**Digging Deeper Reaching Further**

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

**Resources and Readings**

**Bibliography**

**MODULE 1**

* Hearst, M. (2003). What is text mining. SIMS, UC Berkeley. <http://people.ischool.berkeley.edu/~hearst/text-mining.html>
* Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. Poetics, 41(6), 750-769. <http://dx.doi.org/10.1016/j.poetic.2013.08.005>
* Juola, P. (2017). 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/>

**MODULE 2.1**

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

**MODULE 2.2**

* 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/>

**MODULE 3**

* 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 Humanties. (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/llc/18.2.209>

**MODULE 4.1**

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

**MODULE 4.2**

* 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/>

**MODULE 5**

* 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://roskylegaled.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>
* 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)

**Further Reading and Resources**

**SUPPORTING DIGITAL SCHOLARSHIP**

* Auckland, M. (2012). Re-skilling for research: An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers. *RLUK Report,*available at: <http://www.rluk.ac.uk/files/RLUK%20Re-skilling.pdf>
* Ayers, E. L. (2013). Does digital scholarship have a future?. *Educause Review*, *48*(4), 24-34. <https://er.educause.edu/articles/2013/8/does-digital-scholarship-have-a-future>
* Babeu, A. (2011). *” Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classics*. Washington, DC: Council on Library and Information Resources. Retrieved October 3, 2017 from <https://www.ianus-fdz.de/attachments/339/Babeu_Rome-Wasnt-Digitized-in-a-Day_2011.pdf>
* Bryson, T., Posner, M., Pierre, A. S., & Varner, S. (2011). *SPEC kit 326: Digital humanities.* Washington, DC: Association of Research Libraries. Retrieved October 3, 2017 from <http://publications.arl.org/Digital-Humanities-SPEC-Kit-326/>
* Johnson, L., Adams Becker, S., Estrada, V. & Freeman, A. (2015). *NMC Horizon Report: 2015 Library Edition*. Austin, TX: The New Media Consortium. Retrieved October 3, 2017 from <https://www.learntechlib.org/p/151822/>.
* Lippincott, J., & Goldenberg-Hart, D. (2014). Digital scholarship centers: Trends & good practice (CNI workshop report). [*https://www.cni.org/wp-content/uploads/2014/11/CNI-Digitial-Schol.-Centers-report-2014.web\_.pdf*](https://www.cni.org/wp-content/uploads/2014/11/CNI-Digitial-Schol.-Centers-report-2014.web_.pdf)
* Maron, N. L. (2015). The digital humanities are alive and well and blooming: Now what?. *Educause Review*, *50*(5), 28-38. <https://er.educause.edu/articles/2015/8/the-digital-humanities-are-alive-and-well-and-blooming-now-what>
* McDonald, D., McNicoll, I., Weir, G., Reimer, T., Redfearn, J., Jacobs, N., & Bruce, R. (2012). The Value and Benefits of Text Mining. *JISC Digital Infrastructure*. Retrieved from <http://www.jisc.ac.uk/media/documents/publications/reports/2012/value-text-mining.pdf>
* Palmer, C. L., & Neumann, L. J. (2002). The information work of interdisciplinary humanities scholars: Exploration and translation. *The Library Quarterly*, *72*(1), 85-117. <https://doi.org/10.1086/603337>
* Searle, S. (2015). Using scenarios in introductory research data management workshops for library staff. *D-Lib Magazine*, *21*(11/12). <http://www.dlib.org/dlib/november15/searle/11searle.html>
* Sukovic, S. (2011). E-Texts in research projects in the humanities. In *Advances in Librarianship* (Vol. 33, pp. 131–202). Emerald Group Publishing Limited. <https://doi.org/10.1108/S0065-2830(2011)0000033009>
* Sula, C. A. (2013). Digital humanities and libraries: A conceptual model. *Journal of Library Administration*, *53*(1), 10-26. <http://dx.doi.org/10.1080/01930826.2013.756680>
* Toms, E. G., & O’Brien, H. L. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, *64*(1), 102-130. <https://doi.org/10.1108/00220410810844178>
* Vinopal, J., & McCormick, M. (2013). Supporting digital scholarship in research libraries: Scalability and sustainability. *Journal of Library Administration*, *53*(1), 27-42. <http://dx.doi.org/10.1080/01930826.2013.756689>
* Walters, T., & Skinner, K. (2011). *New Roles for New Times: Digital Curation for Preservation*. Washington, DC: Association of Research Libraries. Retrieved October 3, 2017 from <http://files.eric.ed.gov/fulltext/ED527702.pdf>.
* Zorich, D. (2012). *Transitioning to a Digital World: Art History, Its Research Centers, and Digital Scholarship. A Report to the Samuel H. Kress Foundation and the Roy Rosenzweig Center for History and New Media, George Mason University*. Retrieved October 3, 2017 from <http://www.kressfoundation.org/uploadedFiles/Sponsored_Research/Research/Zorich_TransitioningDigitalWorld.pdf>

**HT, THE HTDL, AND THE HTRC**

* Downie, S. J., Furlough, M., McDonald, R. H., Namachchivaya, B., Plale, B. A., & Unsworth, J. (2016). The HathiTrust Research Center: Exploring the full-text frontier. *Educause Review*, *51*(3), 50-51. <http://er.educause.edu/~/media/files/articles/2016/5/erm1638.pdf>
* HTRC “About” page: <https://www.hathitrust.org/htrc_about>
* HathiTrust Research Center Documentation: <https://wiki.htrc.illinois.edu/display/COM/HathiTrust+Research+Center+Documentation>
* Jett, J. et al., (2016). The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. Journal of Open Humanities Data. 2, p.e1. <http://doi.org/10.5334/johd.3>
* York, J., & Schottlaender, B. E. (2014). The Universal Library Is Us: Library Work at Scale in HathiTrust. *Educause Review*, *49(3), 48-49*. <http://er.educause.edu/articles/2014/5/the-universal-library-is-us-library-work-at-scale-in-hathitrust>

**OTHER TEXT ANALYSIS EXAMPLES**

* Digging For Nuggets Of Wisdom – The New York Times. October 10, 2003. Retrieved January 25, 2017, from <http://www.nytimes.com/2003/10/16/technology/digging-for-nuggets-of-wisdom.html>
* Lancashire, I., & Hirst, G. (2009). Vocabulary changes in Agatha Christie’s mysteries as an indication of dementia: A case study. In *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice,*8-10.  <ftp://ftp.cs.toronto.edu/pub/gh/Lancashire+Hirst-extabs-2009.pdf>

**BASH COMMANDS**

* Introduction to Bash: <http://programminghistorian.org/lessons/intro-to-bash>
* Curl and wget: <https://daniel.haxx.se/docs/curl-vs-wget.html>

**PYTHON**

* Official Python FAQ: <https://docs.python.org/3/faq/index.html>
* List of Python beginner’s guides for non-programmers: <https://wiki.python.org/moin/BeginnersGuide/NonProgrammers>

**DATA VISUALIZATION**

**General:**

* Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.
* Steele, J., & Iliinsky, N. (2010). *Beautiful visualization: looking at data through the eyes of experts*. ” O’Reilly Media, Inc.
* Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. Indianapolis, IN: Wiley Pub.
* The Data Visualization Catalogue developed by Severino Ribecca: <http://www.datavizcatalogue.com/index.html>
* Introduction to Data Visualization: Visualization Types: <http://guides.library.duke.edu/datavis/vis_types>

**Culturomics:**

* Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713-716.
* Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

**Tag clouds:**

* Waldner, M., Schrammel, J., Klein, M., Kristjánsdóttir, K., Unger, D., & Tscheligi, M. (2013, May). FacetClouds: exploring tag clouds for multi-dimensional data. In *Proceedings of Graphics Interface 2013* (pp. 17-24). Canadian Information Processing Society.

**Data visualization examples:**

* Visualizing Data: [http://www.visualisingdata.com](http://www.visualisingdata.com/)
* FlowingData: [http://flowingdata.com](http://flowingdata.com/)
* Information is Beautiful: <http://www.informationisbeautiful.net/visualizations>
* Text Visualization Browser: [http://textvis.lnu.se](http://textvis.lnu.se/)

**DATA CURATION AND MANAGEMENT**

* DH Curation Guide: [http://guide.dhcuration.org](http://guide.dhcuration.org/)
* Digital Curation Centre: <http://www.dcc.ac.uk/>
* Research Data Alliance: [https://www.rd-alliance.org](https://www.rd-alliance.org/)
* Research Data and Preservation symposium (RDAP) 2011 Summer Humanities Data Curation Summit, Muñoz and Renear, “Issues in Humanities Data Curation” discussion paper: <http://cirssweb.lis.illinois.edu/paloalto/whitepaper/premeeting/>
* ACLS, Our Cultural Commonwealth, The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences (2006): <http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf>
* Data life cycle: <http://data.library.virginia.edu/data-management/lifecycle/>
* Examples of Data Management Plans from previous successful NEH grant applications: <https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available>

**DATA COLLECTIONS AND TOOLS**

**Collecting data:**

* Text and data mining at MIT (a guide for MIT affiliates on rights and restrictions for using licensed resources for text and data mining): <https://libraries.mit.edu/scholarly/publishing/text-and-data-mining-at-mit/>
* JSTOR Data for Research: [http://dfr.jstor.org](http://dfr.jstor.org/)
* DocSouth Data: <http://docsouth.unc.edu/docsouthdata/>
* Folger Digital Texts: <http://www.folgerdigitaltexts.org/download/>
* Internet Archive: <https://archive.org/index.php>
* Twitter API Overview*:*<https://dev.twitter.com/overview/api>
* Ethical use of social media data: Research Ethics for Students & Teachers: Social Media in the Classroom:  this handout was created by the Digital Alchemists and collaborators and produced by The Center for Solutions to Online Violence (CSOV). <http://femtechnet.org/wp-content/uploads/2016/06/Research-Ethics-For-Students-Teachers_Social-Media-in-the-Classroom_DA-CSOV_2016-1.pdf>
* Bailey, M. (2015). # transform (ing) DH Writing and Research: An Autoethnography of Digital Humanities and Feminist Ethics. *DHQ: Digital Humanities Quarterly,* 9(2). http://www.digitalhumanities.org/dhq/vol/9/2/000209/000209.html

**Preparing data:**

* OpenRefine: [http://openrefine.org](http://openrefine.org/)

**Analyzing data:**

* Voyant: [http://voyant-tools.org](http://voyant-tools.org/)
* Lexos: <http://lexos.wheatoncollege.edu/upload>
* AntConc: <http://www.laurenceanthony.net/software/antconc/>
* Weka: <http://www.cs.waikato.ac.nz/ml/weka/>
* Mallet: [http://mallet.cs.umass.edu](http://mallet.cs.umass.edu/)
* HTRC Algorithm: <https://analytics.hathitrust.org/statisticalalgorithms>

**Visualizing data:**

* Voyant: [http://voyant-tools.org](http://voyant-tools.org/)
* Wordle: [http://www.wordle.net](http://www.wordle.net/)
* ArcGIS Online/StoryMaps: <https://storymaps.arcgis.com/en/>
* Google Ngram Viewer: <https://books.google.com/ngrams>
* HathiTrust+Bookworm: <https://bookworm.htrc.illinois.edu/develop/>
* Tableau: [https://www.tableau.com](https://www.tableau.com/)
* Gephi: [https://gephi.org](https://gephi.org/)
* NodeXL: <http://www.smrfoundation.org/nodexl/>
* DH Press: [http://dhpress.org](http://dhpress.org/)

**Managing and sharing data:**

* Figshare: [https://figshare.com](https://figshare.com/)
* Github: [https://github.com](https://github.com/)
* Jupyter Notebook*:*[http://jupyter.org](http://jupyter.org/)
* Journal of Open Humanities Data: [http://openhumanitiesdata.metajnl.com](http://openhumanitiesdata.metajnl.com/)

**PROJECTS AND INITIATIVES SIMILAR TO DDRF**

* Data Carpentry: [http://www.datacarpentry.org](http://www.datacarpentry.org/)
* Rochester DH Institute for Mid-Career Librarians: <http://humanities.lib.rochester.edu/institute/>
* U Mass Data Management Lessons: <http://library.umassmed.edu/necdmc/index>
* Data Carpentry: <http://www.datacarpentry.org/>
* Software Carpentry:  <http://software-carpentry.org/>
* Library Carpentry: <https://github.com/data-lessons>
* DataCamp: <https://www.datacamp.com/courses>