Python

- Python is a scripting language
- Good for working with data
 - Interpreted language → follows step-by-step directions
- Relatively straightforward syntax
 - Avoids excess punctuation



Using Python: Interactive programming

- Using a **python interpreter**
- Run each step at the prompt
- If you enter “Python” on the command line, you will start the interpreter
- *We aren't using it today!*



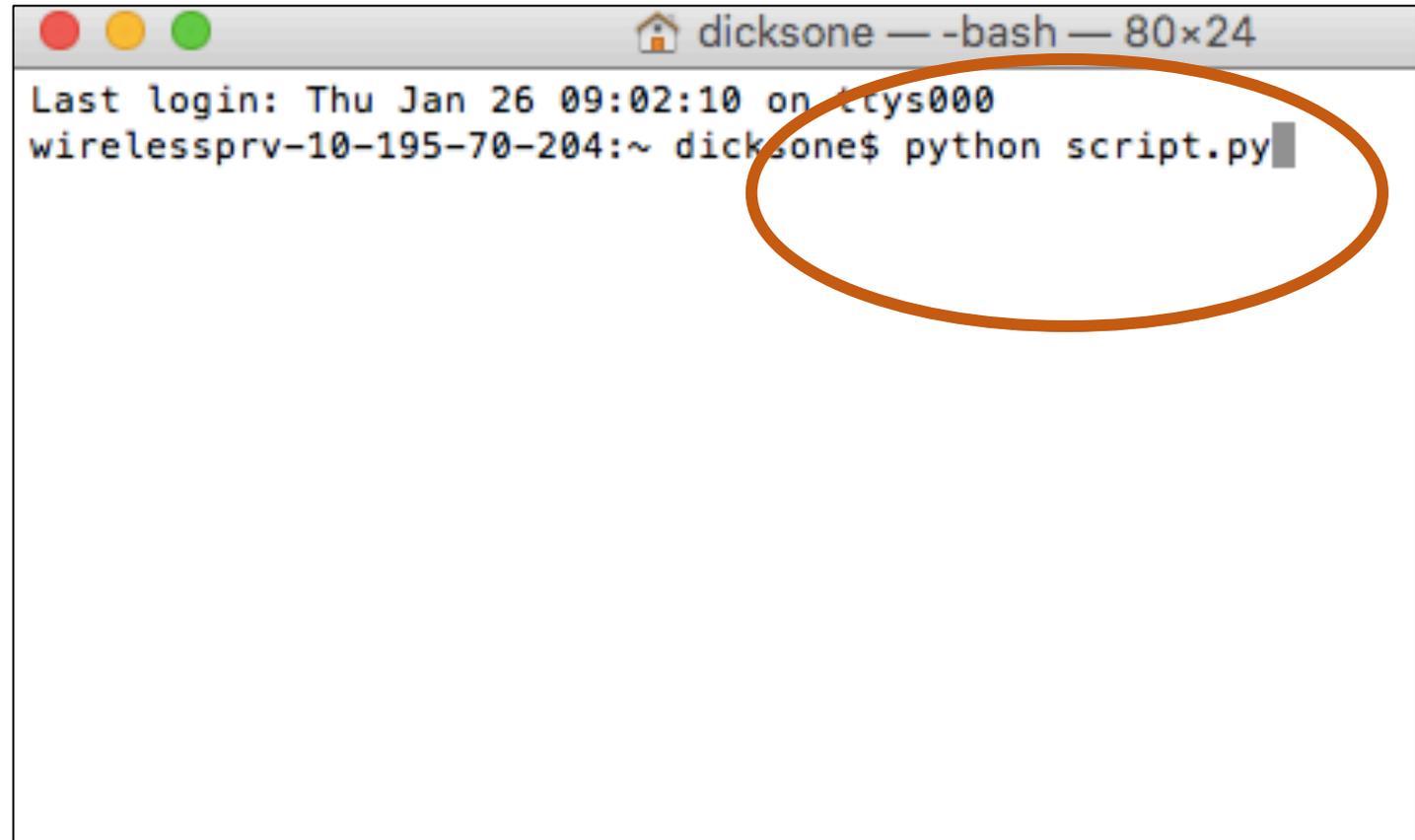
```
dicksone — python — 80x24
Last login: Thu Jan 26 09:02:10 on ttys000
[wirelessprv-10-195-70-204:~ dicksone$ python
Python 3.5.2 |Continuum Analytics, Inc.| (default, Jul 2 2016, 17:52:12)
[GCC 4.2.1 Compatible Apple LLVM 4.2 (clang-425.0.28)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> print('hello world')
hello world
>>> █
```



Using Python: Write & run scripts

- Scripts are directions for your computer to follow
- Save the script as a file ending in .py
- On the command line, run the script

→ This is how we'll do it!

A terminal window titled 'dicksone — -bash — 80x24' with a home icon. The window shows the following text: 'Last login: Thu Jan 26 09:02:10 on ttys000', 'wirelessprv-10-195-70-204:~ dicksone\$', and 'python script.py' with a cursor. The command 'python script.py' is circled in orange.

```
dicksone — -bash — 80x24
Last login: Thu Jan 26 09:02:10 on ttys000
wirelessprv-10-195-70-204:~ dicksone$ python script.py
```



Anatomy of running a Python script

```
python <script filename> <arguments>
```

1. Tells the computer to be ready for Python
2. Directs which file to run
3. The file may ask for additional information to be given at runtime, called arguments

Example: `python myscript.py newfile.txt`



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:

Use a Python script to prepare the text data we scraped from Wikisource.



Hands-on activity

 See Handout p. 14

In this activity, you will run a Python script that will remove the HTML tags from George Washington's *Fourth State of the Union Address* you scraped earlier from Wikisource.



Strip HTML tags

- What You Need:
 - PythonAnywhere bash console
 - washington_4.txt
 - remove_tag.py script



Go to Files to view script

Files

Enter new file name, eg hello.py New file

 .bash_history	  	2017-03-16 06:40	802 bytes
 .bashrc	  	2016-03-28 15:42	559 bytes
 .gitconfig	  	2016-03-28 15:42	266 bytes
 .my.cnf	  	2017-04-13 14:12	40 bytes
 .mysql_history	  	2017-04-13 14:17	0 bytes
 .profile	  	2016-03-28 15:42	79 bytes
 .python_history	  	2017-02-09 18:20	56 bytes
 .pythonstartup.py	  	2016-07-01 20:53	77 bytes
 .sqlite_history	  	2017-04-14 19:53	2.1 KB
 .vimrc	  	2016-07-01 20:53	4.6 KB
 README.txt	  	2016-10-05 16:32	235 bytes
 get_verbs.py	  	2017-03-21 17:31	903 bytes
 mdp.49015002221837.json.bz2	 	2017-03-08 14:27	988.0 KB
 remove_stopwords.py	  	2017-05-05 14:37	353 bytes
 remove_tag.py	  	2017-07-07 13:17	374 bytes
 see_nouns_new.py	  	2017-07-05 15:58	590 bytes
 stopwords.txt	  	2017-05-05 14:37	621 bytes
 tagless_file.txt	  	2017-07-07 15:10	13.2 KB
 washington_4.txt	  	2017-06-22 19:00	41.1 KB

 Upload a file



View script

```
1 import sys
2 import re
3
4 file_name = sys.argv[1]
5
6 with open(file_name, 'r') as myfile:
7     contents = myfile.read().replace('\n', '')
8     pattern = r"<p>(.*?)</p>"
9     new_text = str(re.findall(pattern, contents))
10    new_text = new_text.replace("<i>", " ")
11    new_text = new_text.replace("</i>", " ")
12    new_text = new_text.replace("<a>", " ")
13    new_text = new_text.replace("</a>", " ")
14    new_text = new_text.replace("<b>", " ")
15    new_text = new_text.replace("</b>", " ")
16    cleaned = open('tagless_file.txt', 'w')
17    cleaned.write(new_text)
18
19
```



Open shell and run script

```
15:09 ~ $ python remove_tag.py washington_4.txt
```

Enter command:

```
python remove_tag.py washington_4.txt
```



Find new file

 .bash_history	   2017-03-16 06:40 802 bytes
 .bashrc	   2016-03-28 15:42 559 bytes
 .gitconfig	   2016-03-28 15:42 266 bytes
 .my.cnf	   2017-04-13 14:12 40 bytes
 .mysql_history	   2017-04-13 14:17 0 bytes
 .profile	   2016-03-28 15:42 79 bytes
 .python_history	   2017-02-09 18:20 56 bytes
 .pythonstartup.py	   2016-07-01 20:53 77 bytes
 .sqlite_history	   2017-04-14 19:53 2.1 KB
 .vimrc	   2016-07-01 20:53 4.6 KB
 README.txt	   2016-10-05 16:32 235 bytes
 get_verbs.py	   2017-03-21 17:31 903 bytes
 mdp.49015002221837.json.bz2	  2017-03-08 14:27 988.0 KB
 remove_stopwords.py	   2017-05-05 14:37 353 bytes
 remove_tag.py	   2017-07-07 13:17 374 bytes
 see_nouns_new.py	   2017-07-05 15:58 590 bytes
 stopwords.txt	   2017-05-05 14:37 621 bytes
 tagless_file.txt	   2017-07-07 15:10 13.2 KB
 washington_4.txt	   2017-06-22 19:00 41.1 KB



Review results

1 [' Fellow-Citizens of the Senate and of the House of Representatives: ',
2 'It is some abatement of the satisfaction with which I meet you on the present occasion that, in felicitating you on a
3 continuance of the national prosperity generally, I am not able to add to it information that the Indian hostilities which
4 have for some time past distressed our Northwestern frontier have terminated.', 'You will, I am persuaded, learn with no less
5 concern than I communicate it that reiterated endeavors toward effecting a pacification have hitherto issued only in new and
6 outrageous proofs of persevering hostility on the part of the tribes with whom we are in contest. An earnest desire to procure
7 tranquillity to the frontier, to stop the further effusion of blood, to arrest the progress of expense, to forward the prevalent
8 wish of the nation for peace has led to strenuous efforts through various channels to accomplish these desirable purposes; in
9 making which efforts I consulted less my own anticipations of the event, or the scruples which some considerations were
10 calculated to inspire, than the wish to find the object attainable, or if not attainable, to ascertain unequivocally that
11 such is the case.', 'A detail of the measures which have been pursued and of their consequences, which will be laid before
12 you, while it will confirm to you the want of success thus far, will, I trust, evince that means as proper and as efficacious
13 as could have been devised have been employed. The issue of some of them, indeed, is still depending, but a favorable one,
14 though not to be despaired of, is not promised by anything that has yet happened.', 'In the course of the attempts which have
15 been made some valuable citizens have fallen victims to their zeal for the public service. A sanction commonly respected even
16 among savages has been found in this instance insufficient to protect from massacre the emissaries of peace. It will, I presume,
17 be duly considered whether the occasion does not call for an exercise of liberality toward the families of the deceased.',
18 'It must add to your concern to be informed that, besides the continuation of hostile appearances among the tribes north of
19 the Ohio, some threatening symptoms have of late been revived among some of those south of it.', 'A part of the Cherokees,
20 known by the name of Chickamaugas, inhabiting five villages on the Tennessee River, have long been in the practice of
21 committing depredations on the neighboring settlements.', 'It was hoped that the treaty of Holston, made with the Cherokee
22 Nation in July, 1791, would have prevented a repetition of such depredations; but the event has not answered this hope. The
23 Chickamaugas, aided by some banditti of another tribe in their vicinity, have recently perpetrated wanton and unprovoked
24 hostilities upon the citizens of the United States in that quarter. The information which has been received on this subject
25 will be laid before you. Hitherto defensive precautions only have been strictly enjoined and observed.', 'It is not understood
26 that any breach of treaty or aggression whatsoever on the part of the United States or their citizens is even alleged as a
27 pretext for the spirit of hostility in this quarter.', 'I have reason to believe that every practicable exertion has been made
28 (pursuant to the provision by law for that purpose) to be prepared for the alternative of a prosecution of the war in the event
29 of a failure of pacific overtures. A large proportion of the troops authorized to be raised have been recruited, though the
30 number is still incomplete, and pains have been taken to discipline and put them in condition for the particular kind of



What happened

- Looks for `< >` to denote html tags
 - Using regular expressions and Python's `replace()`
- Writes the characters that are NOT html tags and their contents to a file
- Note: this script is simple, but not the most robust way to remove tags



On your own

 See Handout p. 14

Can you run a script to remove stop words?

- The script is called `remove_stopwords.py`
- Hint: The `remove_stopwords.py` script requires three arguments:
 - The input filename of your tagless file
 - The filename of the list of stop words
 - The output filename you make up
- Edit the `stopwords.txt` file to customize your list



Case Study: *Inside the Creativity Boom*

After downloading the Extracted Features data for the relevant volumes, used scripting to:

- Narrow corpus to individual pages that contained creativ*
 - Discarded all other pages
- Discard certain tokens such as pronouns and conjunctions
 - To keep only to most "meaningful" terms



Read and Reflect...

👉 *See Handout p. 15*

- Passage from “[Against Cleaning](#)” by Katie Rawson and Trevor Muñoz
- They suggest a strategy for dealing with humanities data:
 - Shared authority control across data sets
 - Indexes for nuance
 - Tidy, not clean data



Read and Reflect...

👉 *See Handout p. 15*

“When humanities scholars recoil at data-driven research, they are often responding to the reductiveness inherent in this form of scholarship. This reductiveness can feel intellectually impoverishing to scholars who have spent their careers working through particular kinds of historical and cultural complexity... From within this worldview, data cleaning is then maligned because it is understood as a step that inscribes a normative order by wiping away what is different. The term “cleaning” implies that a data set is ‘messy.’ “Messy” suggests an underlying order. It supposes things already have a rightful place, but they’re not in it—like socks on the bedroom floor rather than in the wardrobe or the laundry hamper.”

- “Against Cleaning” (Rawson and Muñoz, 2016)



Discussion

- *What does this excerpt suggest about the nuances of data cleaning?*
- *What does “clean” imply?*
- *How might you talk to researchers on your campus who would be uncomfortable with the idea of clean v. messy data?*



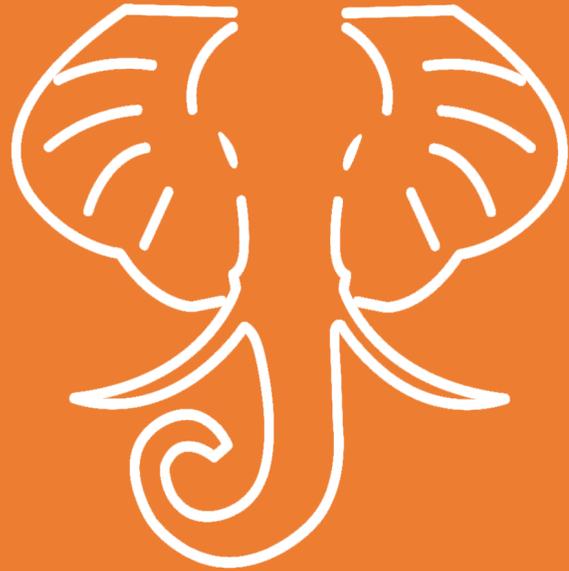
Questions?



References

- Denny, M. J. and Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. <https://ssrn.com/abstract=2849145> .
- National Endowment for the Humanities. (2017) *Data Management Plans for NEH Office of Digital Humanities Proposals and Awards*. Retrieved October 1, 2017, from https://www.neh.gov/files/grants/data_management_plans_2018.pdf .
- Rawson, K., & Muñoz, T. (2016). Against Cleaning. Retrieved August 16, 2017, from <http://curatingmenus.org/articles/against-cleaning/> .
- Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), 209–219. <https://doi.org/10.1093/lc/18.2.209> .





**Digging Deeper,
Reaching Further**

Module 4.1: Analyzing Textual Data

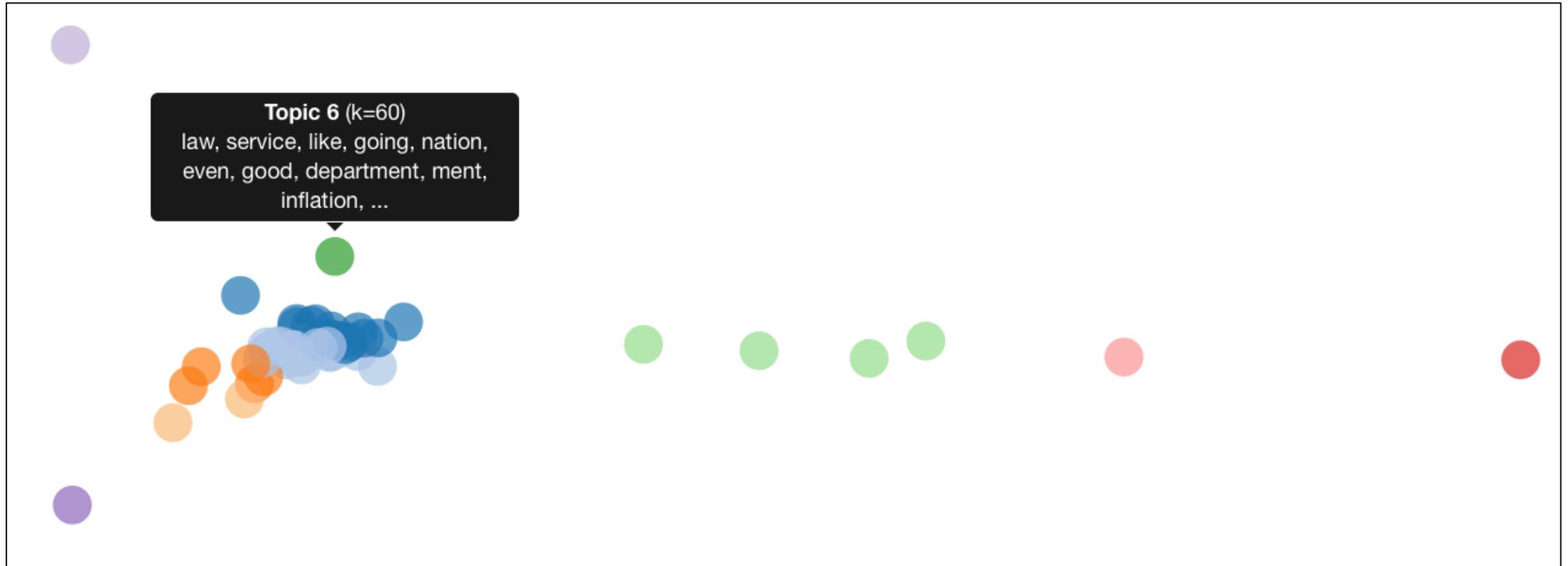
Using Off-the-Shelf Tools

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. 16

- 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. 17-18*

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. 17

The screenshot shows the HathiTrust Research Center Analytics website. At the top, there is a navigation menu with links for "HTRC Analytics", "Algorithms", "Data Capsules", "Worksets", "Datasets", and "Explore". On the right side of the menu, there are links for "Help", "About", and "rhan11". Below the navigation menu is the HathiTrust Research Center logo, which features an elephant head icon and the text "HATHI TRUST Research Center". To the right of the logo, the main heading reads "HathiTrust Research Center Analytics". Below this heading, a sub-heading states: "Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research." Below the main content area, there is a section titled "Featured Services" which contains three geometric diagrams: a cube with various labels, a rectangular prism with labels, and a pyramid with labels.



Analysis in the HTRC

 See Handout p. 17

Algorithms

Extracted Features Download Helper (v3.0.2)

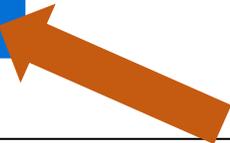
Generate a script that allows you to download extracted features data for your workset of choice. The script is a file containing a list of the rsync commands to access the volumes of the workset. After you download the script from HTRC Analytics, it can be run locally (from your computer), which will 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@lisalibrarian
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...

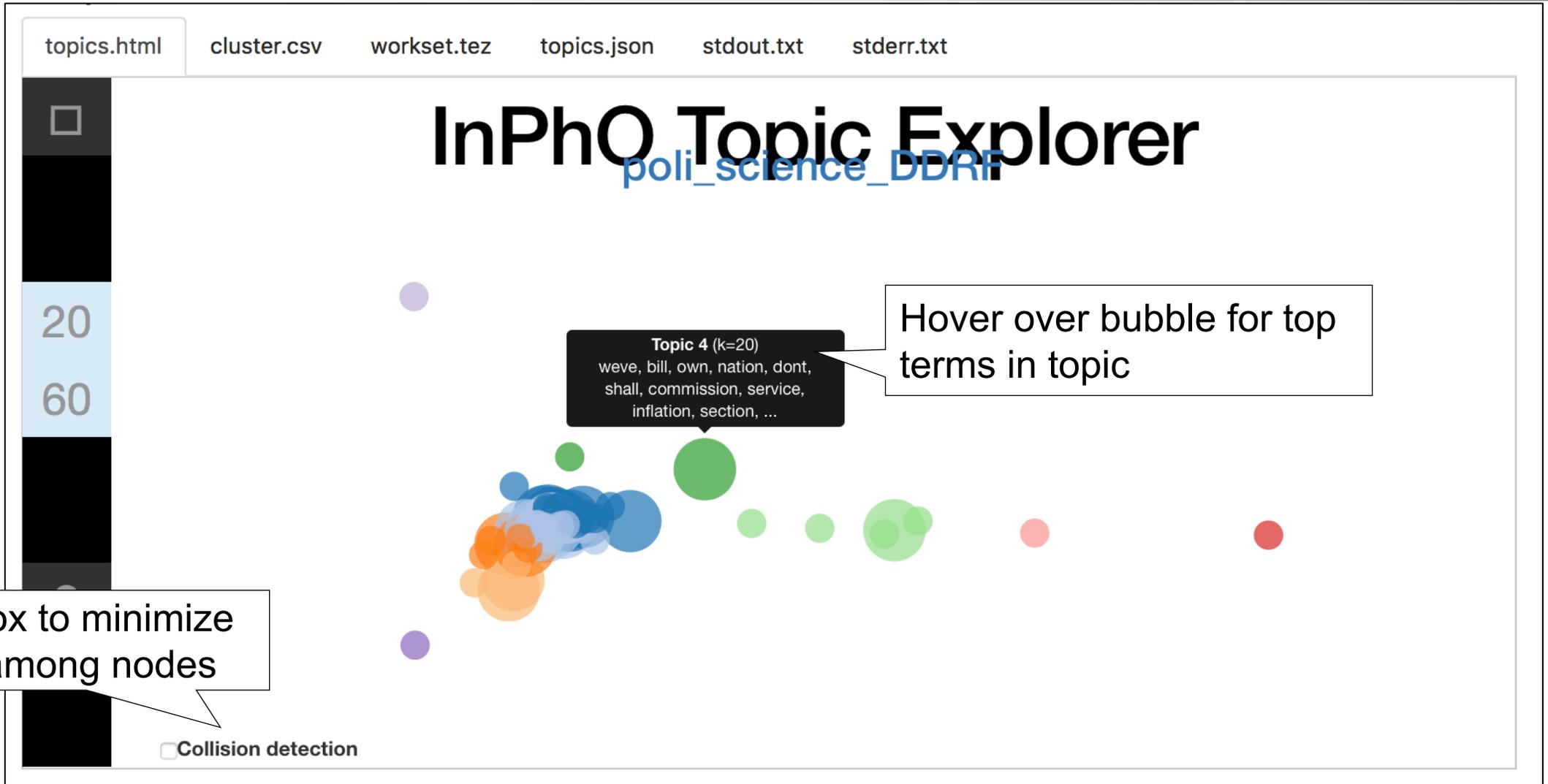
Job Name	Algorithm	Last Updated	Status	Actions
TestJobName	InPhO_Topic_Model_Explorer	2018-08-06 16:51:59	Staging	

Showing 1 to 10 of 1 entries

First << 1 >> Last



View results



Check box to minimize overlap among nodes

Hover over bubble for top terms in topic



Topics visualized



Results files

Output

topics.html cluster.csv worksheet.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?



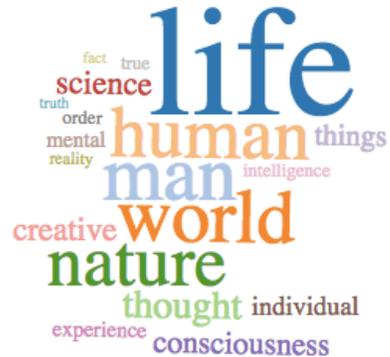
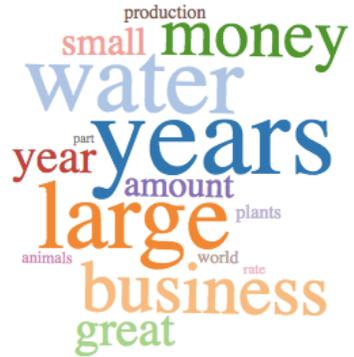
Case Study: *Inside the Creativity Boom*

- 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?



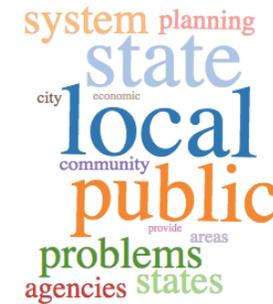
Tips for topic modeling

- 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*

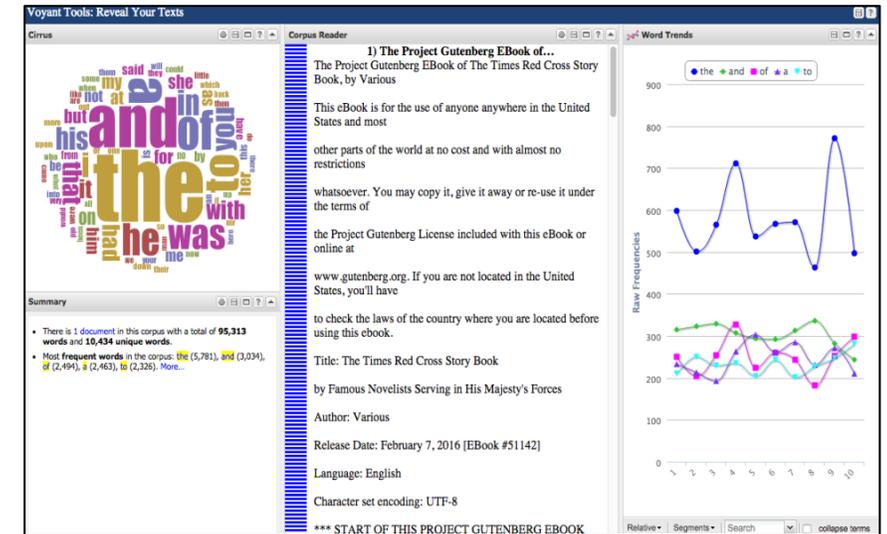
- Sam ended up using HTRC Extracted Features to get the data needed to analyze contemporary material
- The fits and starts of his project are a great real-world example!



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?

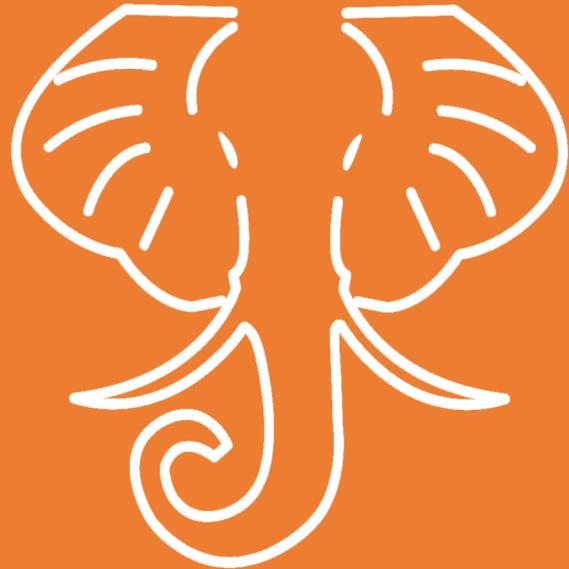


Questions?

References

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





**Digging Deeper,
Reaching Further**

Module 4.2: Performing Text Analysis

Basic Approaches with Python

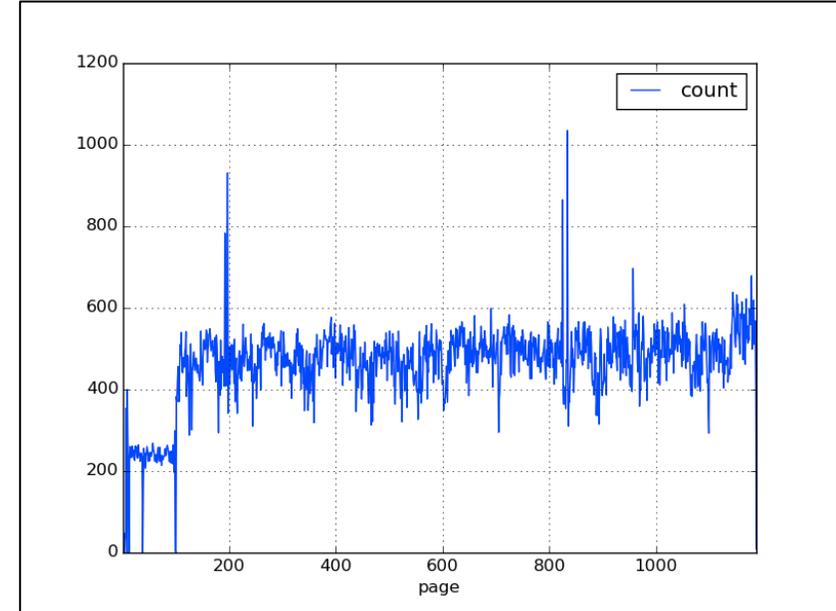
In this module we'll...

- Review text analysis strategies for advanced researchers
 - *Make skill-appropriate recommendations*
- Explore text analysis methods in-depth
 - *Understand the kinds of research available in the field*
- Use an HTRC dataset to conduct exploratory data analysis
 - *Practice programming skills for data-driven research*
- See how Sam analyzed his *Creativity Corpus*
 - *Learn how a researcher used text analysis methods*



Where we'll end up

```
13:45 ~ $ python top_adjectives.py 1970
count
token
_ 21248
other 17824
own 11552
new 10400
good 9984
great 9440
American 7936
many 7840
major 6480
last 6288
public 6224
important 6096
first 5808
such 5760
economic 5504
human 5424
international 5216
national 4864
same 4688
next 3920
nuclear 3840
local 3744
foreign 3696
political 3648
comprehensive 3488
few 3472
sure 3440
possible 3408
```



Use a Python library to create a list of top adjectives in a volume and a graph that visualizes word count across a volume.



Key approaches to text analysis

Broad Area: Natural Language Processing (NLP)

Using computers to understand the meaning, relationships, and semantics within human-language text

- **Specific Methods:**
 - **Named entity extraction:** what names of people, places, and organizations are in the text?
 - **Sentiment analysis:** what emotions are present in the text?
 - **Stylometry:** what can we learn from measuring features of style?



Key approaches to text analysis

Broad Area: Machine Learning

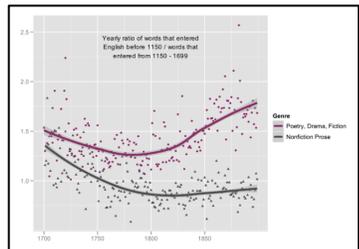
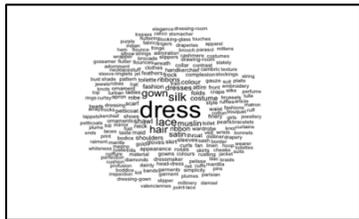
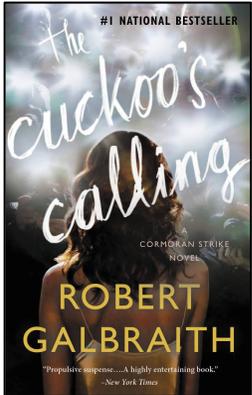
Training computers to recognize patterns.

- **Specific Methods**
 - **Topic modeling** – What thematic topics are present in the text?
 - **Naïve Bayes classification** – Which of the categories that I have named does the text belong to?



Activity: Identify the method

👉 See Handout p. 19



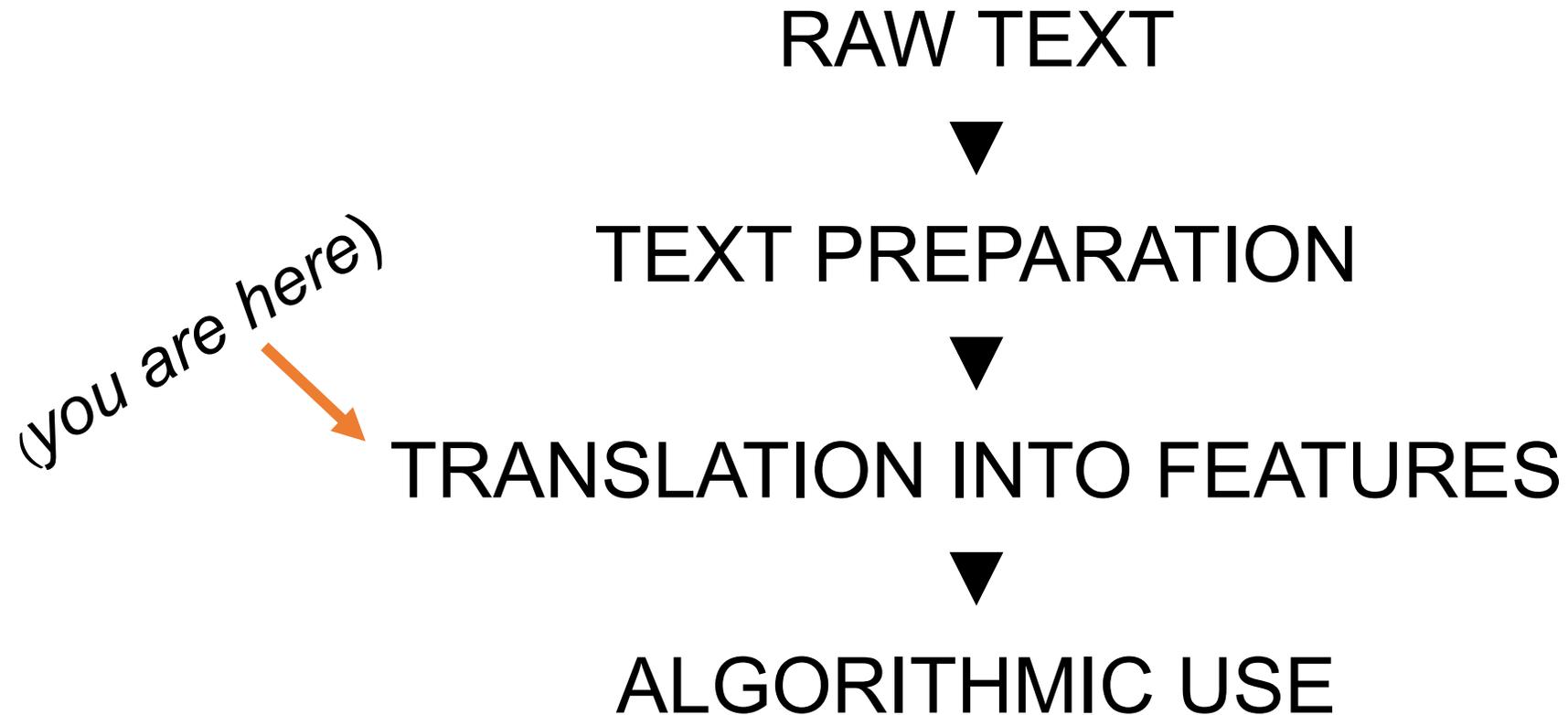
	Broad area	Specific method
'Rowling and "Galbraith"': an authorial analysis		
<i>Significant Themes in 19th Century Literature</i>		
<i>The Emergence of Literary Diction</i>		

Note:
Broad areas/specific methods are those defined in the previous two slides

Link to project summaries: <http://go.illinois.edu/ddrf-research-examples>



Text analysis workflow



Features in the HTRC

- HTRC Extracted Features dataset
- Downloadable
- Structured data consisting of features
- 5 billion pages, in 13.6 million volumes

<https://analytics.hathitrust.org/datasets#ef>



HTRC Extracted Features (EF)

- The features are
 - Selected data and metadata
 - Extracted from raw text
- Position the researcher to begin analysis
 - Some of the preprocessing is already done
- Form of non-consumptive access



Per-volume features

- Pulled from bibliographic metadata
- Title
- Author
- Language
- Identifiers

```
1 {
2   "id":"uc1.b3419888",
3   "metadata":{
4     "schemaVersion":"1.2",
5     "dateCreated":"2015-02-12T13:30",
6     "title":"Zoonomia = or The laws of organic life / by Erasmus Darwin.",
7     "pubDate":"1809",
8     "language":"eng",
9     "htBibUrl":"http://catalog.hathitrust.org/api/volumes/full/htid/uc1.b3419888.json",
10    "handleUrl":"http://hdl.handle.net/2027/uc1.b3419888",
11    "oclc":"3679915",
12    "imprint":"Thomas and Andrews, 1809."
13  },
14  "features":{
15    "schemaVersion":"2.0",
16    "dateCreated":"2015-02-20T23:58",
17    "pageCount":616,
18    "pages":[
```



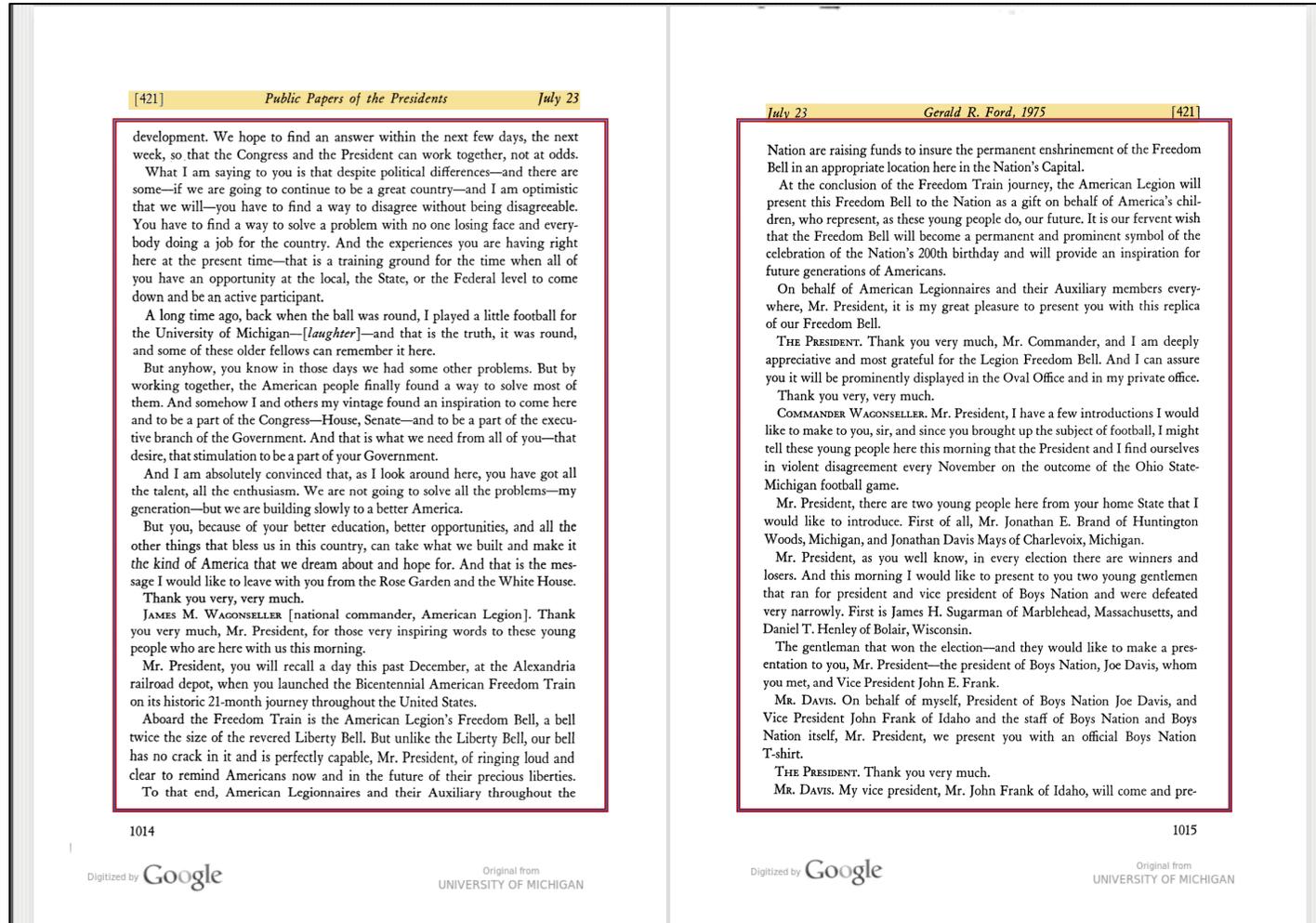
Per-page features

- Page sequence
- Computationally-inferred metadata
 - Word, line, and sentence counts
 - Empty line count
 - Language

```
20 {  
21   "seq": "00000035",  
22   "tokenCount": 507,  
23   "lineCount": 44,  
24   "emptyLineCount": 0,  
25   "sentenceCount": 14,  
26   "languages": [  
27     {  
28       "en": "1.00"}],
```



Page section features



Public papers of the presidents of the United States (Gerald R. Ford, book 2)



Page section features

LIST OF ITEMS	vii
CABINET	lxvii
PUBLIC PAPERS OF GERALD R. FORD, JULY 21—DECEMBER 31, 1975	1005
<i>Appendix A</i> —Additional White House Releases	2021
<i>Appendix B</i> —Presidential Documents Published in the Federal Register	2049
<i>Appendix C</i> —Presidential Reports to the 94th Congress, 1st Session . .	2057
<i>Appendix D</i> —Rules Governing This Publication	2061
INDEX	A-1

*Public papers of the
presidents of the
United States
(Gerald R. Ford,
book 2)*



Page section features

Header, body, footer

- Line, empty line, and sentence count
- Counts of beginning- and end-line characters
- Token counts
 - Homonyms counted separately
 - Part-of-speech codes are from the Penn Tree Bank

```
40 "body":{
41   "tokenCount":504,
42   "lineCount":43,"
43   "emptyLineCount":0,"
44   "sentenceCount":12,
45   "tokenPosCount":{
46     "fynthefis":{"NNP":1},
47     "Laws":{"NNP":1},
48     "beautiful":{"JJ":1},
49     "philofopher":{"NN":1},
50     "uponthe":{"IN":1},
51     "for":{"IN":1},
```



Using HTRC Extracted Features

- Identify parts of a book
 - From descriptive metadata
- Perform any method that works with bags-of-words
 - Topic modeling
 - Dunning's log-likelihood
- Classify volumes
 - Compare with bibliographic metadata



Do-it-yourself text analysis

- Some researchers won't use off-the-shelf tools
 - Want more control over processes
 - Set their own parameters
- Mix-and-match approaches to create “toolkit” of strategies



The toolkit

- Researcher-dependent
- Requires understanding of statistics
- Often draws on expert collaborators
- Consists of command line tools and programming languages



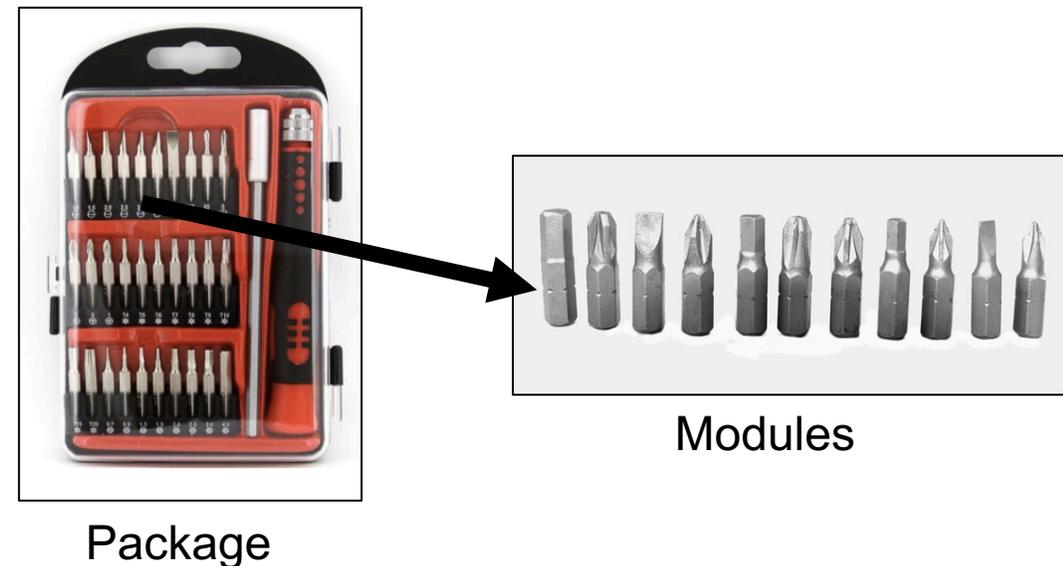
Command-line tools

- Tools you download and run from the command line
- MALLET (Java-based)
 - Topic modeling and classification
- Stanford NLP (Java-based)
 - Natural language processing

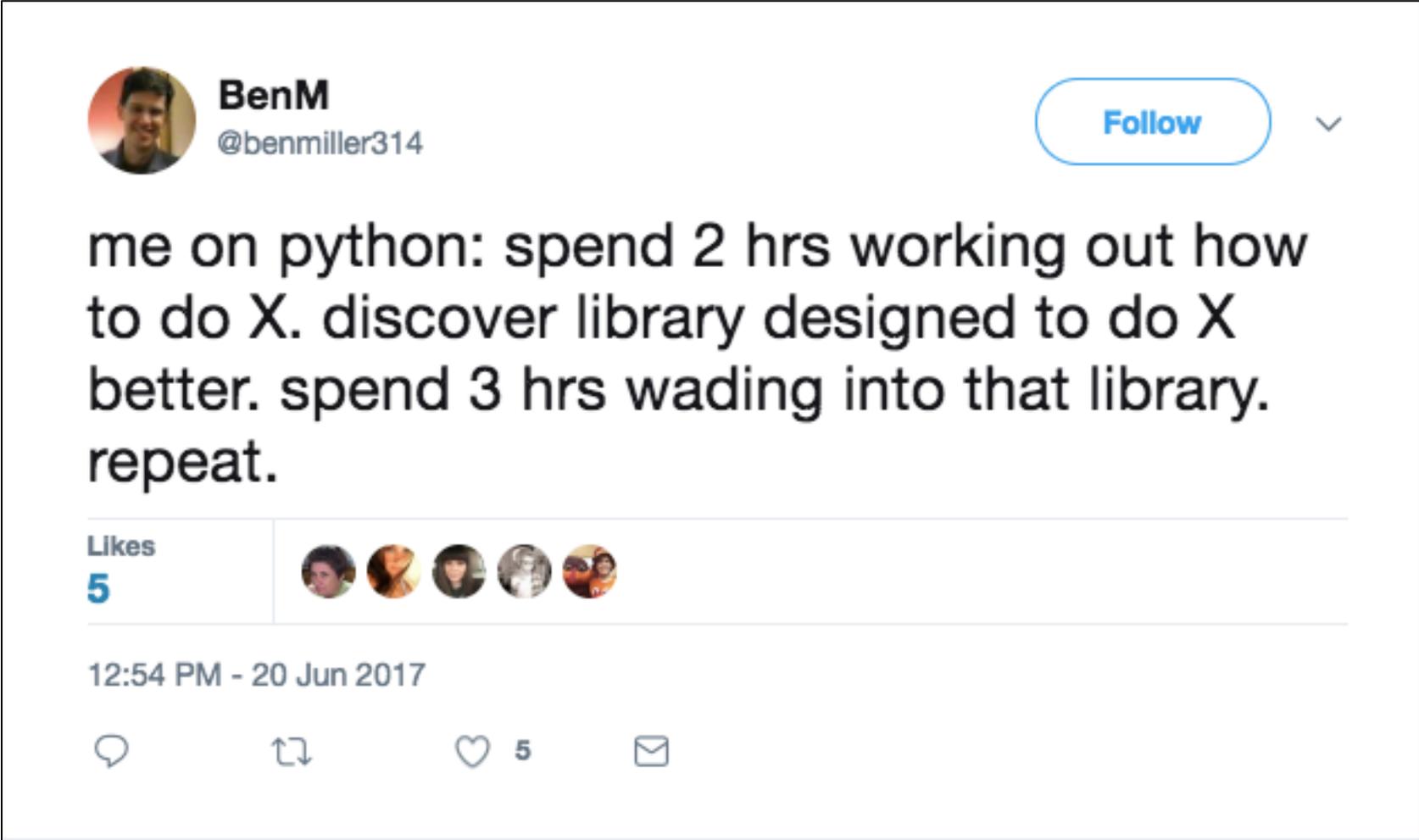


Code libraries

- Part of programming is re-using code, often shared as ***libraries***
- The existing code is installed via ***packages***
- Packages are collections of bits of code, made up of ***modules***
- Modules facilitate programming tasks



Using libraries in practice



 **BenM**
@benmiller314 Follow

me on python: spend 2 hrs working out how to do X. discover library designed to do X better. spend 3 hrs wading into that library. repeat.

Likes **5**

12:54 PM - 20 Jun 2017

   5 



Installing packages for Python

- Install using a package manager
- Example: pip
 - Package manager for Python
 - Generally comes with Python install
- Basic syntax
 - `pip install <name>`
- Other installers: Homebrew, Conda



Using a library's module in a script

- Once library is installed...
- Must declare that it will be used at the top of the script
 - “Import” the module
- Use the desired module in the body of the script

```
 / home / elfdickson / get_verbs.py Keyboard shortcuts: Normal  
1 from htrc_features import FeatureReader  
2 import pandas as pd  
3 import sys  
4  
5  
6 # if __name__ == '__main__':  
7 #     if len(sys.argv) < 2:  
8 #         print("Missing filename argument")  
9 #         sys.exit(1)  
10  
11  
12 filename = sys.argv[1]  
13 idx = pd.IndexSlice  
14  
15 def get_proper_nouns(vol):  
16     tl = vol.tokenlist(pages=False)  
17     tl.index = tl.index.droplevel(0)  
18     #     tl['date'] = vol.year  
19     #     tl = tl.set_index('date', append=True).reorder_levels(['date', 'token', 'pos'])  
20     #     tl = tl.set_index#.reorder_levels(['token', 'pos'])  
21     try:  
22         proper_nouns = tl.loc[idx[:, :, ('VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ')],]  
23         proper_nouns.index = proper_nouns.index.droplevel(2)  
24         return proper_nouns[proper_nouns['count'] < 150]  
25     except:  
26         return pd.DataFrame()
```

Imports modules, i.e. the components of the package

Use the module

Python text analysis libraries

- Machine learning
 - SciKit Learn
- Data science
 - Pandas
- Natural language processing
 - Natural Language Toolkit (NLTK)
- **Example:** `nltk.words_tokenize()`
 - *Can you guess what this does?*



HTRC Feature Reader Python library

- Python library for working with HTRC Extracted Features
 - Bits of pre-written code to parse the JSON
- Install using a package manager, like pip
 - Source code lives on Github
- Also need to have Pandas installed
 - Remember: pandas = Python library for working with data
 - Luckily it's included in PythonAnywhere!



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: view adjectives in Extracted Features file



Hands-on activity

 *See Handout pp. 19-20*

In this activity, you will install the Feature Reader library and run a Python script to create a list of the most-used adjectives and the number of times they occur in different sets of volumes of presidential papers (using the Extracted Features files of these volumes).

What You Need:

Script: `top_adjectives.py`

Files for analysis, in 2 directories:

- 1970
- 1930



Files and directories

- 1930 and 1970 directories and required files should already be in your files in PythonAnywhere
- Check using the “Browse files” button
- 1970/ = Extracted Features files of presidential speeches from the 1970s (corresponds to our political science workset!)
- 1930/ = Extracted Features files of presidential speeches from the 1930s



Install Feature Reader library

Bash console 4179031

```
19:56 ~ $ pip install --user htrc-feature-reader
```

```
pip install --user htrc-feature-reader
```



Examine code

```
1 from htrc_features import FeatureReader
2 import glob
3 import pandas as pd
4 import sys
5
6 decade = sys.argv[1]
7
8 paths = glob.glob(decade+'/*.json.bz2')
9 idx = pd.IndexSlice
10 fr = FeatureReader(paths)
11 vol = next(fr.volumes())
12 tl = vol.tokenlist(pages=False)
13 tl.index = tl.index.droplevel(0)
14 adjectives = tl.loc[idx[:,('JJ')],]
15 adj_dfs = [adjectives for vol in fr.volumes()]
16 all_adj = pd.concat(adj_dfs).groupby(level='token').sum().sort('count', ascending=False)[:50]
17
18 print(all_adj)
```

Prepares us to use the libraries needed for this task.

'JJ' is the code for adjectives from the Penn TreeBank.



Open shell and run command

```
13:45 ~ $ python top_adjectives.py 1970
```

```
python top_adjectives.py 1970
```



View results

```
13:45 ~ $ python top_adjectives.py 1970
count
token
_ 21248
other 17824
own 11552
new 10400
good 9984
great 9440
American 7936
many 7840
major 6480
last 6288
public 6224
important 6096
first 5808
such 5760
economic 5504
human 5424
international 5216
national 4864
same 4688
next 3920
nuclear 3840
local 3744
foreign 3696
political 3648
comprehensive 3488
few 3472
sure 3440
possible 3408
```



On your own

 See Handout p. 20

- **Compare the adjectives used by presidents in the 1970s with those used in the 1930s.**
- **Work and discuss with your neighbor.**
- *How do you need to change your command?*
- *What differences do you see? Similarities?*

Challenge

Try modifying the script to search verbs or another part of speech.

(Hint: Don't forget the Penn Tree Bank!)



Exploratory data analysis

- Approach for getting familiar with data
- Easier to recognize patterns (and problems)
 - Hard to see trends in a spreadsheet or text file!
- There are whole books about it
- Strategies:
 - Plot raw data
 - Plot simple statistics
 - Compare plots to look for patterns



Visualization libraries

▪ Python

- *Matplotlib* library, *pyplot()* function
- *ggplot* library

▪ Others:

- R: *ggplot2*
- D3.js: JavaScript library for visualizations



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:

View word count over a volume using an Extracted Features file



Hands-on activity

👉 *See Handout p. 21*

- In this activity, you will create a visualization using the the native plotting functionality in Pandas, called pyplot, to see the word count over a volume based on its Extracted Features file

Script: `word_count.py`

Files for analysis: `mdp.49015002203033.json.bz2` (*Presidential Papers of the Presidents of the United States: Gerald Ford*)



Examine code

```
1 from htrc_features import FeatureReader
2 #import os
3 import glob
4 #import pandas as pd
5 import matplotlib
6 import matplotlib.pyplot as plt
7 matplotlib.use("Agg")
8
9
10 path = glob.glob('1970/mdp.49015002203033.json.bz2')
11 fr = FeatureReader(path)
12 vol = fr.first()
13
14 tokens = vol.tokens_per_page()
15
16 tokens.plot()
17 plt.savefig("words.png")
18
19
```

Imports our modules so we can call functions later, including matplotlib's pyplot

Set variable "tokens"

Plot tokens and save plot as png

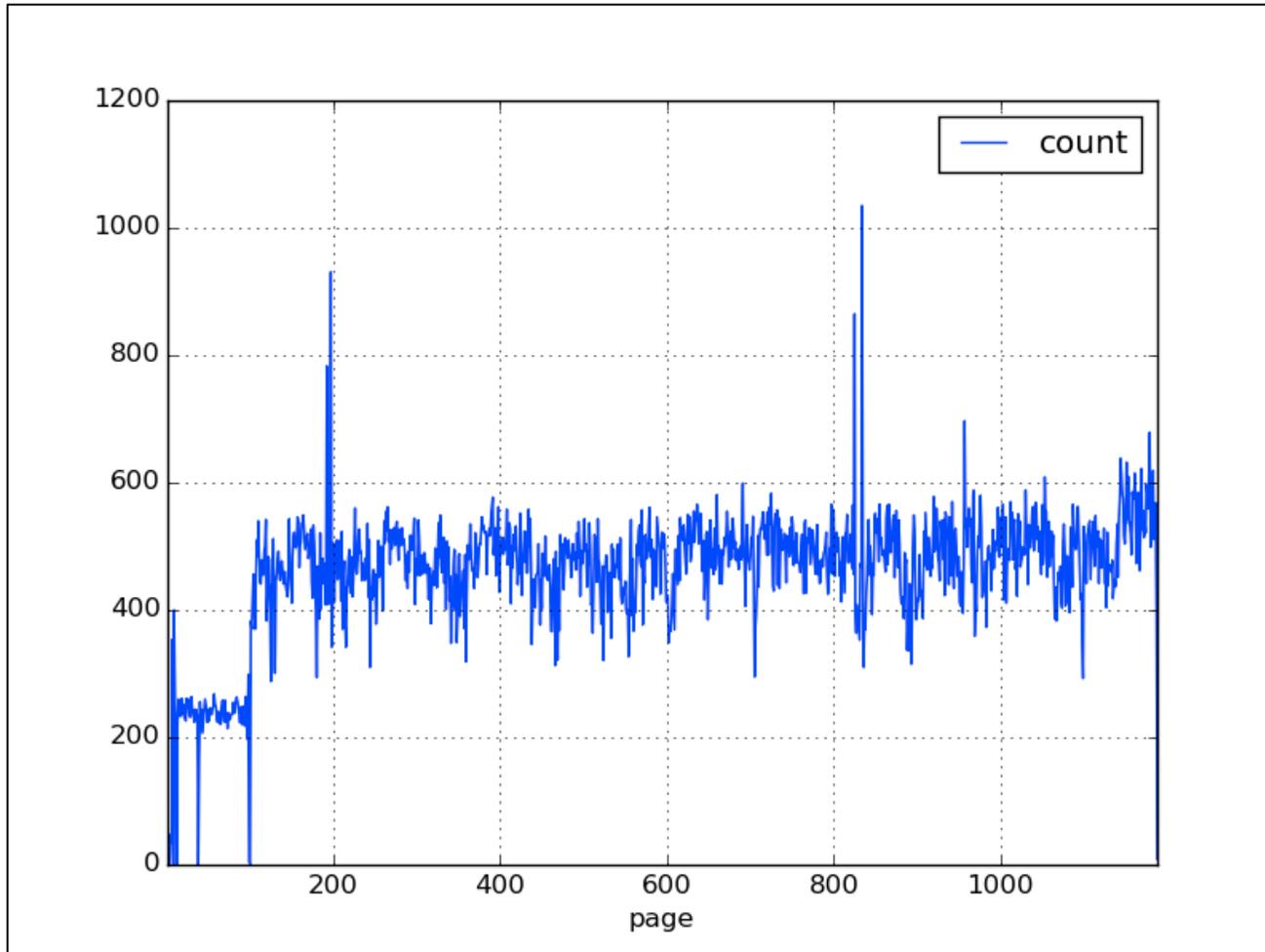


Open shell and run command

```
12:28 ~ $ python word_count.py
```



View graph



The graph visualizes how many words appear on each page in the volume.

Page numbers are on the horizontal axis, and the word count is on the vertical axis.

E.g. The 600th page has about 350 words.



On your own

 *See Handout p. 21*

- Can you modify the script to look at another volume?
- Edit, save, and run the file
 - Hint: Change the path to another volume in the script
- View and discuss with your neighbor

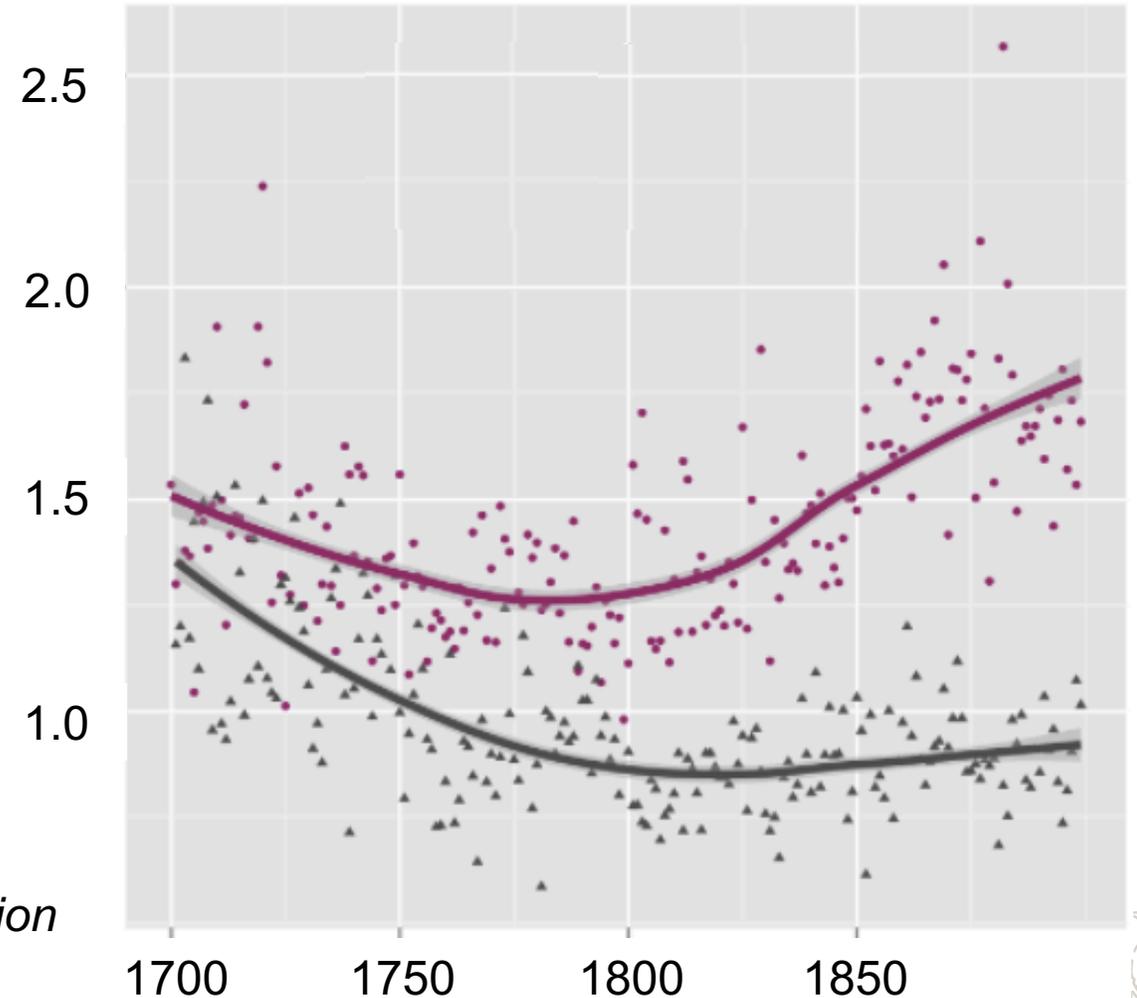


Another look

From paper - graph of diction patterns
between genres, using frequency counts
(Underwood and Sellers, 2012)

- Even basic counts can lead to sophisticated research
- Ted used HTRC EF
- Classified genre at the page level

Genre
— Poetry, Drama, Fiction
— Nonfiction Prose



Case Study: *Inside the Creativity Boom*

After reducing Creativity Corpus to pages containing forms of creativ*:

- Performed topic modeling on those pages
- Ended up with topics that reflect what themes are prevalent around concept of “creativity” in the 20th century
- Graphed the topics over time to see how their usage changed

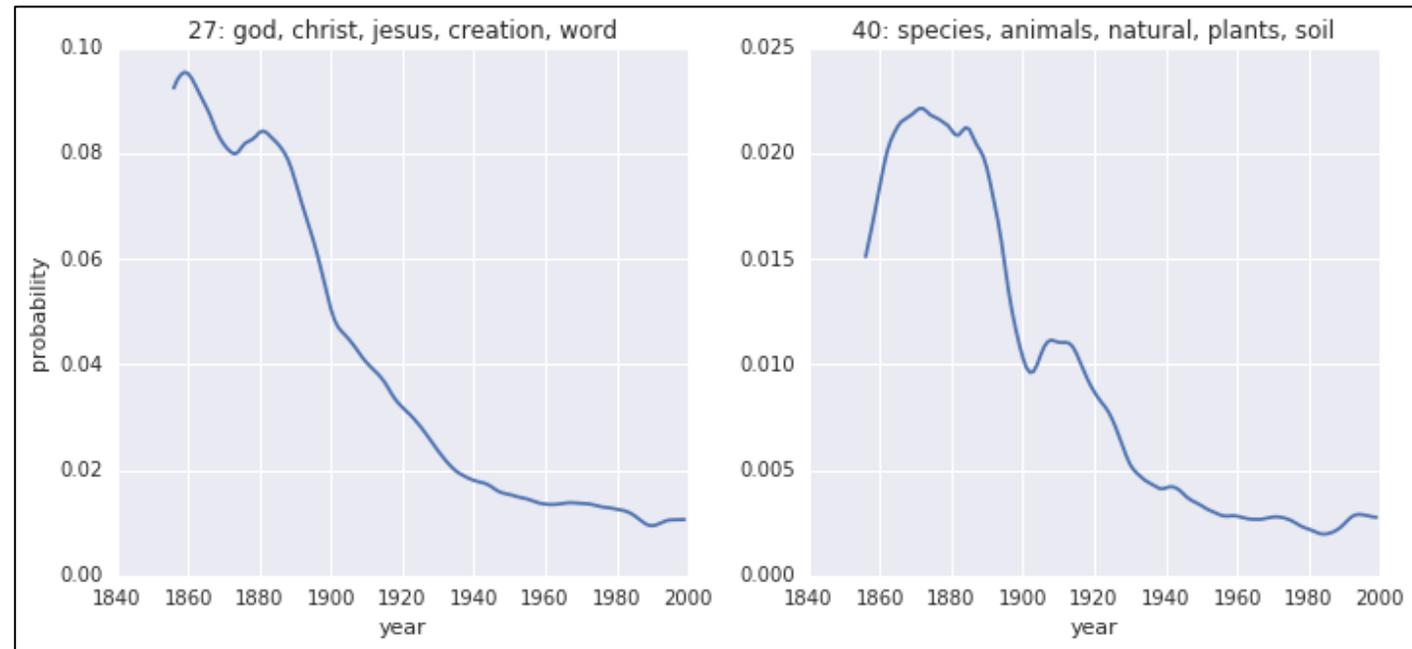


Case Study: *Inside the Creativity Boom*

■ Topics that decreased in usage over time

- god, christ, jesus, creation, word
- species, animals, natural, plants, soil
- nature, mind, creative, world, human
- invention, power, creative, own, ideas

Creativity topics with falling usage



Case Study: *Inside the Creativity Boom*

- Topics that increased in usage over time
 - advertising, media, marketing, sales, television
 - economic, development, capital, economy, production
 - poetry, language, poet, poets, poems
 - social, creative, study, development, behavior

Creativity topics with increasing usage



Discussion

- *In what ways can librarians support advanced text analysis research?*
- *What additional skills would you need to learn in order to do so?*

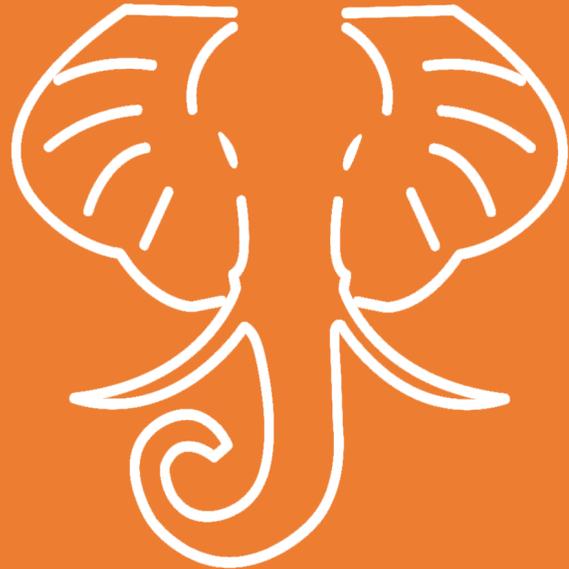


Questions?

References

- Underwood, T., & Sellers, J. (2012). The emergence of literary diction. *Journal of Digital Humanities*, 1(2), 1-2.
<http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/> .





**Digging Deeper,
Reaching Further**

Module 5: Visualizing Textual Data

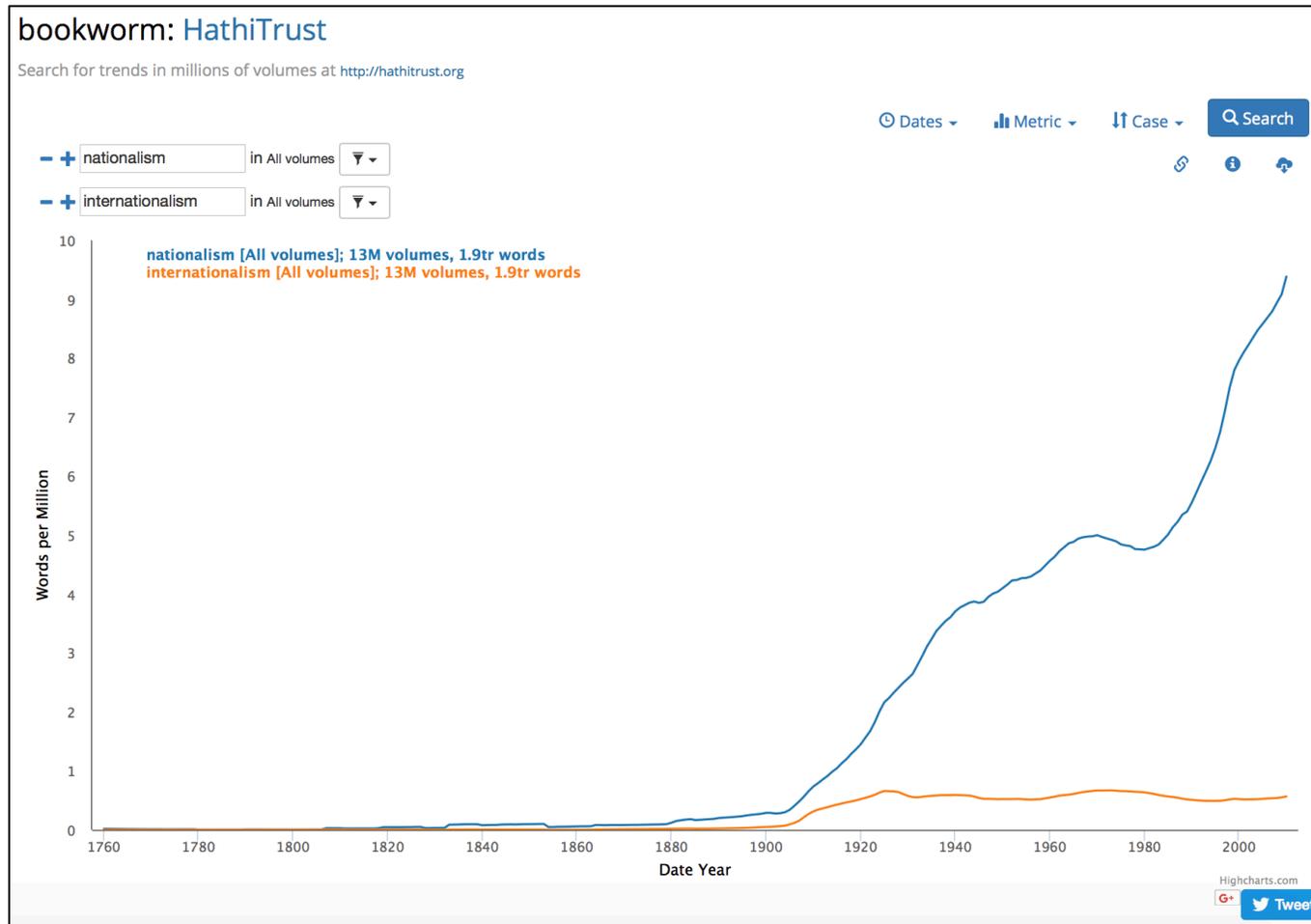
An Introduction

In this module we'll...

- Introduce common visualization strategies for text data
 - *Communicate with researchers about their options*
- Use a web-based visualization tool, HathiTrust+Bookworm
 - *Gain experience creating and reading data visualizations*
- See how Sam used HathiTrust+Bookworm for his project
 - *Learn how HT+BW was utilized in research*



Where we'll end up



Create visualization of word usage trends across the HathiTrust corpus.



Data visualization

- Data visualization is the process of converting data sources into a visual representation
- Visualizations present particular ways of interpreting data
- Data visualization is an entire field of study; we're barely scratching the surface



Why visualize text data?

- Understand broader themes of a dataset
- Explore patterns in the data
- Cluster texts for overview or classification
- Compare data to other data (e.g., correlating with social networks)



Place in research process

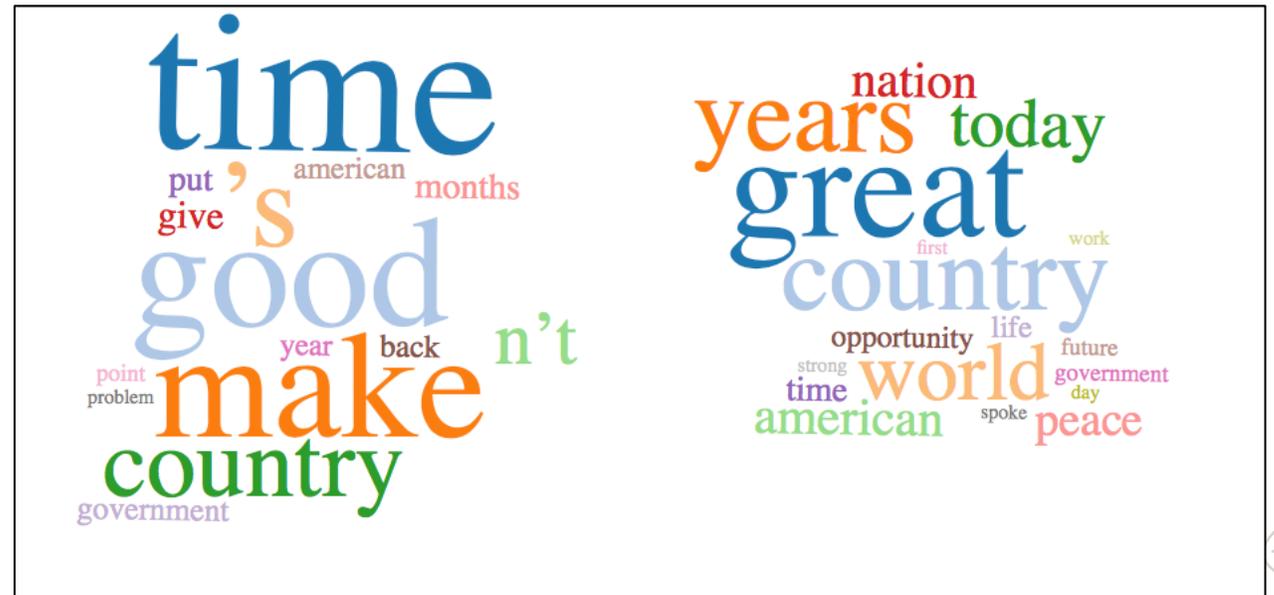
- In the earlier exploration stage of a project:
 - Explore full range of data
 - Discover characteristics and themes in data
- In the later explanation stage of a project:
 - Communicate findings to others in a clearer and more efficient way



Common text data visualizations

Word cloud

- Relatively unsophisticated, but effective
- Size of word relates to prominence or salience



Topic models from
HTRC Algorithms

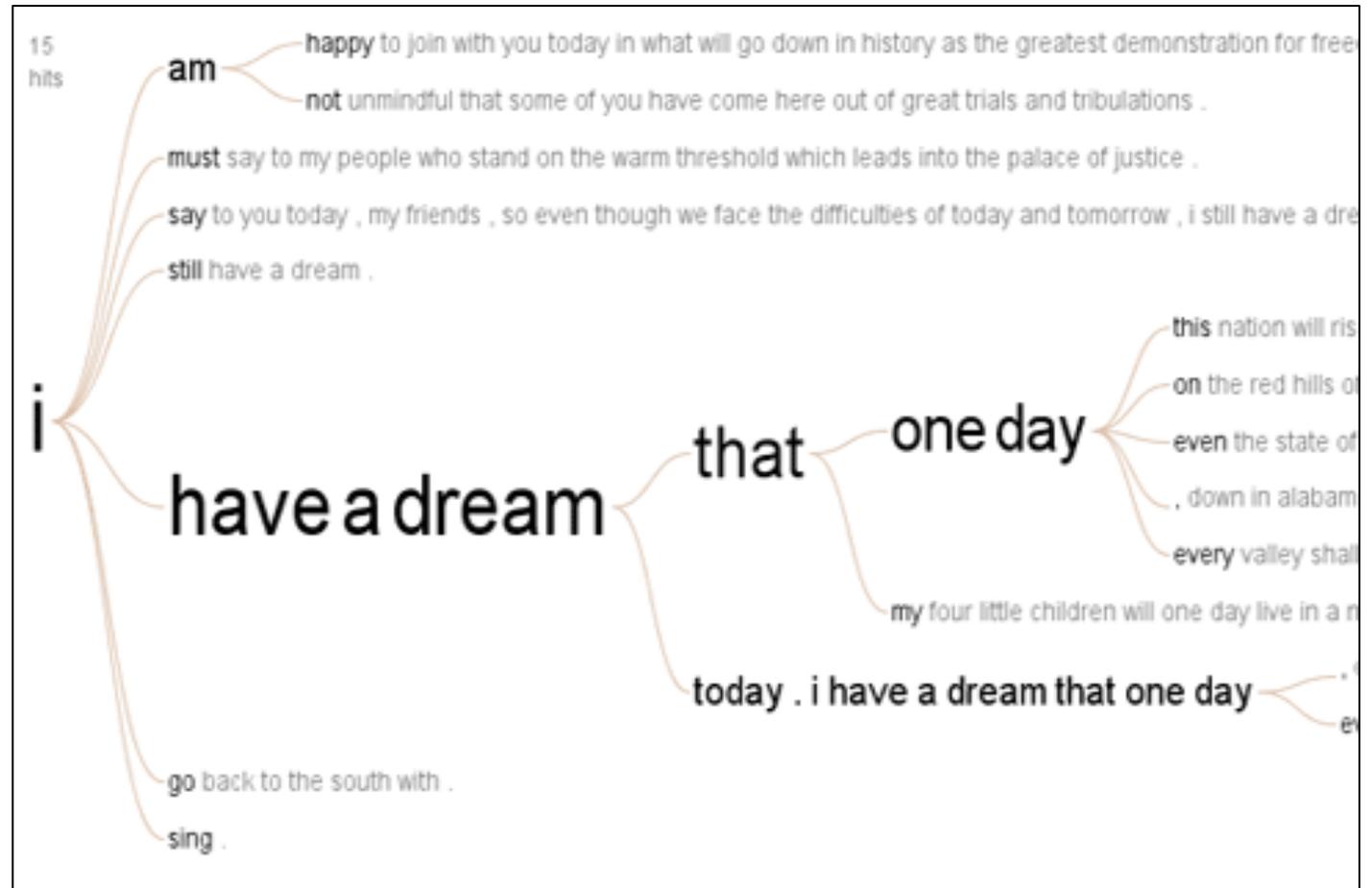


Common text data visualizations

Trees or hierarchies

■ Word trees

Occurrences of “I have a dream” in
Martin Luther King’s historical
speech.
(Wattenberg and Viégas, 2008)

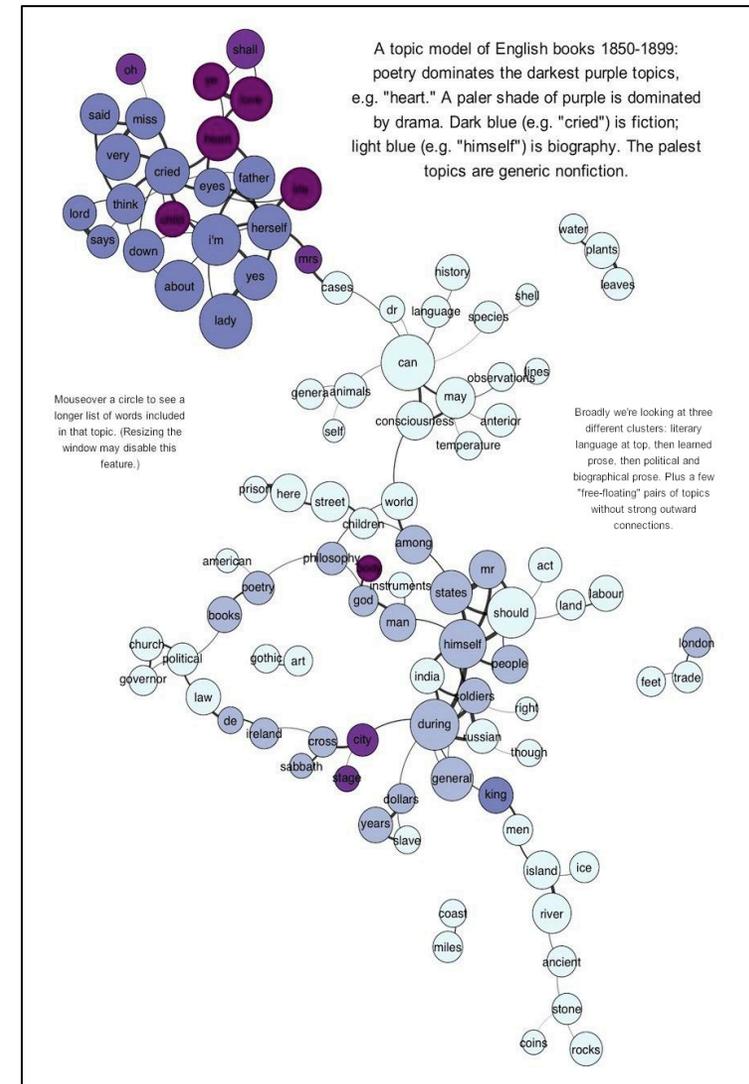


Common text data visualizations

Networks

- Node-link diagrams
- Good for representing topic models
- Visualize connections between named entities

Topic model of English books,
1850-1899
(Underwood, 2012)



Common text data visualizations

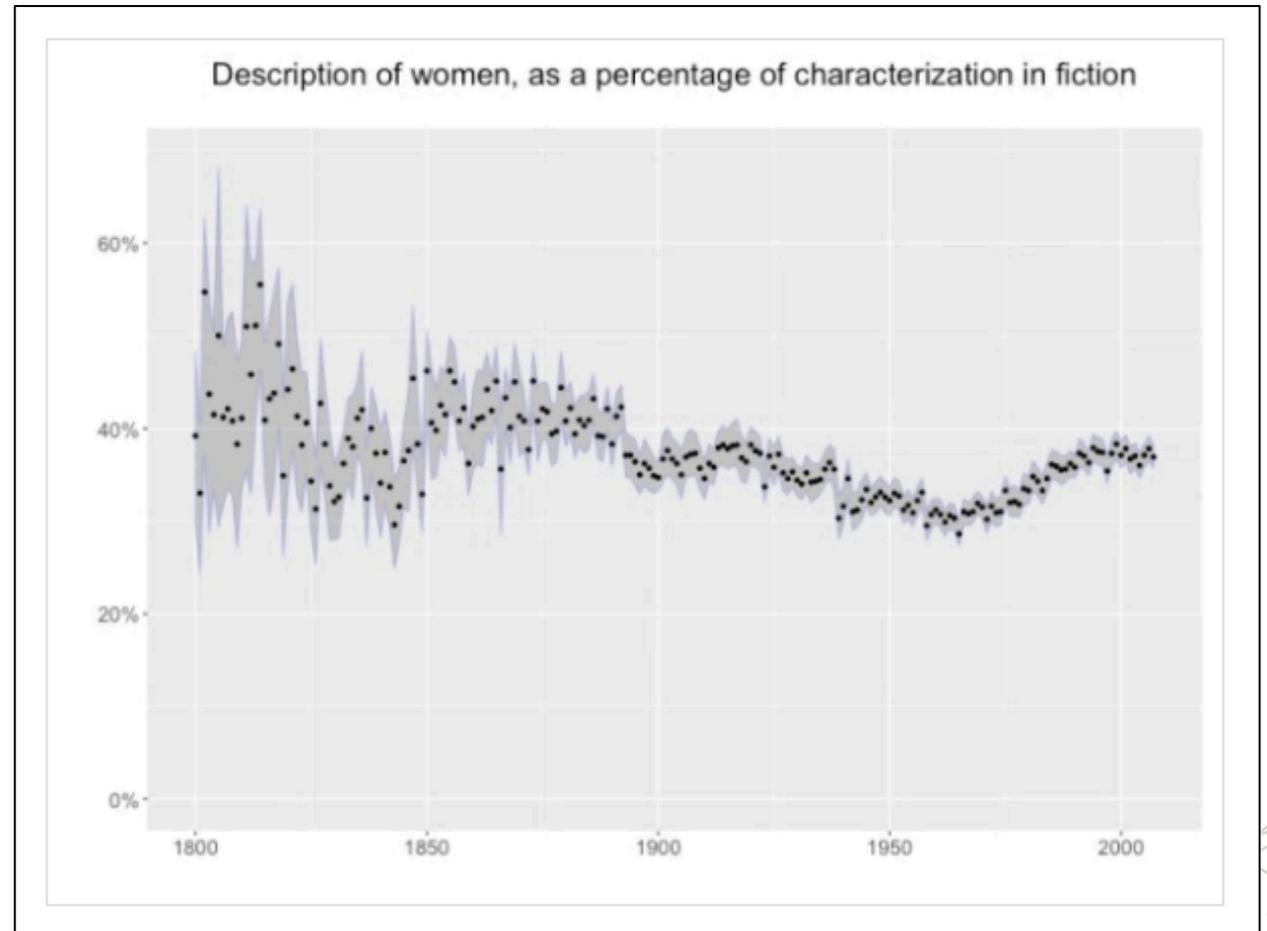
Temporal- or spatial-based visualizations

- Temporal visualizations

Percent representation of female characters
in English literature

(Underwood and Bamman, 2016)

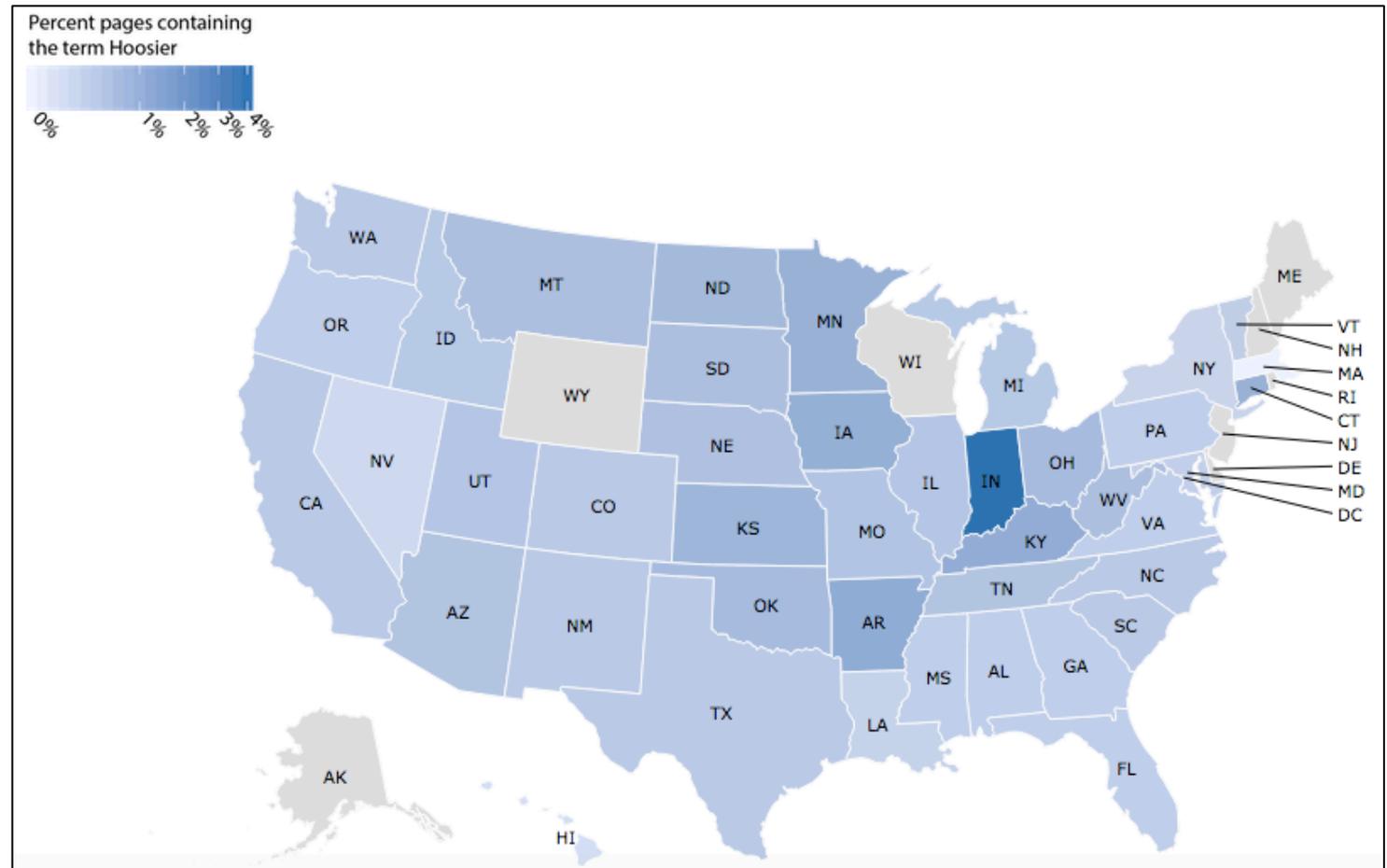
<https://tedunderwood.com/2016/12/28/the-gender-balance-of-fiction-1800-2007/>



Common text data visualizations

Temporal or spatial visualizations

- Maps



Percent of newspaper pages containing the term “hoosier” (Palmer, Polley, & Pollock, n.d.)

<http://centerfordigschol.github.io/chroniclinghoosier/map1.html>

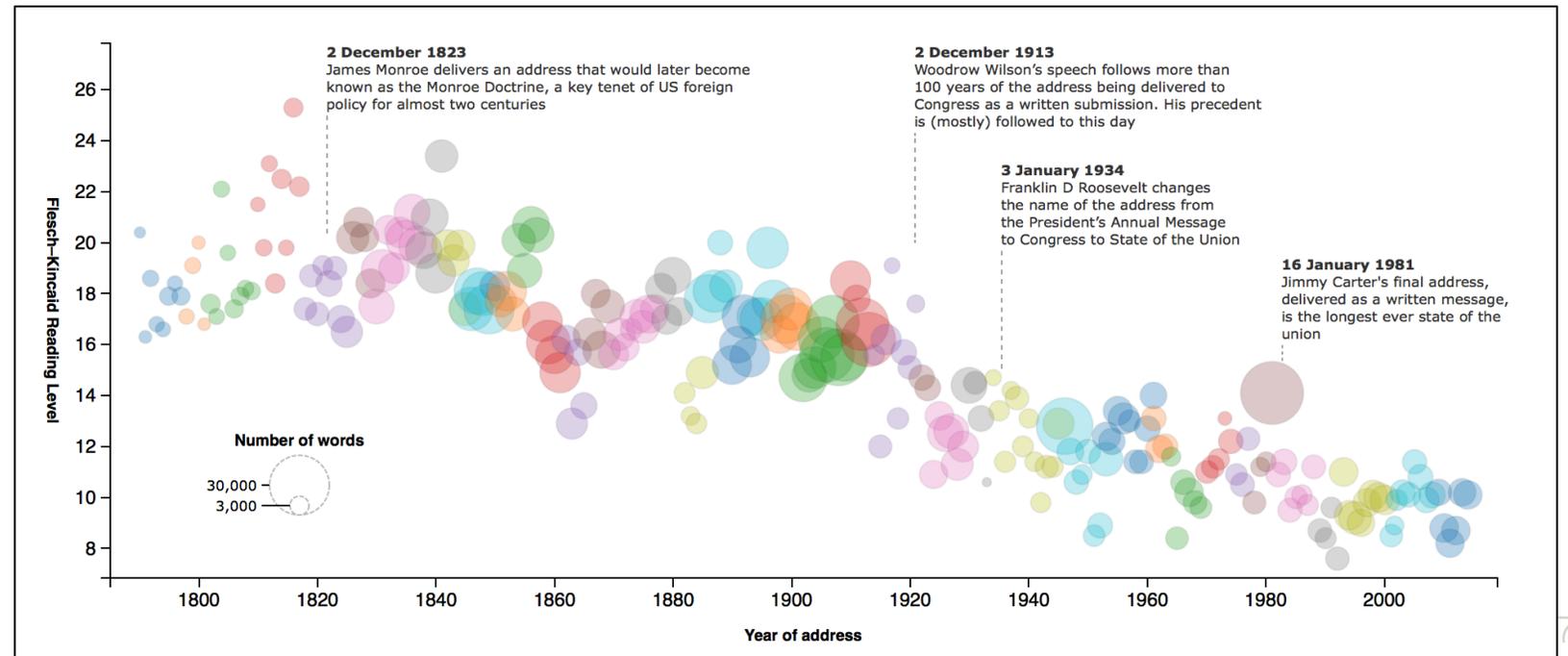


Common text data visualizations

Other “multi-dimensional” visualizations

- Bubble charts
- Heat maps

Bubble chart: readability of U.S. presidential speeches (The Guardian, 2013)



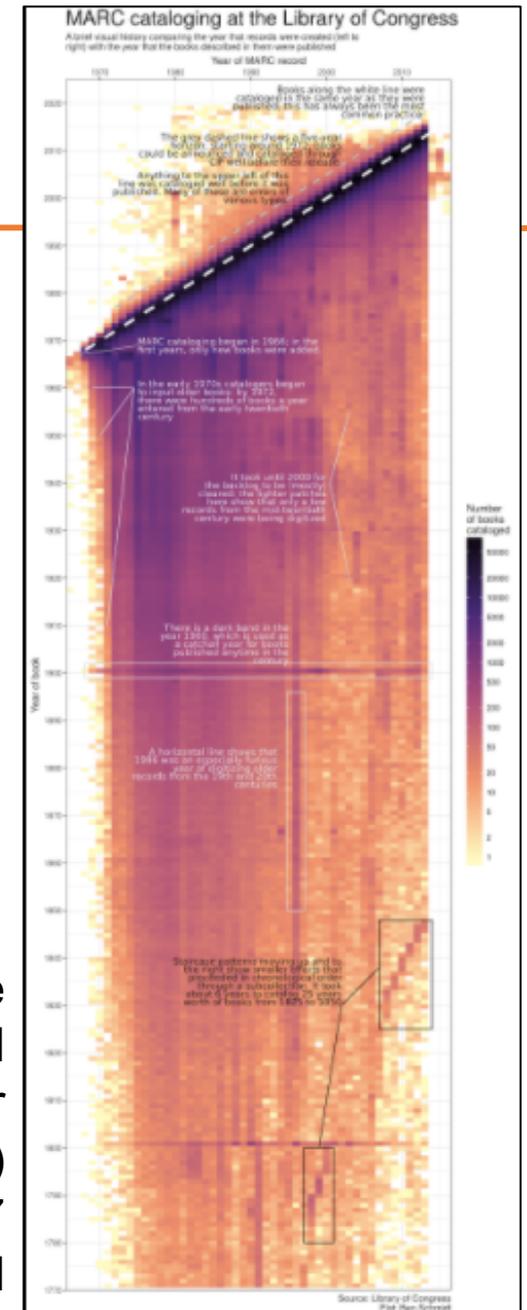
Common text data visualizations

Other “multi-dimensional” visualizations

- Heat maps

Heatmap of MARC cataloging at the Library of Congress by book year and cataloging year (Schmidt, 2017)

<http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html>



Activity

👉 See Handout p. 22

Match type of use to the type of visualization:

Visualization	What would it be good for?	Uses
Word cloud		Change over time
Trees or hierarchies		Spatial
Networks		Topical density
Timeline		Relationships
Map		Word distribution
Bubble chart		
Heatmap		

** Bonus: what kinds of variables (i.e. data points) you would need for each visualization?



Common visualization tools

- **Word clouds**

- Voyant
- Wordle

- **Word use trends**

- Google Books Ngram Viewer
- HathiTrust+Bookworm

- **Tabular data visualization**

- Tableau

- **Mapping**

- ArcGIS Online with StoryMaps
- Tableau

- **Network graphs**

- Gephi
- NodeXL
- DH Press



Common visualization libraries

- **Python**

- matplotlib, pyplot
- ggplot library

- **R**

- ggplot2

- **D3.js**

- Javascript library for visualizations



Review: key terms in text analysis

N-gram

A contiguous chain of n items from a sequence of text where n is the number of items. Example: Bigram.

four score, score and, and seven, seven years, years ago, ago our, our fathers, fathers brought, brought forth, forth on, on this, this continent, continent a, a new, new nation, nation conceived, conceived in, in liberty, liberty and...



N-gram visualization: HathiTrust + Bookworm

Brings together:

- Text data (unigrams)
 - Bibliographic metadata
- HathiTrust
- Visualization tool
 - Track trends in a repository
- Bookworm

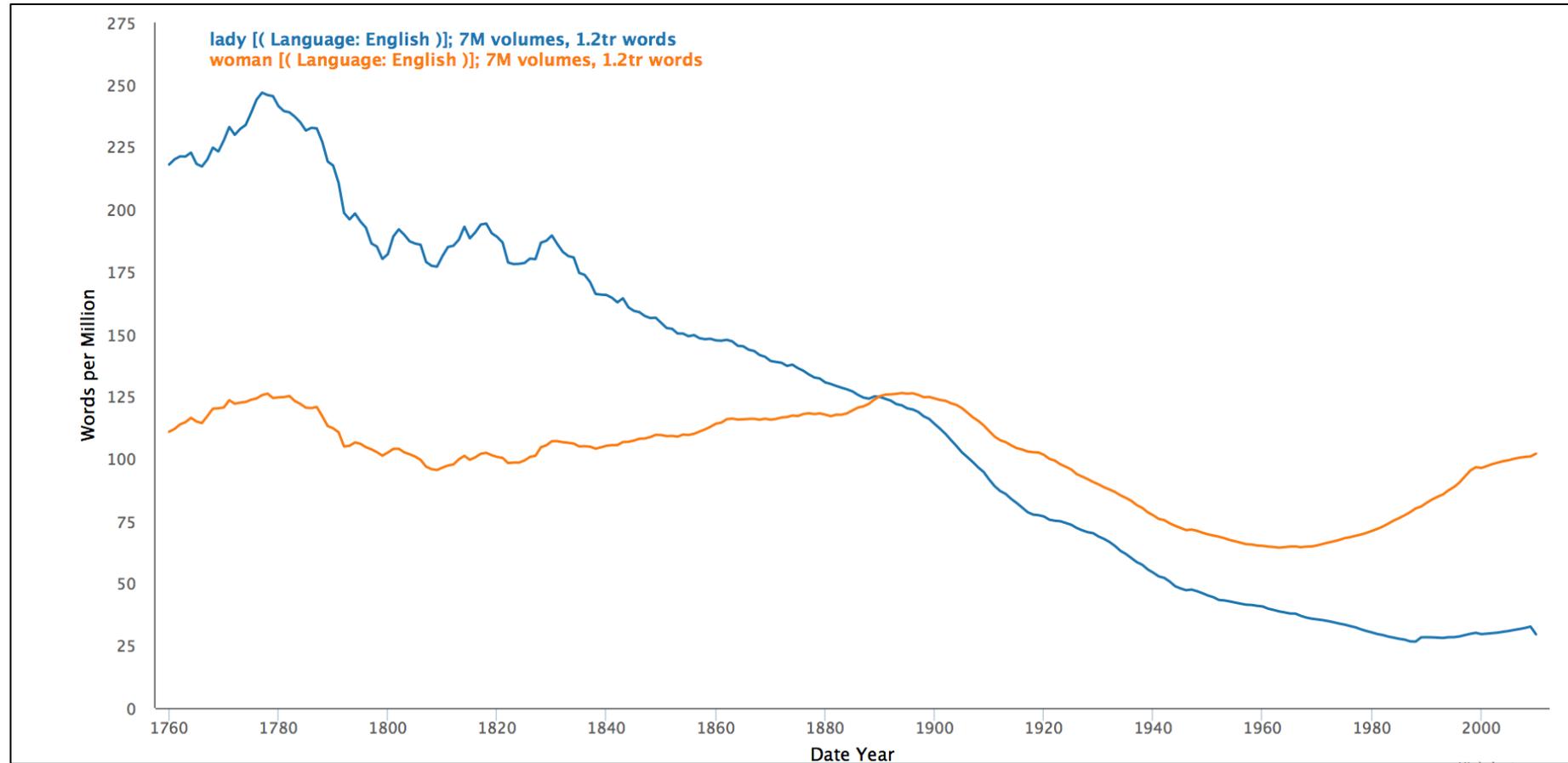


Bookworm framework

- Visualizes categories
- The category is plotted along the x-axis
 - Often plot years along the x-axis
 - Can plot other things!
- HathiTrust+Bookworm is just one implementation of the framework



Example HT+Bookworm view

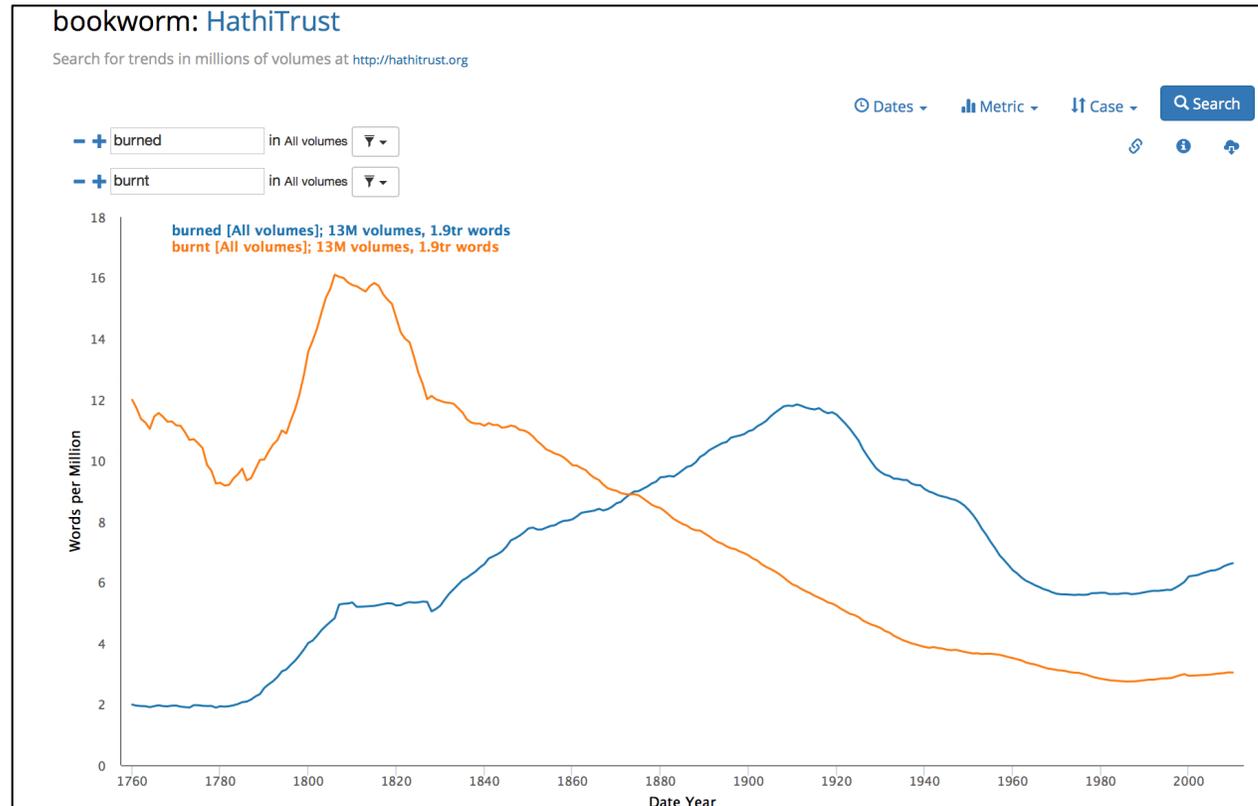


Track social change: lady vs. woman over time



Reading an HT+BW graph

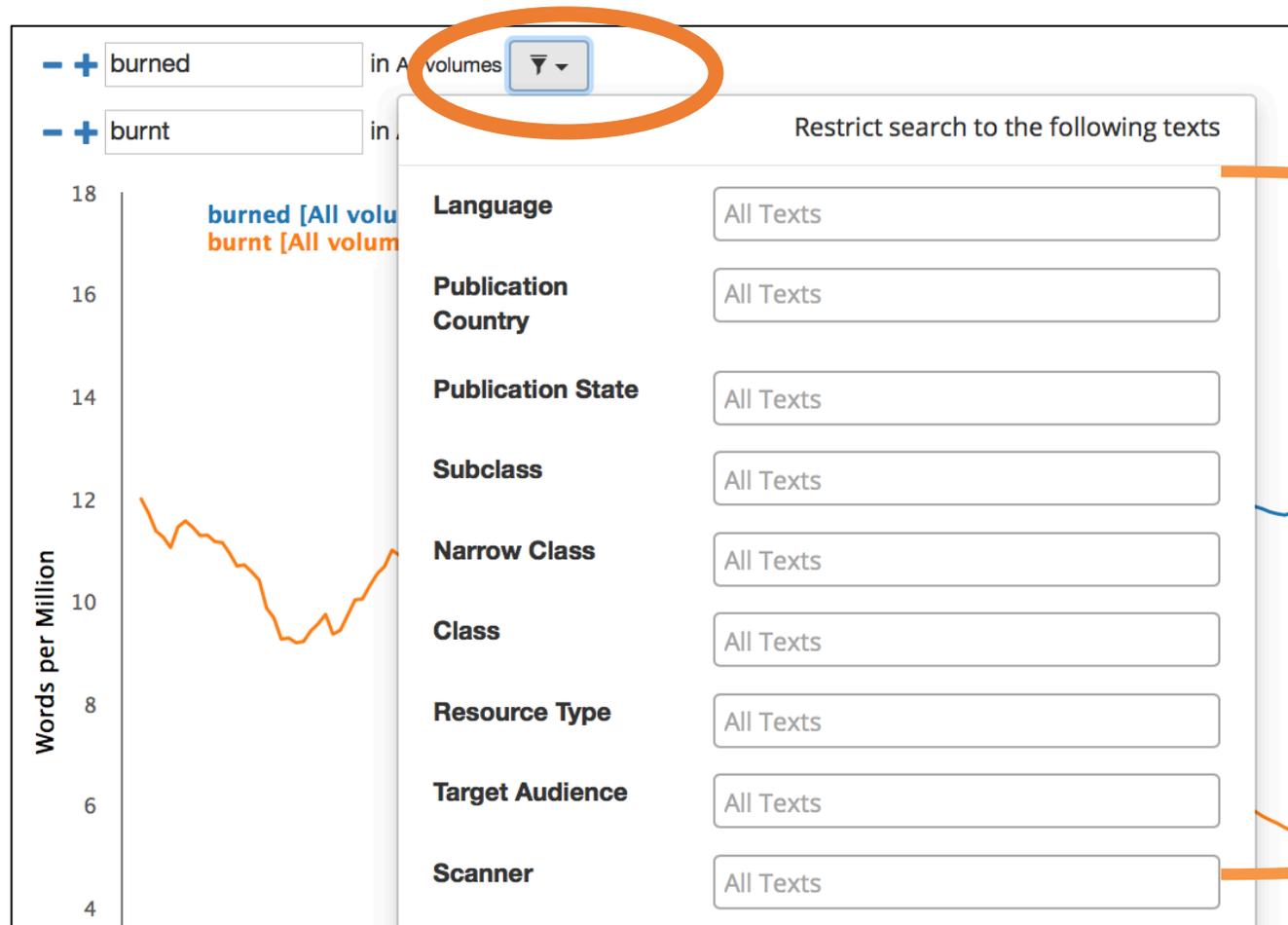
- Let's look at how verbs change over time
 - Eg. Burned vs. burnt



Do you see any trends?



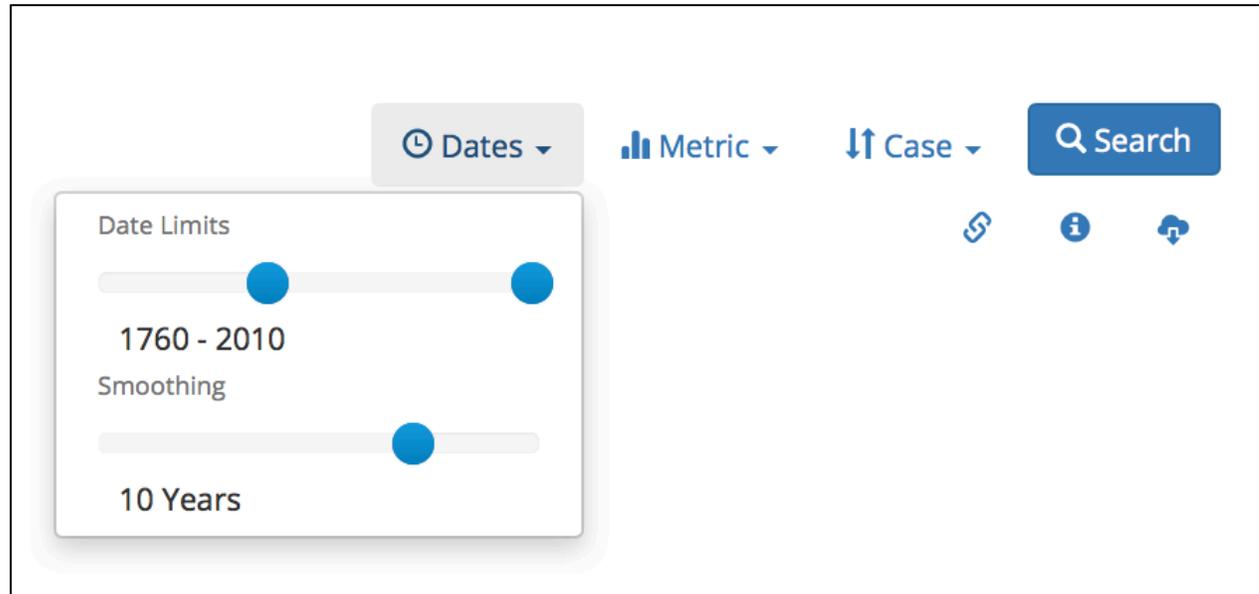
Bookworm interface



Limit
your
search
with
facets



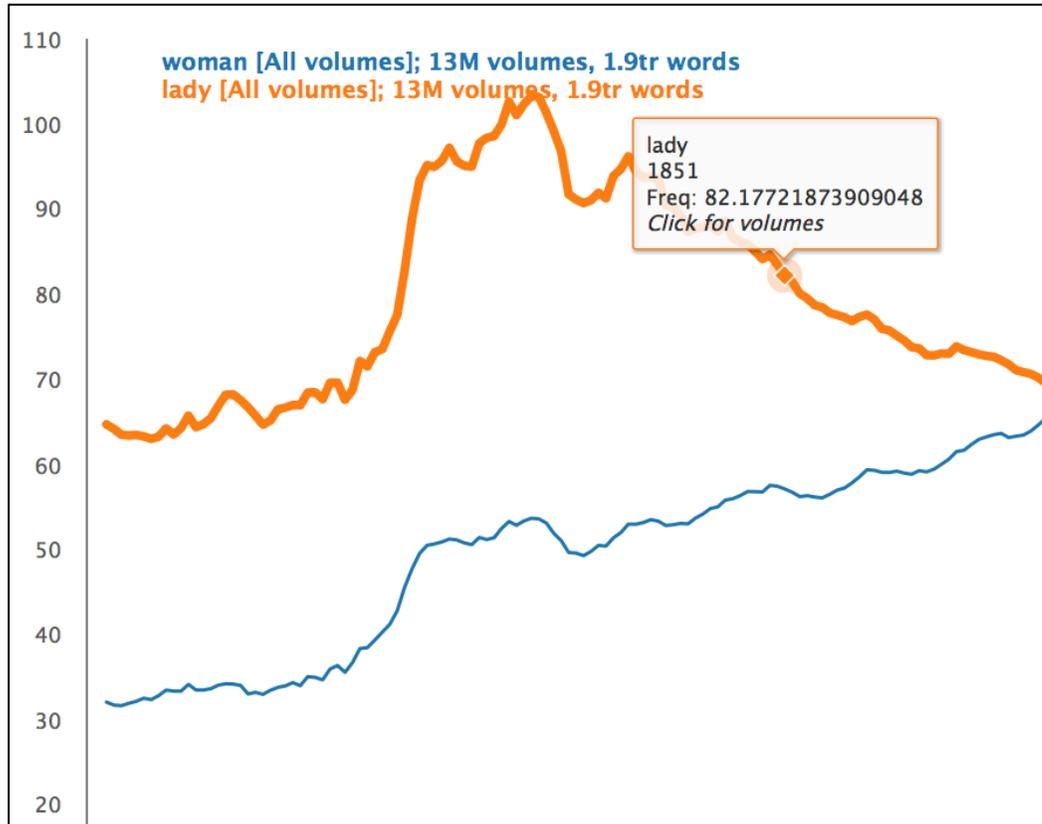
Bookworm interface



Fine-tune
your
results



Bookworm interface



About Collections Help Feedback

Search words about the items

Jump to 3 Go

Search in this text Find

8.22
C19412

LADY DI'S MINUET.

Enter Sir JOHN WILDUCK, speaking as he enters.

SIR J. Tell Lord Mulligatawney that Sir John Wilduck is in the drawing-room. Come, that's settled! I must make an end of it to-day. I don't know what to make of this Mulligatawney—a man that takes such a desperate fancy to one of a sudden, all about a shooting adventure—and will have one marry his daughter, whether one likes it or not. Every morning I come here, with my mind made up to break the thing off; but the moment Mulligatawney sees me, he rushes at me, seizes me by the hand, and calls me his “Dear Sir John—his good Sir John!” I should like to know how I'm to tell such a father as that—“Your daughter's not the thing for me; look out for another son-in-law.” Accordingly, I hesitate—I put it off to the next time. The day's gone by, and if this goes on, I shall find myself a married man before I know where I am; not that there is anything to be said against Lady Diana—she's pretty, witty, rich! Yes, by-the-bye, she has one fault—she's too short—not like my cousin Louisa, with her five feet eight. I forgot my rule—I never fall in love under my own height. How can two people step well together in harness, if one's a foot taller than the other? And then, they call it a good match. But Louisa's up to my shoulder already, and growing visibly—and the taller she grows, the better I like her; besides, our marriage is settled between the families. Well, I'm very sorry for Lady Diana, but I must tell her father to-day.

Mavis Sam

Links directly to texts in the HTDL



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:

Explore word usage trends of political concepts within the HathiTrust using HT+BW



Hands-on activity

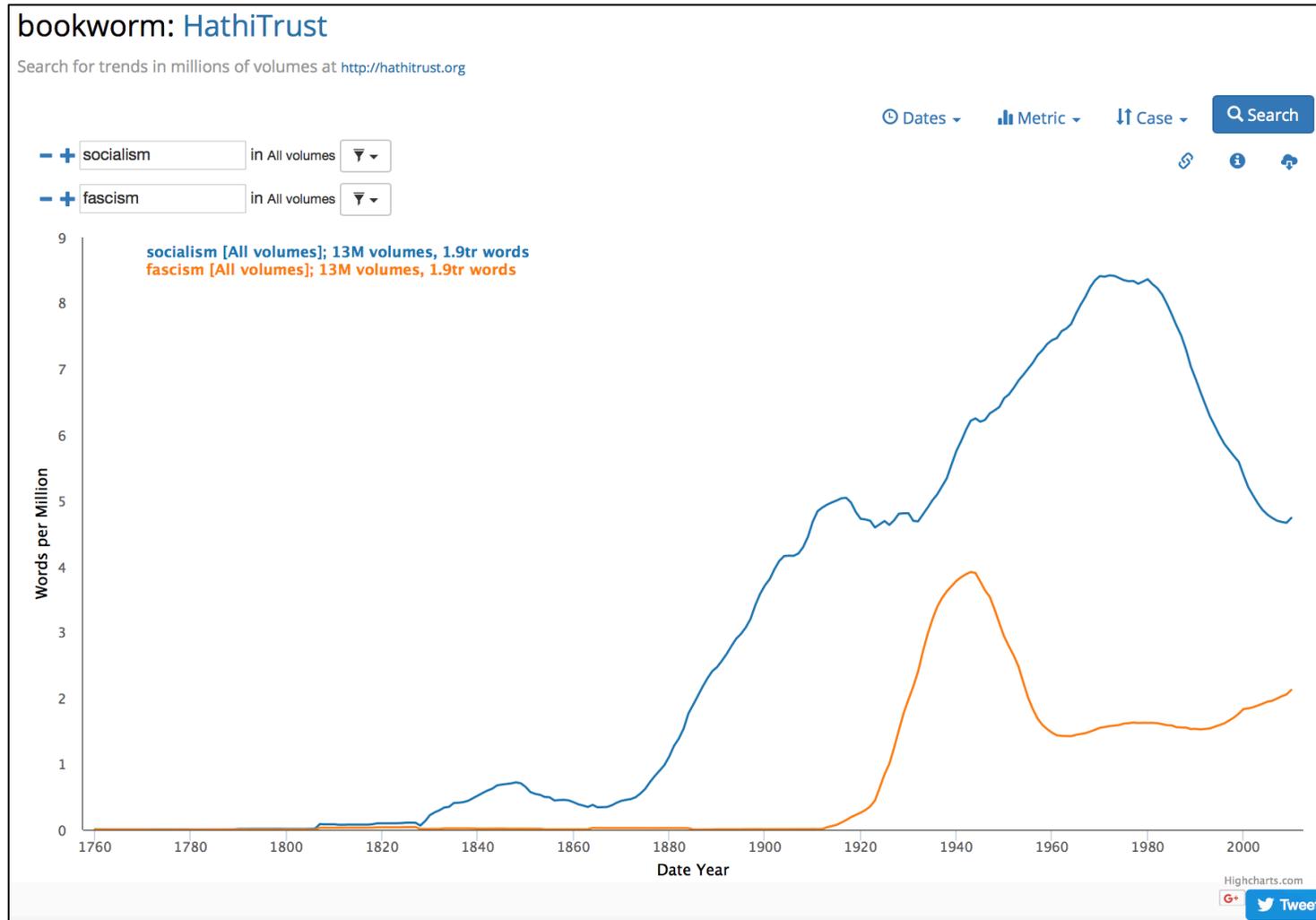
 *See Handout p. 22*

- In this activity, you will use HT+BW to explore lexical trends

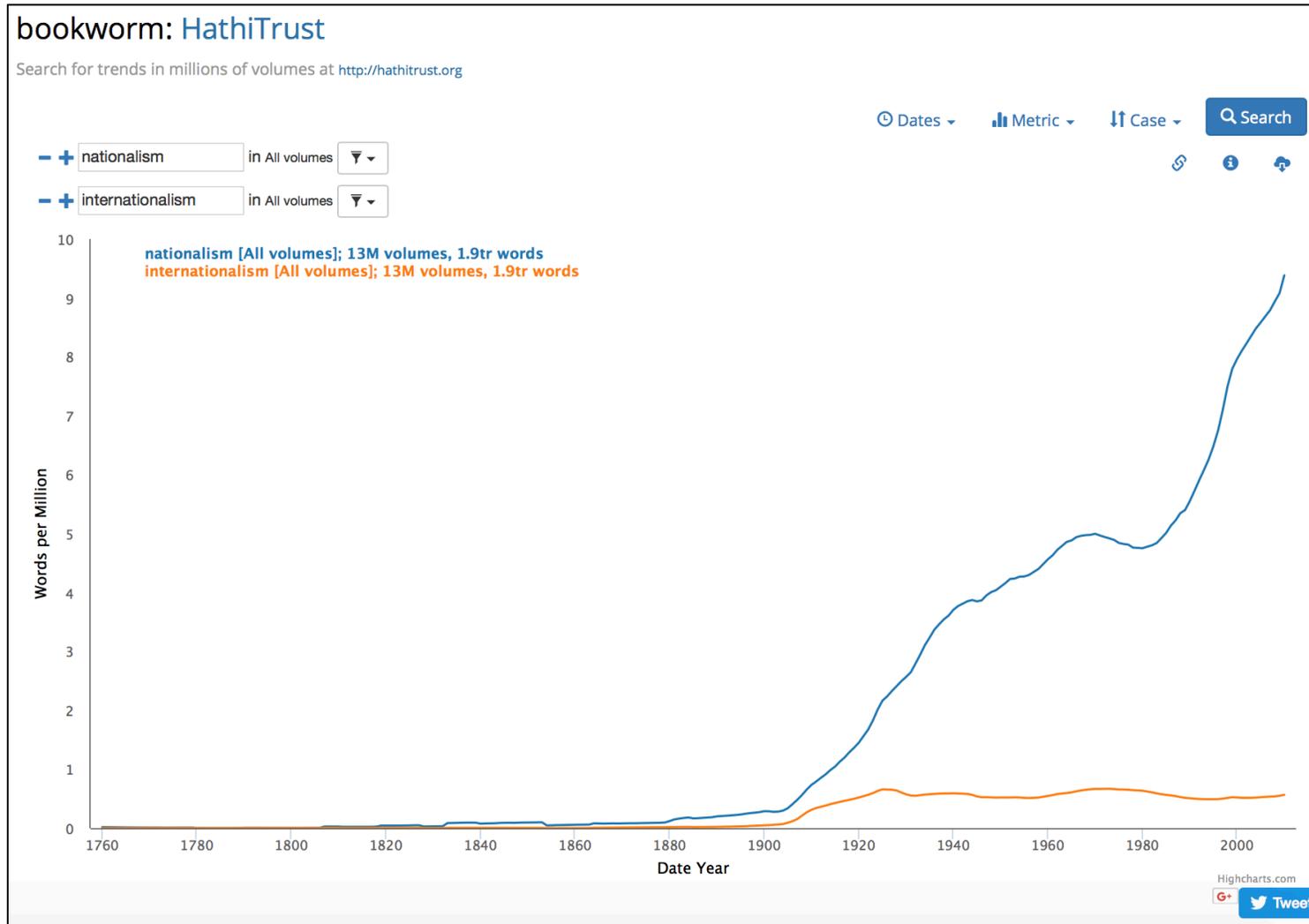
Website: <https://bookworm.htrc.illinois.edu/develop>



Examples



Examples



Discussion

- *What trends did you discover?*

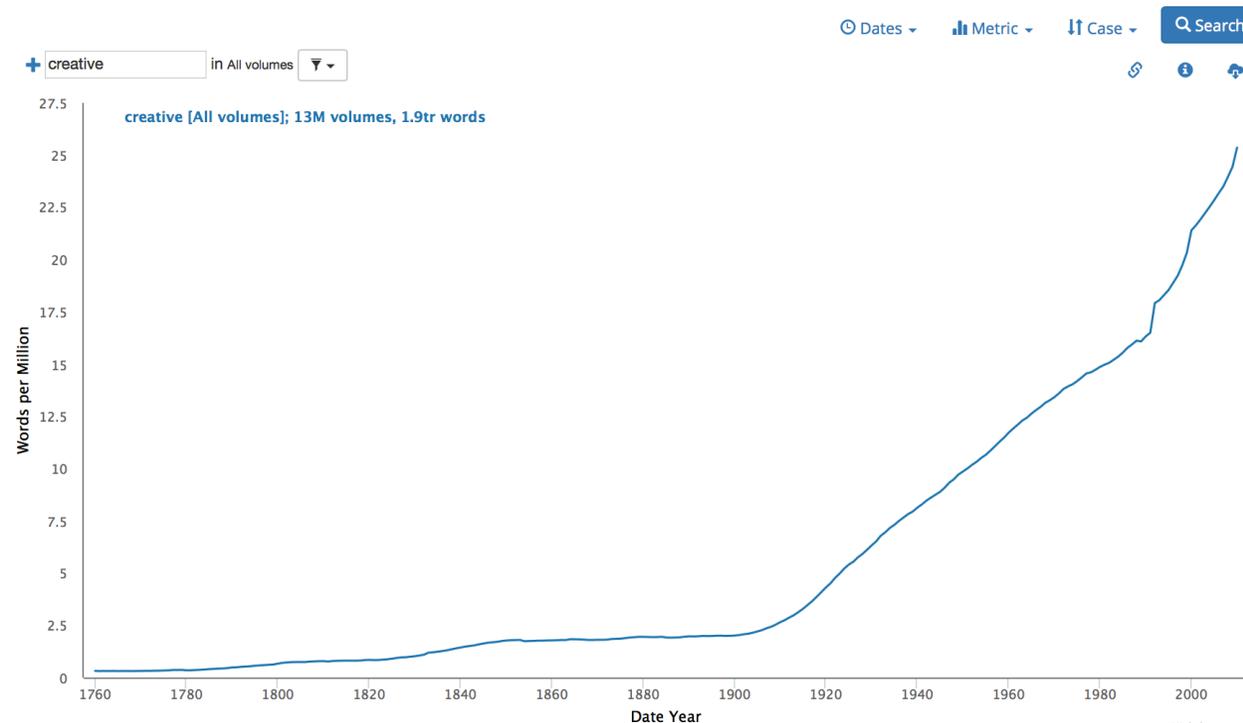


Case Study: *Inside the Creativity Boom*

- Sam used HT+Bookworm to visualize the use of “creative” in the HTDL over time

bookworm: [HathiTrust](#)

Search for trends in millions of volumes at <http://hathitrust.org>



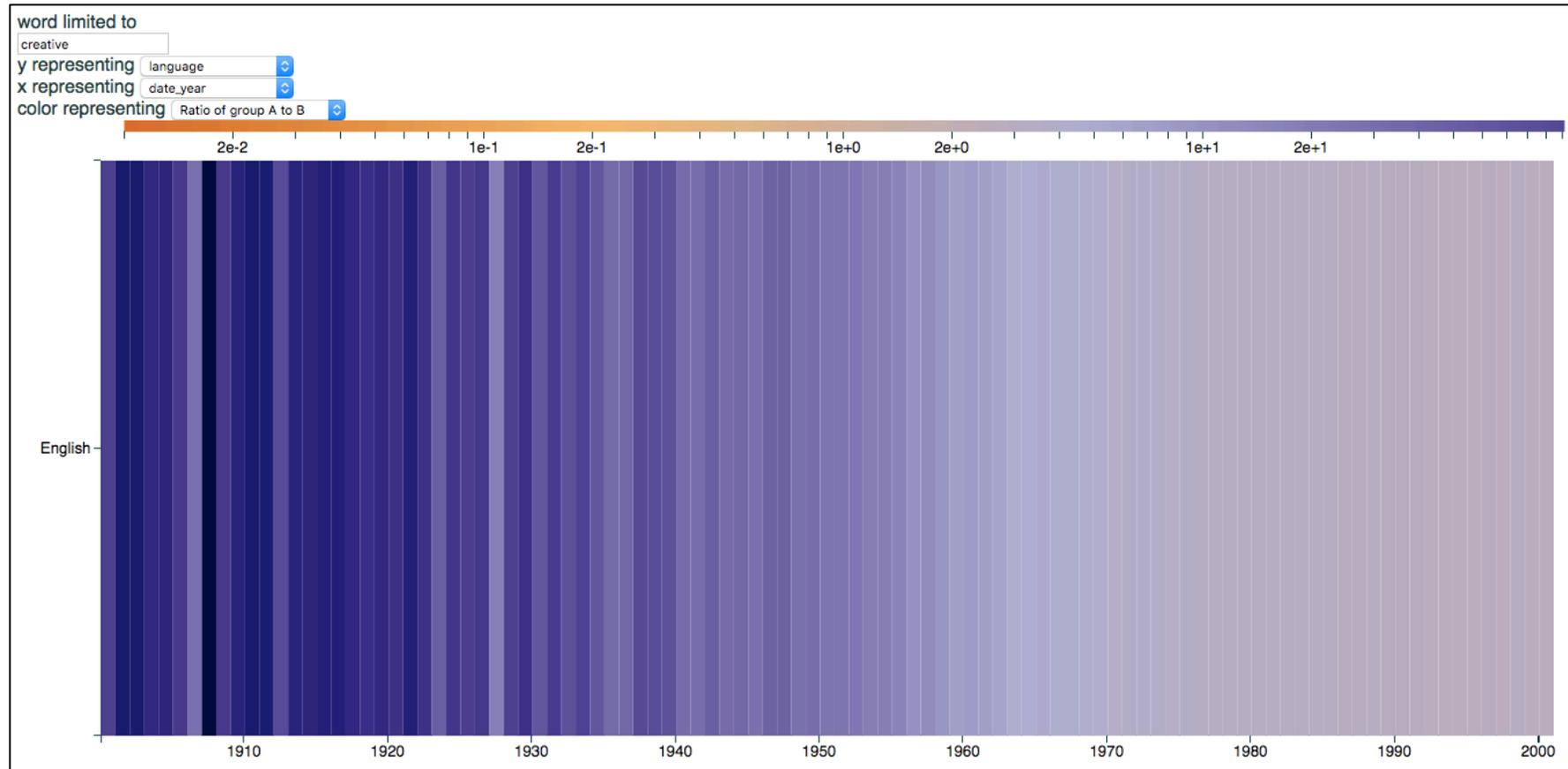
Case Study: *Inside the Creativity Boom*

- Sam also used an experimental HT+BW interface to create different kinds of visualizations...



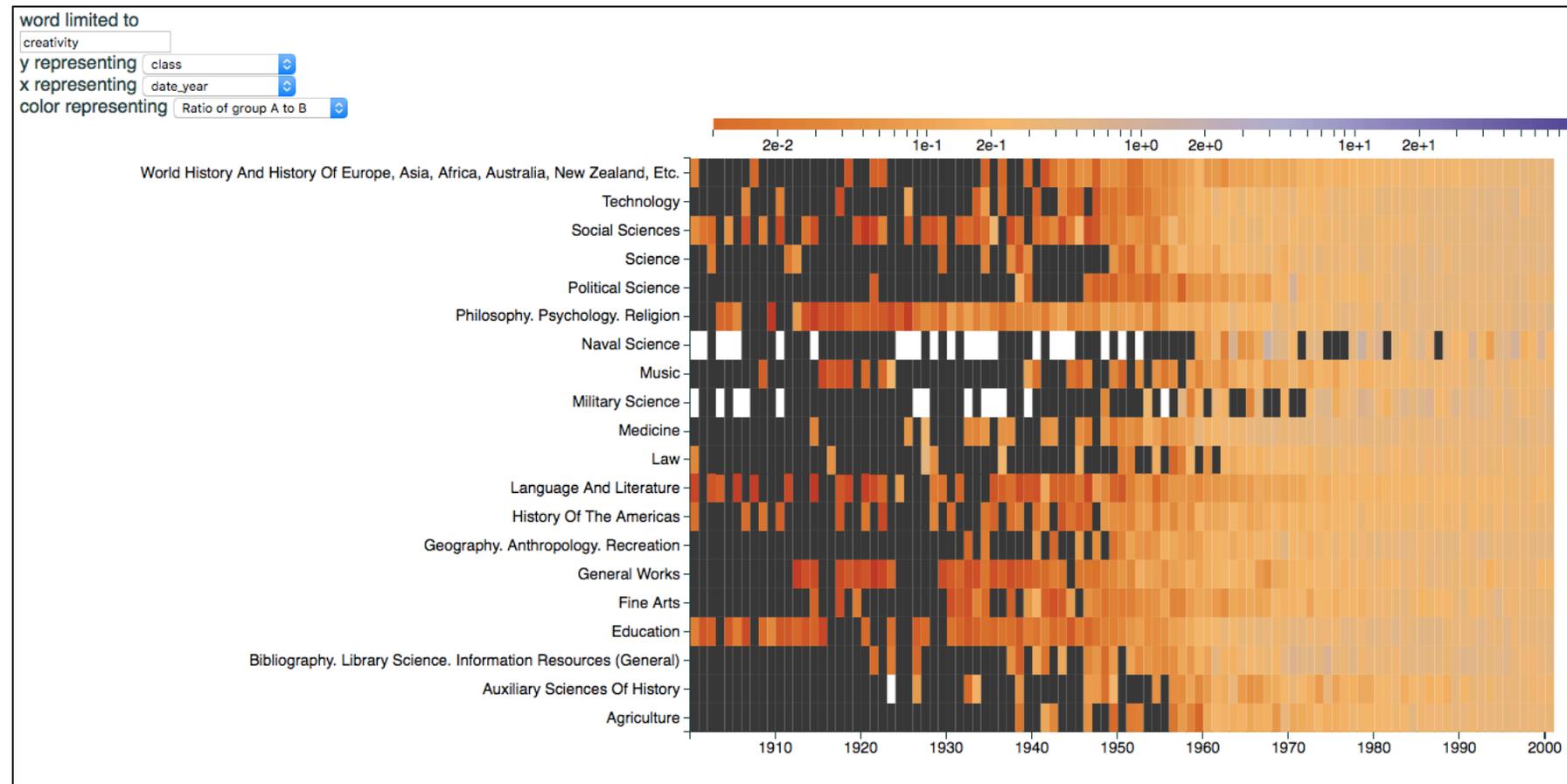
Case Study: *Inside the Creativity Boom*

- “Creative” by language and year



Case Study: *Inside the Creativity Boom*

- “Creativity” by library classification and year



Discussion

- *Where does visual literacy fit into data literacy overall?*
- *What would it mean to be visually literate, particularly with regard to text analysis?*



Questions?



References

- Chuang, J. (2011). Text Visualization. November 2011. Retrieved January 25, 2017, from <http://hci.stanford.edu/courses/cs448b/f11/lectures/CS448B-20111117-Text.pdf> .
- Palmer K., Polley T., & Pollock C. (n.d.). Chronicling Hoosier. Retrieved August 16, 2017, from <http://centerfordigschol.github.io/chroniclinghoosier/map1.html> .
- Roskey Legal Education Blog. (2011, July 15). Martin Luther King, Jr.'s "I have a dream" speech as a word tree. Retrieved August 16, 2017, from <http://roskeylegaled.com/blog/post/martin-luther-king-jr-s-i-have-a-dream-speech-as-a/> .
- Schmidt, B. (2017, May 16). A brief visual history of MARC cataloging at the Library of Congress. Retrieved August 16, 2017, from <http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html> .
- Schmidt, B. (n.d.). API Philosophy | Bookworm. Retrieved August 16, 2017, from https://bookworm-project.github.io/Docs/api_philosophy.html .



References

- Theguardian.com. (2013, February 12). The state of our union is ... dumber: How the linguistic standard of the presidential address has declined. Retrieved August 16, 2017, from <https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level>
- Underwood, T., & Bamman, D. (2016, November 28). The Gender Balance of Fiction, 1800-2007 | The Stone and the Shell. Retrieved August 16, 2017, from <https://tedunderwood.com/2016/12/28/the-gender-balance-of-fiction-1800-2007/> .
- Underwood, T. (2012, November 11). Visualizing topic models. | The Stone and the Shell. Retrieved August 16, 2017, from <https://tedunderwood.com/2012/11/11/visualizing-topic-models/> .
- Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, 14(6). [10.1109/TVCG.2008.172](https://doi.org/10.1109/TVCG.2008.172) .



Thanks to all our team members!

- **PI**

Harriett Green, [University of Illinois at Urbana-Champaign Library](#)

- **Co-PIs**

Angela Courtney, [Indiana University-Bloomington Libraries](#)

J. Stephen Downie, [Graduate School of Library and Information Science, University of Illinois](#)

Terese Heidenwolf, [Lafayette College Library](#)

Geoffrey Morse, [Northwestern University Libraries](#)

Amanda Henley, [University of North Carolina at Chapel Hill Library](#)

- **Project Members (Past and Present)**

University of Illinois at Urbana-Champaign

Eleanor Dickson Koehl

Ruohua Han

*Rachel Blomer

*Gabriella Stuardi

*Beth Sandore Namachchivaya

*Ryan Dubnicek

*Sayan Bhattacharyya

Indiana University

Leanne Nay

Robert McDonald

Nicholae Cline

*Ewa Zegler-Poleska

Lafayette College

Lijuan Xu

Ben Jahre

*Neil McElroy

University of North Carolina at Chapel Hill

Sarah Morris

*Stewart Varner

*Jacob Hill

Northwestern University

*Geoffrey Swindells

* = former members



QUESTIONS?

HTRC_workshop@library.illinois.edu

<https://teach.htrc.illinois.edu>



award #RE-00-15-0112-15

