Syllabus: Programming for Cultural Data Analysis

University of Texas at Austin School of Information

Spring 2018, INF 315E

Thursday 3–6 p.m., UTA 1.210A

Instructor: Stephen McLaughlin

Office hours: Thursday 1–3 p.m., UTA 5.558

Course Objectives

Prerequsites: previous coursework in the humanities; no or very little programming experience preferred

In the data, information, knowledge, wisdom (DIKW) hierarchy that circulates through Knowledge Management (KM) and Information Science (IS) discussions, data appears at the base of a pyramid of which wisdom is the pinnacle. In this schematic, data is "raw" and lacking in meaning, while information, the next higher level of the pyramid—just below knowledge and then wisdom—represents the presence of added links and relationships; information is higher up on the wisdom chain because it is data made meaningful. In the humanities, students are taught that data is not found in the "raw" but has rather been cooked all along, taken and constructed and seasoned according to our situated contexts including access issues (Where is the data?); media, format, and technology constraints (How is the data?); and perspectives (What is the data? Who is involved in and impacted by its creation and use?).

Learning to think critically about data as information means rejecting common illusions about data more generally, including its objectivity, impersonality, atemporality, and authorlessness. To teach students to think about information from this more critical perspective means first understanding how a culture tends to understand what is informative.

Towards these ends, this course takes on "data wrangling" in the context of humanist perspectives.

Learning goals:

- Exploration of cultural implications of large-scale preservation of cultural materials.
- Writing using perspectives in critical data studies;
- Familiarity with scripting-style programming in Python and Unix-like systems, emphasizing literacy in finding and using free and open source software; techniques for collecting, transforming, and analyzing media and metadata available on the Web; with commonly used data models and their standard formats, including CSV, JSON, and

XML; with text analysis techniques such as natural language processing (NLP), sentiment analysis, and machine learning classification; and with tools for analyzing cultural data via visualization and statistical tests, emphasizing critical reflection on the limitations of these approaches.

Course Principles

- Writing critically about data requires both a level of knowldege about data and data wrangling, as it requires a level of knowledge about thinking and writing from critical perspectives learned in cultural studies. While this course does not *teach* cultural studies, an understanding of and experience in humanities theory and research and the principles of cultural studies are essential.
- Imitating and modifying others' code is essential in learning to program. You can find many examples and explanations on Stack Overflow and similar online forums. Taking one or two lines without attribution is OK; if you use a longer chunk of code found online, add a #comment with the source's URL.
- Begin assignments early. If you realize what you had in mind is more difficult than expected, talk to the instructor about choosing an alternative.
- We'll be focusing on a scripting approach to programming. This course is not oriented toward developing large, complex programs or writing perfectly optimized code.
- Learning to code takes trial and error. Work through weekly programming tutorials before class, and continue experimenting with in-class material at home.

Assignments

Final Project: Critical Data Analysis (50%)

For your final project, you will use a dataset drawn from online sources and analyze the data in a critical essay. You may either present an argument about the data (e.g., describing bias in the way the data were chosen and arranged) or you may use your dataset as the basis for an argument about culture (e.g., tracing a stylistic shift in a literary community). You should conceive and execute your project with a specific audience in mind, such as literary scholars, newspaper readers, or policy advocates.

Your dataset should comprise at least 200 texts or other media files, or at least 2000 metadata records. The size of your collection should be appropriate to your technical skills and the complexity of each record. Rather than using an entire pre-existing dataset, you may choose to extend or limit the dataset in some way. This might mean curating material from multiple sources, mashing up two or more datasets, augmenting records using machine learning or natural language processing, or using a creative technique to organize messy data.

Your final project will include the following elements:

- Proposal (7%)

- Proposal Peer Review (3%)
- In-class presentation (week 14) (10%)
- 10 page critical essay, with an appendix of 3–4 data visualizations (30%)

Weekly Assignments (WA) (50%)

Except when indicated, there will be required readings each week. The required readings will either be available online and linked below or posted on Canvas, so there are no books to buy or papers to acquire for the class.

Assignments should be posted on Canvas by noon on the day of class.

Week 1 (1/18): Introductions & Command Line Basics

Week 2 (1/25): Python Intro and Text Manipulation

Readings

Readings in Canvas

- Nick Montfort (2016) “Why Program?” In *Exploratory Programming for the Arts and Humanities*, 267–77. Cambridge, MA: The MIT Press.
- danah boyd & Kate Crawford (2012) "Critical Questions for Big Data," *Information, Communication & Society*, 15:5, 662-679.

Optional

- Oualline, Steve. “The End of Line Puzzle.” *The Practical Programmer*. http://www.oualline.com/practical.programmer/eol.html
- Piper, Andrew. "There will be Numbers." *Journal of Cultural Analytics* 1, no. 1 (May 23, 2016). http://culturalanalytics.org/2016/05/there-will-be-numbers/
- Stephenson, Neal. “In the Beginning Was the Command Line.” Cryptonomicon, 1999. http://www.cryptonomicon.com/beginning.html. TXT.
- Neff, Gina, Tanweer, Anissa, Fiore-Gartland, Brittany, Osburn, Laura Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data* 5, no. 2, 2017.

Assignment

Weekly Assignment #1

Week 3 (2/1): Collections as Data: Meaning Making

Readings

Canvas

- Posner, Miriam. "Humanities Data: A Necessary Contradiction." *Miriam Posner's Blog*, June 25, 2015. http://miriamposner.com/blog/humanities-data-a-necessary-contradiction
- "The Jupyter Notebook." http://jupyter-notebook.readthedocs.io/en/latest/notebook.html

Read pages 1–28 of Shieber's Python tutorial and work through the code examples:

- Shieber, Stuart M., *Programming for Humanists* pages 1–28, 2014. http://blogs.harvard.edu/programmingforhumanists/files/2014/12/proghum.pdf

Optional

- Padilla, Thomas. "Humanities Data in the Library: Integrity, Form, Access.".
- Allardice, Simon. "Foundations of Programming: Fundamentals, parts 1-3; part 5, just "part 5, Breaking your code apart"; and part 14, just "Python" and "Libraries and frameworks". http://www.lynda.com/JavaScript-tutorials/Foundations-of-Programming-Fundamentals/83603-2.html [To access Lynda.com. follow links below, click "Log in," then "Organizational Login," and enter your UT EID and password.]
- Marini, Joe. "Up and Running with Python." Lynda.com. http://www.lynda.com/Python-tutorials/Welcome/122467/142550-4.html
- Introna, L. D. "The Enframing of Code: Agency, Originality and the Plagiarist." *Theory, Culture & Society* 28, no. 6 (November 1, 2011): 113–41.
- Liu, Alan. "The Meaning of the Digital Humanities." *PMLA* 128, no. 2 (March 2013): 409–23.

Assignment

Weekly Assignment #2

Week 4 (2/8): Collections as Data: Data Models

Readings

Canvas

- Fortune, Stephen. "A Brief History of Databases." *Avant*, February 27th 2014. https://web.archive.org/web/20150220031213/http://avant.org/media/history-of-databases
- Sacasas, Michael. "Do Artifacts Have Ethics?" *The Frailest Thing*, November 29, 2014. http://thefrailestthing.com/2014/11/29/do-artifacts-have-ethics
- van Hooland, Seth, and Ruben Verborgh. "Modelling." In *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*, 11–70. Chicago: Neal-Schuman, 2014.

Optional Readings

- Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–36.
- Joerges, B. "Do Politics Have Artefacts?" *Social Studies of Science* 29, no. 3 (June 1, 1999): 411–31.
- Albon, Chris. "Parse JSON File." http://chrisalbon.com/python/json_parse_file.html
- Lundh, Fredrik. "Elements and Element Trees." http://effbot.org/zone/element.htm [Python XML tutorial]
- Beazley, David, and Brian K. Jones. "Chapter 6: Data Encoding and Processing." In Python Cookbook: *recipes for Mastering Python 3*, 3. ed., 175–216. Bejing: O'Reilly, 2013.
- Zhuang, Atima Han, Ishita Vedvyas, and Rishikesh Dole. "Tutorial: OpenRefine," 2013. http://casci.umd.edu/wp-content/uploads/2013/12/OpenRefine-tutorial-v1.5.pdf

Assignment

Weekly Assignment #3

Week 5 (2/15): An Algorithmic Criticism: Word-Level Text Analysis

Readings

Canvas

- Ramsay, Stephen. "Chapter 1: An Algorithmic Criticism." In *Reading Machines: Toward an Algorithmic Criticism*, 1–17. Topics in the Digital Humanities. Urbana: University of Illinois Press, 2011.

- Montfort, Nick. “Text III.” In *Exploratory Programming for the Arts and Humanities*, 185–213. Cambridge, MA: The MIT Press, 2016.
- “Alphabetical list of part-of-speech tags used in the Penn Treebank Project.” https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Optional

- Fellenbaum, Christiane. “Wordnet(s).” In The *Encyclopedia of Language & Linguistics*, edited by E. K. Brown and Anne Anderson, 2nd ed., 14:665–79. Amsterdam ; Boston: Elsevier, 2005.
- Burrows, John. “Textual Analysis.” In *Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Link.

Assignment

Weekly Assignment #4

Week 6 (2/22): Web Scraping & APIs

Readings

Canvas

- Albon, Chris. "Beautiful Soup Basic HTML Scraping." http://chrisalbon.com/python/beautiful_soup_html_basics.html
- Peters, Justin. *The Idealist: Aaron Swartz and the Rise of Free Culture on the Internet*, Chapters 7 and 8. New York: Scribner, 2016.
- Swartz, Aaron. “Building a Platform: Providing APIs.” In *Aaron Swartz’s ‘A Programmable Web’: An Unfinished Work*, 31–39. San Rafael, CA: Morgan & Claypool Publishers, 2013.
- Kelly, Chelsea Emelie. “Beyond Digital: Open Collections & Cultural Institutions,” 2014. https://artmuseumteaching.com/2014/11/06/beyond-digital-open-collections-cultural-institutions

Optional Readings

- Pomerantz, Jeffrey. "The Future of Metadata." In *Metadata*. The MIT Press Essential Knowledge Series. Cambridge, MA ; London, England: The MIT Press, 2015.
- “HTML Introduction” and “HTML5 Introduction.” W3Schools.
  - http://www.w3schools.com/html/html_intro.asp
  - http://www.w3schools.com/html/html5_intro.asp

- Kazil, Jacqueline, and Katharine Jarmul. "PDFs and Problem Solving in Python." In *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*, 91–126. O'Reilly, 2016.
- Sanger, David E., and Eric Schmitt. "Snowden Used Low-Cost Tool to Best N.S.A." The New York Times. February 8, 2014. http://www.nytimes.com/2014/02/09/us/snowden-used-low-cost-tool-to-best-nsa.html

Assignment

Weekly Assignment #5

Week 7 (3/1): The Politics of Open Data

Readings

Canvas

- Greenwald, Glenn. "Chapter 1: Contact." In *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*, 2015.
- American Civil Liberties Union. "First Amendment Lawsuit Brought on Behalf of Academic Researchers and Journalists Who Fear Prosecution Under the Computer Fraud and Abuse Act." https://www.aclu.org/news/aclu-challenges-law-preventing-studies-big-data-discrimination
- Christen, Kim. "Does Information Really Want to be Free? Indigenous Knowledge Systems and the Question of Openness." International Journal of Communication 6 (2012), 2870–2893.

Optional Readings

- Freelon, Deen Goodwin, Charlton D. McIlwain, and Meredith D. Clark. "Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice," 2016. http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf
- Day, Ronald E. "Governing Expression: Social Big Data and Neoliberalism." In *Indexing It All: The Subject in the Age of Documentation*, Information, and Data, 123–44. History and Foundations of Information Science. Cambridge, Massachusetts: The MIT Press, 2014.
- Hitchcock, Tim. "Digital Searching and the Re-formulation of Historical Knowledge" 2008. In *The Virtual Representation of the Past*, edited by Mark Greenglass and Lorna Hughes, 81-90. Ashgate: 2008.

Assignment

Weekly Assignment #6

Week 8 (3/8): Statistics and Visualization

Readings

Canvas

- Krumme, Coco. “What Data Doesn’t Do.” In *Beautiful Data: The Stories behind Elegant Data Solutions*, edited by Toby Segaran and Jeff Hammerbacher, 1st ed. Beijing ; Sebastopol, CA: O’Reilly, 2009.
- Moretti, Franco. "Graphs." In *Graphs, Maps, Trees: Abstract Models for Literary History*, 3–33. London ; New York: Verso, 2007.
- McCandles, David. *Information is Beautiful*. *http://www.informationisbeautiful.net*

Optional Readings

- Montfort, Nick. "Statistics and Visualization." In *Exploratory Programming for the Arts and Humanities*, 215–40. Cambridge, MA: The MIT Press, 2016.
- Gries, Stefan. “Useful statistics for corpus linguistics.” http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.9846&rep=rep1&type=pdf
- Thompson, Clive. “The Surprising History of the Infographic.” http://www.smithsonianmag.com/history/surprising-history-infographic-180959563/?no-ist
- Manovich, Lev. “What Is Visualisation?” *Visual Studies* 26, no. 1 (March 15, 2011): 36–49. http://www.tandfonline.com/doi/abs/10.1080/1472586X.2011.548488.
- Schmidt, B. "Do Digital Humanists Need to Understand Algorithms? http://dhdebates.gc.cuny.edu/debates/text/99

Week 9 (3/22): Your Data, Your Culture

No Readings

Assignment

Due: Proposal

Week 10 (3/29): Machine Learning

Readings

Canvas

- Brew, Chris. “Language Processing: Statistical Methods.” In Encyclopedia of Language & Linguistics, edited by Keith Brown, 2nd ed., 12:597–604. Elsevier, 2006.
- Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica. “Machine Bias.” *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Geitgey, Adam. “Machine Learning is Fun!” *Medium*. https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471
- **Revisit:** Montfort, Nick. "Text III." In *Exploratory Programming for the Arts and Humanities*, 185–213. Cambridge, MA: The MIT Press, 2016.

Optional Readings

- Norvig, Peter. “Natural Language Corpus Data.” In *Beautiful Data: The Stories Behind Elegant Data Solutions*, edited by Toby Segaran and Jeff Hammerbacher, 1st ed. Beijing ; Sebastopol, CA: O’Reilly, 2009.
- Baharudin, Baharum, Lam Hong Lee, and Khairullah Khan. “A Review of Machine Learning Algorithms for Text-Documents Classification.” *Journal of Advances in Information Technology* 1, no. 1 (February 1, 2010).
- Berendt, Bettina, Preibusch, Soren. Toward Accountable Discrimination-Aware Data Mining:The Importance of Keeping the Human in the Loop--and Under the Looking Glass.Big DataVolume 5, Number 2, 2017.
- Wolfram, S. Machine Learning for Middle Schoolers. Stephen Wolfram Blog. 11 May 2017. http://blog.stephenwolfram.com/2017/05/machine-learning-for-middle-schoolers/#comments
- “Working With Text Data.” scikit-learn. http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

Assignment

WA #7

Week 11 (4/5): Critical Text Analysis

Readings

Canvas

- Hammond, Adam. "The double bind of validation: distant reading and the digital humanities' 'trough of disillusionment." *Literature Compass* 14, no. 8 (August 1, 2017): no. pg.

- Marche, Stephen. “Literature Is not Data: Against Digital Humanities.” *Los Angeles Review of Books*, October 28th, 2012. https://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities

- Jockers, Matthew Lee. "Chapter 8: Theme." In *Macroanalysis: Digital Methods and Literary History*, 118–53. Topics in the Digital Humanities. Urbana: University of Illinois Press, 2013.

Optional Readings

- Ramsay, Stephen. “Chapter 3: Potential Readings.” In *Reading Machines: Toward an Algorithmic Criticism*, 33–57. Topics in the Digital Humanities. Urbana: University of Illinois Press, 2011.

- Hall, Gary. "Toward a Postdigital Humanities: Cultural Analytics and the Computational Turn to Data-Driven Scholarship." *American Literature* 85, no. 4 (January 1, 2013): 781–809.

Assignment

Weekly Assignment #8

Week 12 (4/12): Peer Production & Crowdsourcing

Readings

Canvas

- Bodó, Balázs. “Set the Fox to Watch the Geese: Voluntary IP Regimes in Piratical File-sharing Communities.” In Piracy: Leakages from Modernity, edited by James Arvanitakis and Martin Fredriksson, 241–63. Sacramento, CA: Litwin Books, 2014.

- Kreiss, D., M. Finn, and F. Turner. “The Limits of Peer Production: Some Reminders from Max Weber for the Network Society.” New Media & Society 13, no. 2 (March 1, 2011): 243–59.

- Manzo, Christina, Geoff Kaufman, Sukdith Punjasthitkul, and Mary Flanagan. “‘By the People, For the People’: Assessing the Value of Crowdsourced, User-Generated Metadata.” *Digital Humanities Quarterly* 9, no. 1 (2015). http://www.digitalhumanities.org/dhq/vol/9/1/000204/000204.html

Optional Readings

- Benkler, Yochai. "Peer Production and Sharing." In *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, 59–90. New Haven [Conn.]: Yale University Press, 2006.

- Benkler, Yochai, and Helen Nissenbaum. "Commons-Based Peer Production and Virtue." Journal of Political Philosophy 14, no. 4 (2006): 394–419. https://www.nyu.edu/projects/nissenbaum/papers/jopp_235.pdf.

Assignment

Weekly Assignment #9

Week 13 (4/19): Copyright and the Information Commons

Readings

Canvas

- Sims, Nancy. "Library Licensing and Criminal Law: The Aaron Swartz Case." *College & Research Libraries News* 72, no. 9 (2011): 534–37. http://crln.acrl.org/content/72/9/534.short.
- "The Digital Public Library of America Policy Statement on Metadata," 2013. http://dp.la/info/wp-content/uploads/2013/04/DPLAMetadataPolicy.pdf
- Band, Jonathan. "The Complexity Dialectic: A Case Study From Copyright Law," 2015. *policybandwidth*. http://infojustice.org/wp-content/uploads/2015/03/band03102015.pdf

Optional Readings

- Code of Best Practices in Fair Use for Academic and Research Libraries*. Association of Research Libraries, 2012*. http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf
- "Creative Commons: About the Licenses." https://creativecommons.org/licenses/

Assignment

Weekly Assignment #10

Week 14 (4/26): Final Project Workshop

Assignment

Weekly Assignment #11

Week 15 (5/3): Final Presentations

Final Presentation due

5/11: Final Project due

Additional resources:

Installation Tutorials

- Jeroen Janssens, Seven Command Line Tools for Data Science (2013) workbench.

- Juola, P. and Ramsay, S. Six Septembers: Mathematics for the Humanist. Zea E-Books.

- Seaver, Nick "Algorithms as culture: Some tactics for the ethnography of algorithmic systems" Big Data and Society. 9 Nov. 2017