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

In pairs or small groups, review the summarized research projects available at [http://go.illinois.edu/ddrf-research-examples](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)
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)
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

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

From paper - Figure 3: Word cloud of topic labeled “Female Fashion.”
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

X axis: Year

Genre
- Poetry, Drama, Fiction
- Nonfiction Prose
HTRC for text analysis

Digitized text
- Scanned & OCR-ed

Computational methods
- E.g. Word counts, classification, topic modeling

Analysis

HathiTrust Research Center

at scale from the digital library

provided tools and services
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 20\textsuperscript{th} century?*

**Approach:**

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

Learn more: [https://wiki.htrc.illinois.edu/x/CADiAQ](https://wiki.htrc.illinois.edu/x/CADiAQ)
A word of caution…

Workshop outline suggests research workflow like:

- Find text
- Prepare text
- Apply algorithm
- Visualize results
A word of caution…

Actual research workflow like:

Search for text

Clean text

Get access to text

Exploratory visualization

Prepare text

Apply algorithm

Visualize results
Discussion

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

  [http://people.ischool.berkeley.edu/~hearst/text-mining.html](http://people.ischool.berkeley.edu/~hearst/text-mining.html)


