Text and Data Mining in the Humanities and Social Sciences

Strategies and Tools
Peter Leonard
Director, DH Laboratory
Yale University Library
peter.leonard@yale.edu
@pleonard

Lindsay King
Assistant Director, Haas Arts Library
Yale University Library
lindsay.king@yale.edu
@mslindsayking
In this webinar:

Overview of text and data mining as a **technique and trend**

**Making sense** of the data

What does this mean for **libraries**?

*Pause for questions*

Putting these ideas into **practice**: Robots Reading Vogue

**Conclusion** and last questions
Part 1: Text & Data Mining as Trend

Predecessors & Allies

Reasons for current interest

Major Approaches
Historical Predecessors & Current Allies

Monastic Tradition - Concordances – Roberto Busa

Astronomy - Radio Telescopes

Corpus Linguistics - Computational Linguistics
Reasons for current interest

Volume of digitized source material
- Google Books / HathiTrust
- JSTOR
- Etc.

APIs for programmatic access
- HathiTrust Research Center
- JSTOR Data for Research
- Various vendor text-mining API's (Elsevier, etc.)

Advances in technique
- Algorithmic sophistication of Machine Learning
- Moore's Law & RAM prices
Major Approaches

Look for something you think is there

Let data organize itself
Bookworm

Bookworm is a simple and powerful way to visualize trends in repositories of digitized texts.
Special issue of Poetics: "Topic Models and the Cultural Sciences"

Editors’ Introduction:

"Topic models: What they are and why they matter."
John Mohr (Sociology, UCSB) and Petko Bogdanov (Computer Science, UCSB)

In this short essay we provide a brief, non-technical introduction to the text mining methodology known as topic modeling. We start with the most basic question—what is a topic model? We review the theory behind the method and then focus in on the concept of a ‘topic.’ Here we have understood and interpreted the term. Then we comment briefly on some of the data that have been analyzed and the results obtained. We then proceed to the second question: 'What can we answer by describing some of the results of the analysis?' For this, we look at the eight research papers collected for this special issue. Section 1 contains two papers that more clearly illustrate the potential of topic models to address social scientific questions. In Paper 1, "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Government Arts Funding in the U.S.,” Paul DiMaggio (Sociology, Princeton University), Manish Nag (Sociology, Princeton University), and David Blei (Computer Science, Princeton University).
Part 2: Making Sense of the Data

What makes a good TDM project?

What do you have?
Making Sense of the Data

What makes a good TDM project?

- Locally meaningful or unique datasets
- Relevant data to your researchers' specific interests
Making Sense of the Data

What do you have?

- Some approaches rely on metadata alone
- Some approaches rely on data alone
- Many approaches operate on the intersection of data and metadata
- In all cases, more is better
Approaches that rely on data, not metadata

Topic Modeling (automatic discovery of themes from uncataloged texts)

Sequence Alignment (automatic discovery of quotation and plagiarism)
Textual Data for TDM: The fine print

How *machine-actionable* are your texts?

- Handwritten?
  - Need to transcribe

- Typewritten or Printed?
  - Ready for OCR

- Already OCR’d, but a long time ago?
  - Consider re-processing with modern algorithms
Textual Data for TDM: More fine print

Did your OCR processes…

• Use language-specific heuristics?

• Re-connect words split with hyphens?

• Disregard running headers and footers?

None of these are show-stoppers, but they are good to be aware of for highest-quality results
Part 3: What does this mean for libraries?

Data Access: A Spectrum

- No access at all (vendor refuses or doesn’t have the rights)
- Vendor-controlled tools (part of their hosted platform)
- API (Application Programming Interface)
- Full Data
  - Download
  - Hard drive through the mail

Different levels of potential and effort
Part 3: What does this mean for libraries?

**Perpetual Access License** as key to getting data for TDM?
- Currently bolsters arguments for obtaining data from vendors
- Not always available, nor affordable

Think about TDM at **point of purchase/licensing** to ensure access to data
- Involve your library copyright/licensing specialists

**Secure storage** and preservation of local hard drive copies of data
Questions and comments?
Robots Reading Vogue

Full access to data: text and images

Vendor-digitized with extensive markup, rather than library-digitized

Copyrighted content and licensing restrictions on use
- Not just ProQuest but Condé Nast
- High value=more concerns about copying

N-gram searching with Bookworm (http://bookworm.culturomics.org)

Topic modeling with MALLET (http://mallet.cs.umass.edu/)

Student projects enabled by controlled access to data files
Conclusion

TDM is facilitated by huge amounts of humanities/social science data – and new algorithmic ways to process it.

Securing data access and licensing can be challenging, but yield powerful results for scholars.

Ideal TDM projects explore data that connect to your local audience.

Two approaches to big datasets:
- Searching for what might be there
- Letting the data organize itself
Questions and comments?
Miscellaneous questions from chat feed

Q: How do you communicate any restrictions on terms of use to researchers? Do you have searchers sign any kind of locally created terms of use agreement?
A: Yes, we’re working with collection development and licensing staff on an agreement for researchers to sign that reminds them of the restrictions on the content, so that everyone is aware.

Q: Do your licensing agreements at Yale allow you to give non-Yale researchers who travel to your campus access to the ProQuest results?
Sadly, we have no money at our library for that [Vogue] subscription, but sometimes our scholars can get a travel stipend (like traveling to see another university’s ‘analog’ archival collections).
A: The content links cited in the project work at any institution which has licensed the ProQuest Vogue database.

Q: Great project! Do you have shareable code?
A: For programming code for the MALLET topic modeling used in the Vogue project see The Programming Historian, at http://programminghistorian.org/lessons/topic-modeling-and-mallet
Miscellaneous comments from chat feed

Comment: We are seeing computational journalism burgeoning in this area as well.

Comment: You can buy annotated NY Times data from 1987 to 2007 from the University of Pennsylvania’s Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC2008T19-

Comment: Chronicling America has an API (and horrible OCR). Virginia Tech's Digging into Data project on the Spanish Flu Pandemic used it. http://vtechworks.lib.vt.edu/handle/10919/19271

Comment: It would be so helpful if libraries negotiating these contracts for perpetual use would keep an eye toward preserving fair uses to the data and not licensing them away.
Robots Reading Vogue

http://dh.library.yale.edu/projects/vogue/

Peter Leonard
Director, DH Laboratory
Yale University Library
peter.leonard@yale.edu
@pleonard

Lindsay King
Assistant Director, Haas Arts Library
Yale University Library
lindsay.king@yale.edu
@mslindsayking
Upcoming CRL events

**Webinar: CRL Licensing and Acquisitions 2015-2016**

*Wednesday, August 19*
*2:00 to 3:00 p.m. Central Time*

Visit [www.crl.edu/events](http://www.crl.edu/events) to register
For More Information

- Fill out our follow-up survey at http://www.surveymonkey.com/s/CRLWebinarFollowup

- This presentation will soon be available on our YouTube channel: www.youtube.com/crldotedu

- Visit the CRL website www.crl.edu

- Sign up for CRL Connect: www.crl.edu/connect

- Find CRL on Facebook and Twitter