Machine Readings: Text Analysis for the Information Age

Winter 2016

ENGL 55 / MATH 5 / QSS 30.02

Time: at the 12 hour

Location: Haldeman Center 28 (lower level of Kemeny Hall)

Instructor: Allen Riddell

Office Hours: Kemeny 333, Tuesdays 2:00-3:00pm

Description

Library digitization has made millions of books, newspapers, and other printed materials accessible to the public. In this course we will learn how to draw on computational resources to analyze a range of materials, including poetry, novels, newspaper articles, and personal diaries. We will explore debates about the representation of literary texts as "data" and consider the challenges "machine reading" poses for research in the humanities and how we think about what it means to read a text. Through case studies we will reflect critically on the history of the digital humanities (formerly known as humanities computing) and will gain practical experience in text analysis.

Prerequisites

This course welcomes students from a diversity of academic backgrounds. Students are expected to be familiar with algebra at the level of Math 1 or Math 2 (http://www.math.dartmouth.edu/courses/by-course/). No prior experience with computer programming is required for this course.

Required Texts

- Hacking, Ian. *An Introduction to Probability and Inductive Logic.* Cambridge University Press, 2001.
- Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History.* London: Verso, 2005.
- Ulrich, Laurel. *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary, 1785-1812.* New York, NY: Knopf, 1990.

It should be possible to find inexpensive copies of these texts. Other readings will be made available.

Having one of the following texts on hand is recommended:

- The Python Tutorial (http://docs.python.org/3.3/tutorial/) (official documentation). Free, very well written.
- Lutz, Mark. *Learning Python.* Sebastopol, CA: O'Reilly, 2013. 5th edition.
- Gries et al. *Practical Programming: An Introduction to Computer Science Using Python 3.* 2013. 2nd edition.

You may wish to browse the table of contents for these texts and select the one that is best suited to you.

The following (interactive) online resources are recommended:

- Computer Science Circles (University of Waterloo) (http://cscircles.cemc.uwaterloo.ca/) (Python 3)
- Introduction to Computer Science and Programming Using Python (MIT 6.00.1x) (https://www.edx.org/course/introduction-computer-science-mitx-6-00-1x-0) (edX, NB: Python 2!)
- Exercism.io (http://exercism.io/) interactive exercises for beginners with peer review.

Assignments

You will complete three written assignments during the term and a final project. The course also features programming assignments (to be completed outside of class) and in-class lab assignments intended to help you learn the Python programming language and to familiarize you with methods of text analysis discussed during the week.

Course Grade

You will be evaluated according to your work on homework (15%) and lab assignments (10%), on written assignments (35%), and on the final project (40%). In grading written assignments and the final project, I will consider your grasp of and ability to engage analytically with the materials encountered in the course.

You are encouraged to discuss course material, including written assignments and homework. The work that you turn in must reflect their own work. Code and writing submitted may not be copied from other students, books, or websites. Plagiarism, the offense of passing off someone else's work as your own, is a violation of Dartmouth's honor code that the instructor must bring to the attention of the disciplinary committee. Such a violation can result in a student's expulsion from Dartmouth College. Please consult the policies concerning the honor principle (http://www.dartmouth.edu/~deancoll/student-handbook/principles.html). If you have any questions regarding plagiarism or the honor principle, please consult your instructor.

Course Policies

You are expected to attend all class sessions. Missing more than three sessions will have consequences for your grade. Students wishing to take part in religious observances that conflict with their participation in the course should meet with me before the end of the second week of the term to discuss arrangements.

The lowest two grades for non-written assignments will be dropped; no credit will be given for late work. If you are unable to finish an assignment by the deadline, submit your work for partial credit.

Written assignments must be typed, double-spaced, with regular font-size (10-12 point) and standard margins (1" top and bottom 1.25" sides). Acknowledge any ideas that are not your own. Students are encouraged to consult Kate L. Turabian's *A Manual for Writers of Research Papers, Theses, and Dissertations* (Chicago Style) or the *MLA Handbook for Writers of Research Papers* (MLA Style). For help with writing assignments, you may wish to consult John R. Trimble's *Writing with Style* (any edition) and/or a writing tutor in the Writing Resource Center RWIT (Baker Berry Library; appointments may be scheduled online (http://www.dartmouth.edu/~rwit/)).

Students with learning, physical, or psychiatric disabilities are encouraged to contact me in order to discuss accommodations. If you think you might need special accommodation but are not currently registered for those provisions, please contact Student Accessibility Services (http://www.dartmouth.edu/~accessibility/), immediately.

Discussion Kickoffs

Beginning Week 2, each member of the class will select a session during which they will inaugurate our discussion. While considerable flexibility is given to the inagurator, a typical discussion kickoff would include: (1) a succinct description of the project of (one of) the text(s); (2) a brief critical response to the material; and (3) a query or claim addressed to the group arising from the text.

"Green" and "White" assignments

Assignments typically contain two sets of exercises. One set is labeled "white" and the other is labeled "green". By the end of the second week you will need to decide which set of assignments is best suited for you. For example, if you decide on "white" then I expect you will complete the "white" exercises for the rest of the class.

Calendar

Subject to revision. Readings not linked directly are available on the discussion site (https://piazza.com/dartmouth/winter2016/engl5502math501qss3002/home) (under the "Resources" tab (https://piazza.com/dartmouth/winter2016/engl5502math501qss3002/resources)).

Week 1 (January 4, 2016)

Reading (before Wednesday)

- Hockey, Susan, "The History of Humanities Computing (http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1)" in Susan Schreibman, Ray Siemens, and John Unsworth. 2005. A Companion to Digital Humanities. Wiley-Blackwell.
- Svensson, Patrik, "Humanities Computing as Digital Humanities" (http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html). *Digital Humanities Quarterly* 3.3 (2009).
- Install Python 3.5. If you're new to Python, installing the distribution of Python provided by Anaconda (http://docs.continuum.io/anaconda/install) is easiest. Please be sure to install Python 3.5. Do not install Python 2.7.
- Read some or all of "Chapter 1: Creating a Digital Library" (http://nbviewer.ipython.org/github/fbkarsdorp/python-course/blob/book/Chapter%201%20-%20Getting%20started.ipynb) from *Programming for the Humanities*.

Session 1: Introduction to the course, grand tour of quantitative text analysis in the human and interpretive social sciences, history of humanities computing/digital humanities

Session 2: Introduction to the programming language, basic data types, representations of texts.

Material discussed in Session 2 is covered in the Python Tutorial "An Informal Introduction to Python" (https://docs.python.org/3/tutorial/introduction.html) and in "Chapter 1: Creating a Digital Library" (http://nbviewer.ipython.org/github/fbkarsdorp/python-course/blob/book/Chapter%201%20-%20Getting%20started.ipynb).

Lab 1

Assignment 1 (due Thursday, January 14 AoE (https://en.wikipedia.org/wiki/Anywhere_on_Earth)))

If you have any problems installing Python, please ask for help on the discussion site. We'll also have a chance to install the software in class on Friday.

Week 2 (January 11, 2016)

Reading (before Monday)

How (and why) we read

- Hayles, N. Katherine. “How We Read: Close, Hyper, Machine.” ADE Bulletin (2010): 62–79.
- Short, Emily. "Galetea" (http://collection.eliterature.org/1/works/short__galatea.html) (interactive fiction) ("read" using browser version (http://parchment.googlecode.com/svn/trunk/parchment.html?story=http://parchment.toolness.com/if-archive/games/zcode/Galatea.zblorb.js))
- Armitage, D., et al. The Teaching of the Arts and Humanities at Harvard College: Mapping the Future (http://artsandhumanities.fas.harvard.edu/files/humanities/files/mapping_the_future_31_may_2013.pdf). Harvard College, Division of Arts and Humanities, 2013. p. 1-11
- Wieseltier, Leon. "Perhaps Culture is Now the Counterculture" (https://newrepublic.com/article/113299/leon-wieseltier-commencement-speech-brandeis-university-2013) *The New Republic* May 28, 2013. (See also related articles: 1 (https://newrepublic.com/article/114127/science-not-enemy-humanities) 2 (https://newrepublic.com/article/114548/leon-wieseltier-responds-steven-pinkers-scientism))
- Marche, Stephen. “Literature Is Not Data: Against Digital Humanities.” (https://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities) The Los Angeles Review of Books. October 28, 2012. (Dartmouth study referenced is: Quantitative patterns of stylistic influence in the evolution of literature (http://www.pnas.org/content/109/20/7682.full))
- Hanlon, Philip. "The Liberal Arts Imperative" (https://www.dartmouth.edu/~president/announcements/2016-0104.html) January 4, 2016.

Session 3: How we read (and write).

Session 4: Loops and other constructs

Lab 2

Assignments on Canvas

Week 3 (January 18, 2016)

NB: January 18, Monday -- Martin Luther King Jr. day - classes moved to x-periods.

Reading (before Tuesday)

- Laurel Ulrich. 1990. A Midwife’s Tale: The Life of Martha Ballard, Based on Her Diary, 1785-1812. New York, NY: Knopf
- Mining the *Dispatch* (http://dsl.richmond.edu/dispatch/) (website). Browse the website and read "Introduction" (http://dsl.richmond.edu/dispatch/pages/intro).
- Recommended: (Obstetrics in 1846, not too long after Ballard) Listen to NPR report The Doctor Who Championed Hand-Washing And Briefly Saved Lives (http://www.npr.org/blogs/health/2016/01/12/375663920/the-doctor-who-championed-hand-washing-and-saved-women-s-lives)
- Recommended: Watch PBS program *A Midwife's Tale*

Session 5: Time and Text Collections: *A Midwife's Tale* **(Meet on Tuesday during x-hour)**

Session 6: Writing and calling functions

Lab 3

Assignments on Canvas

Week 4 (January 25, 2016)

Reading (before Monday)

- "Introduction" and "Graphs" from Franco Moretti, *Graphs, Maps, Trees* (2005)
- Moretti, Franco. “Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850) (https://www.jstor.org/stable/10.1086/606125).” *Critical Inquiry 36, no. 1* (2009): 134–158. doi:10.1086/606125.
- Cosma Shalizi, "Graphs, Maps, Trees, Fishing" (http://www.thevalve.org/go/valve/article/graphs_trees_materialism_fishing/)
- Recommended: Scott McLemee, "Literature to Infinity" (http://www.insidehighered.com/views/mclemee/mclemee193)
- Recommended: "Adventures of a Man of Science" (https://nplusonemag.com/issue-3/reviews/adventures-of-a-man-of-science/) (2006)

Python resources for this week:

- Matplotlib: plotting (http://scipy-lectures.github.io/intro/matplotlib/matplotlib.html)

Session 7: Sociology of Literature I: Moretti

Session 8: Program design

Lab 4

Assignments on Canvas

Week 5 (February 1, 2016)

NB: Lab moved to Monday, discussion on Wednesday

X-hour review moved to Thursday. E-mail me to schedule a time between 2pm and 5pm.

Reading (before Monday)

- James F. English, "Literary Studies"
- Chapter 2 from Janice A. Radway. 1991. Reading the Romance: Women, Patriarchy, and Popular Literature. 2nd ed. Chapel Hill: University of North Carolina Press
- Larry Isaac. 2009. “Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel.” American Sociological Review 74 (6): 938–965. doi:10.1177/000312240907400605
- Recommended: The VIDA Count 2013 (http://www.vidaweb.org/the-count-2013/)

Session 9: Sociology of Literature II: Radway, Tuchman and Isaac

Session 10: Manipulating text, bag-of-colors, probability

- Recommended: pandas Tutorials (http://pandas.pydata.org/pandas-docs/stable/tutorials.html) (e.g., "10 minutes into pandas")
- Recommended: Chapter 4: Programming Principles (http://nbviewer.ipython.org/github/fbkarsdorp/python-course/blob/book/Chapter%204%20-%20Programming%20principles.ipynb) in *Programming for the Humanities*

Lab 5

Assignments on Canvas

Week 6 (February 8, 2016)

Class on Wednesday is canceled. Lecture will be on Friday. Lab moved to next week.

X-hour review moved to Thursday. E-mail me to schedule a time between 2pm and 5pm.

Readings

- Chapters 4-6 from Ian Hacking. 2001. An Introduction to Probability and Inductive Logic. Cambridge University Press. If this material is new to you, I recommend doing the odd (or even) exercises at the end of each chapter. (Answers are supplied at the end of the book).
- Federalist No. 6 (http://thomas.loc.gov/home/histdox/fed_06.html)
- Chapter 4, "Who Wrote the Disputed Federalist Papers, Hamilton or Madison?" from Mosteller, Frederick and Fienberg, Stephen. *The Pleasures of Statistics: The Autobiography of Frederick Mosteller.* New York, NY: Springer, 2010.
- Mosteller, Frederick. “A Statistical Study of the Writing Styles of the Authors of ‘The Federalist’ Papers.” *Proceedings of the American Philosophical Society* 131, no. 2 (June 1, 1987): 132–140. (Condensed version of Frederick Mosteller and David L. Wallace. 1963. “Inference in an Authorship Problem.” Journal of the American Statistical Association 58 (302): 275–309. doi:10.2307/2283270)
- In lieu of Anti-Federalist essays (recommended):
    - Interview with Sanford Levinson. 2015. Do We Need a New Constitutional Convention?: Reading The Federalist in the twenty-first century (https://bostonreview.net/us/richard-kreitner-sanford-levinson-federalist) *Boston Review*
    - Christian Parenti. 2014. "Reading Hamilton from the Left" (https://www.jacobinmag.com/2014/08/reading-hamilton-from-the-left/) *Jacobin Magazine* (online).

Session 11: Mosteller & Wallace: The Federalist Papers

Session 12: Probability

Lab 6

Assignments on Canvas

Week 7 (February 15, 2016)

Lab during x-hour on Tuesday. Lab may also be done outside of class anytime on Tuesday.

Readings

- Chapters 7, 11, 15 from Ian Hacking. 2001. *An Introduction to Probability and Inductive Logic.* Cambridge University Press. Consider doing a few of the exercises at the end of each chapter. (Answers are in the back of the book.)
- Patrick Juola. “How J.K. Rowling Was Exposed as Robert Galbraith.” (http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/) *Scientific American*, Aug 20, 2013.
- Chapters 1 and 2 from Patrick Juola. 2007. “Authorship Attribution.” *Foundations and Trends® in Information Retrieval* 1 (3): 233–334. doi:10.1561/ 1500000005
- David L. Hoover. 2007. “Corpus Stylistics, Stylometry, and the Styles of Henry James.” (http://www.jstor.org/stable/10.5325/style.41.2.174) Style 41 (2)
- (recommended) D. H. Mellor discusses the meaning of probability (http://philosophybites.com/2014/12/hugh-mellor-on-probability.html).

- (recommended) Video: Rachel Greenstadt "What is the value of anonymous communication?" (https://media.ccc.de/v/32c3-7324-what_is_the_value_of_anonymous_communication) from 32nd Chaos Communication Congress (https://events.ccc.de/congress/2015/wiki/Main_Page)
- (recommended) Video: Deceiving Authorship Detection (https://www.youtube.com/watch?v=C9SgAOcCm0I) from 28th Chaos Communication Congress (https://events.ccc.de/congress/2011/wiki/Main_Page) -->
- (recommended) Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2011. “Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity.” ACM Transactions on Information and System Security 1 (1): 1:1–1:21

Session 13: Authorship attribution, adversarial stylometry, Henry James

Session 14: Probability (cont.), Bayesian Inference

Lab 7

Assignments on Canvas

Final project ({filename}/pages/hw/final-project/final-project.md) (due Sunday, March 13, 2016, AoE)

Week 8 (February 22, 2016)

Readings

- Thomas L. Haskell. 1975. “The True and Tragic History of ’Time on the Cross’.” (https://machinereadings.org/static/readings/haskell1975true_time_on_cross_nyrbooks.pdf). New York Review of Books 22 (15)
- Franco Moretti. 2011. “Network Theory, Plot Analysis (http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf).” New Left Review (68): 80–102
- Cameron Blevins. "Text Analysis of Martha Ballard's Diary" part 1 (http://historying.org/2009/08/31/text-analysis-of-martha-ballards-diary-part-1/), part 2 (http://historying.org/2009/09/09/text-analysis-of-martha-ballards-diary-part-2/), part 3 (http://historying.org/2009/10/19/text-analysis-of-martha-ballards-diary-part-3/)
- Mining the *Dispatch* (http://dsl.richmond.edu/dispatch/) (reprise). Browse the website and read "Introduction" (http://dsl.richmond.edu/dispatch/pages/intro).
- (Recommended) D. Blei. Probabilistic topic models. (http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf) Communications of the ACM, 55(4):77–84, 2012.

Session 15: Networks, 'Time on the Cross'

Session 16: Networks, Topic Models

Lab 8 ({filename}/pages/labs/08/lab-08.md)

Week 9 (February 29, 2016)

Session 17: Lab (Final project work)

Session 18: Lab (Final project work)

Lab 9: Lab (Final project work)

Week 10 (March 7, 2016)

Session 19: Lab (Final project work)

Acknowledgments

The course's subtitle is borrowed from Matt Erlin and Anupam Basu's "Introduction to Digital Humanities: Cultural Analysis in the Information Age". Inspiration for the organization of the programming content came originally from Cosma Shalizi's *Statistical Computing* (http://www.stat.cmu.edu/~cshalizi/statcomp/).