Text-as-Data Workshop

Draft Syllabus for Morning Sessions

(see below for readings for joint lunch discussions and afternoon seminars)

Day 1: Introduction to Archives as Data

What is Digital Text?: In the first session we will introduce digital text. How does it differ from numeric text? Can the same types of methods be used for both? What is metadata and how is it used? We will get a first look at the data available through the History Lab website. We will talk about the different collections available as well how the collections are formatted.

- Topics: Introduction to Data Science; Introduction to R/Python
- Lab: Getting started with Python/R/Stata, data processing and summary
- Readings: Justin Grimmer and Brandon Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts"
  Kenneth Benoit, et. al. "Crowd-sourced text analysis: reproducible and agile production of political data"
  Reference card for R

Day 2: Formatting Text Data, Part 1

Bringing text into software: The History Lab data is available from an SQL database. We will learn how to use SQL queries to extract data from the database.

- Topics: Structured data: txt files, SQL
- Lab: Load textual data; Using different types of textual data; explore History Lab data
- Readings: Kenneth Benoit, "Text Analysis in R"

Day 3: Formatting Text Data, Part 2

Continuing the introduction to bringing text into software, we will focus on how to deal with unstructured and tagged data. The class will learn a bit about web scraping and how to parse HTML and JSON data in Python.

- Topics: Unstructured and tagged data: web scraping, json data
- Lab: Load textual data; Using different types of textual data; explore History Lab data
- Readings: BeautifulSoup, part 1: Trump statements
  BeautifulSoup, part 1: AFL-CIO

Day 4: Examining Text Data

Getting around the data: In this session, we will focus on string functions in different packages to more easily get a sense of the data. We will also learn the basic building blocks of textual analysis–the bag of words approach to represent text.

- Topics: string functions, bag of words
- Lab: Using string functions; creating bags of words

Day 5: Principles of Text Analysis

Day 5 will look at identifying different patterns in text using regular expressions. We will also explore Named Entity Recognition.

- Topics: regular expressions, NER
- Lab: regular expressions, understanding NER

Day 6: Introduction to Quanteda

There are several packages available in R to analyze text data. Day 6 will be an introduction to using the Quanteda package.

- Topics: Introduction to Quanteda
- Lab: Bringing text data into Quanteda
- Readings: Quanteda Quickstart

Day 7: Basic Textual Analysis

In this session we will explore techniques which can give a better sense of the information within the text, including word clouds, keywords in context, and frequency tables.

- Topics: Summary statistics, graphs
- Lab: Key words in context; wordclouds

Day 8: Data Coding & Unsupervised Learning

In this session, we will examine the principles of coding data. We will also focus on how to use unsupervised learning to classify text into different groups.

- Topics: Quantitative versus qualitative coding, k-means clustering

Day 9: Supervised Learning

Day 9 will focus on classifying text using supervised learning. After giving an algorithm training data, we can apply the model to unseen data to group data into different categories.

- Topics: Regression and classification
- Lab: Logit/probit; Naive Bayesian classifiers
- Readings: Replicating Federalist Paper Author attribution
  Vincent D'Orazio, et. al. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines"
  Kosuke Imai “Discovery”. Chapter 5 of *Quantitative Social Science*. (You can skip the section on mapping.)

Day 10: Advanced Topics

In the final session, we will explore topic models and social network analysis

- Topics: Social network analysis, topic models
- Lab: Network analysis, running topic models
- Readings: John Patty and Elizabeth Penn "Network Theory and Political Science"
  Christopher Lucas, et. al. "Computer-Assisted Text Analysis for Comparative Politics"

Archiving Digital Records Workshop

Draft Syllabus for Morning Sessions
(see below for list of speakers and readings for joint lunch discussions and afternoon seminars)

Day 1: Introduction to PDFs

Introductions and objectives. Course syllabus and toolkit overview. PDF History. PDF vs. PDF/A. Basic PDF structure. Introduction to the command line.

- Labs:
  - Make your PDF from scratch
  - Introduction to the command line
- Readings: A Textual and Cultural Assessment of the Mueller Report, Duff Johnson; PDF Specifications, ISO & Adobe

Day 2: PDF Processing

Available PDF metadata elements. How to extract, add, update, and delete metadata. Extracting text from a PDF and measuring text quality. Automated extraction of item-level metadata from a PDF. Metadata and users.

- Labs:
  - Extracting, Adding, Deleting, and Updating PDF Metadata
  - PDF Text Extraction Case Study
  - Identifying Item Level Metadata from a PDF

Day 3: OCR Processing

What is OCR, and when do you need it? Steps in the OCR process. Using Tesseract. Implementing an OCR workflow and improving OCR results.

- Labs:
  - Extracting text from a scanned PDF
  - Improving OCR results

Day 4: Named Entity Recognition & Text Analysis

Named entity recognition (NER) fundamentals. Referencing people, organizations, governments, and places.

- Labs:
    - Applying NER to a PDF
    - Introduction to Voyant

Day 5: Utilizing New Metadata. Recap.

Adding metadata to repository systems and finding aids. Applying what you’ve learned.

- Lab: Applying the toolkit to your collection

Lunchtime Talks and Afternoon Seminar Readings

The exact order of the lunchtime talks will depend on when we can schedule speakers. Afternoon seminar readings will be organized to support and expand on the lunch discussions, and speakers will also be invited to suggest readings. The following texts address the core concerns of the Institute:

- Connelly, Matthew. "State Secrecy, Archival Negligence, and the End of History as We Know It." Knight First Amendment Institute. September 13, 2013, accessed February 23, 2022.

- Drake, Jarrett. "Liberatory Archives: towards Belonging and Believing (Part 1 and 2)." Medium. October 22, 2016, accessed February 23, 2022.

- Handel, Dinah, and Mark Matienzo, *Facilitating and Illuminating Emergent Futures for Archival Discovery and Delivery: The Final Report of the Lighting the Way Project*. Stanford, CA: Stanford University Libraries, 2021.

- Hintz, Carrie, and Sarah Quigley, "A Matter of Trust: Practical Strategies for Writing User Centered, Values-Driven Description." *Journal of Archival Organization* (2021).

- Hitchcock, Tim. "Confronting the Digital, or How Academic History Writing Lost the Plot." *Cultural and Social History* 10, no. 1 (March 2013): 9-23, https://doi.org/10.2752/147800413X13515292098070.

- Hughes-Watkins, Lae'l. "Moving Toward a Reparative Archive: A Roadmap for Holistic Approach to Disrupting Homogenous Histories in Academic Repositories and Creating Inclusive Spaces for Marginalized Voices." *Journal of Contemporary Archival Studies* 5, no. 1 (2018).

- Jules, Bergis. "Confronting Our Failure of Care Around the Legacies of Marginalized People in the Archives." Medium. November 11, 2016, accessed February 23, 2022.

- Jules, Bergis, Ed Summers, and Vernon Mitchell, Jr. "Ethical Considerations for Archiving Social MediaGenerated by Contemporary Social Movements." White Paper. *Documenting The Now*, April 2013.

- Matienzo, Mark. "To Hell with Good Intentions: Linked Data, Community and the Power to Name." Medium. February 11, 2016, accessed February 23, 2022.

- Miller, Larisa. "All Text Considered: A Perspective on Mass Digitizing and Archival Processing."

The American Archivist 76, no. 2 (2016): 521–541, https://doi.org/10.17723/aarc.76.2.6q005254035w2076.

- Moravec, Michelle. This is not a Digital Archive: How Digitized* Changed Historical Research. Medium. August 23, 2016, accessed February 23, 2022.

- Moss, Michael. *Is Digital Different? How Information Creation, Capture, Preservation and Discovery Are Being Transformed.* London: Facet Publishing, 2015. Introduction and Chapter One.

- Nowatzki, Robert. "From Datum to Databases: Digital Humanities, Slavery, and Archival Reparations," *The American Archivist* 83, no. 2 (2016): 429–448, https://doi.org/10.17723/0360-9081-83.2.429.

- Owens, Trevor. "Defining Data for Humanists: Text, Artifact, Information or Evidence?" *Journal of Digital Humanities* 1, no. 1 (Winter 2011).

- Pozen, David. "Crisis in the Archives: State Secrecy, Archival Negligence, and the End of History as We Know It." Knight First Amendment Institute. September 13, 2018, accessed February 23, 2022.

- Putnam, Lara. "The Transnational and the Text Searchable: Digitized Sources and the Shadows They Cast." *The American Historical Review* 121, no. 2 (2016): 377–402.

- Theimer, Kate. "Archives in Context and as Context." *Journal of Digital Humanities* 1, no. 2 (Spring 2012).