INFO 3350/6350: Text Mining for History and Literature

Fall 2017
Prof. David Mimno

[Last edit 8/23/2017]
Location: Upson 142
Time: MWF 11:15--12:05 / Grad Discussion: F 12:20--1:10
Credits: 3
Websites: http://mimno.infosci.cornell.edu/info3350/, CMS, Piazza
Prereq: Familiarity with Python (eg CS 1110) and a contemporary or historical written culture will be useful, but we expect most students to have one or the other.

Course Description

The course will introduce students to research methods in computer-assisted scholarship. We will learn to represent text documents in computational forms, and to appreciate the effect of choices we make in this process. We will cover a selection of popular tools such as classification, clustering, and topic modeling. Each week we will discuss both the details of computational methods and how each method can be applied in the context of scholarly research.

Contact Information

Email Prof. Mimno for course administrative questions. Post to Piazza about course content questions.

Professor:
David Mimno, Gates 205. 607-255-8919. Best contact method is email mimno@cornell.edu. I will reply within 24 hours.

TA: Laure Thompson, laurejt@cs.cornell.edu; Jason Liao

Grading

Grades will be based on class participation and attendance (10%), in-class assignments (50%), reading responses (20%), and a take-home final exam OR independent research project (20%). Graduate students must do a research project, undergraduates may opt to substitute a research project for the final exam. Work will be turned in through CMS. Regrade requests should go in writing to mimno@cornell.edu.

Graduate Section

Students enrolled under INFO 6350 will attend an additional weekly discussion section and conduct an independent research project alone or in teams of up to three students. Expectations for groups will be raised proportionally. This project will include a research report of approximately 10 pages describing motivation, methods, and findings. The project will also consist of any code and curated data.

Absences and late/missing work

Class time will mix discussion and hands-on programming exercises, so attendance is important. If you will be absent, write to mimno@cornell.edu with an explanation. Late homework will not be accepted, but your lowest homework grade will be discarded. In the unlikely event that you are having difficulty with CMS, work received by email before the deadline will be accepted.

A typical week

We will assign "technical" readings due Monday. On Monday and Wednesday we will provide in-class programming exercises. These will focus on student-directed experimentation. We will be particularly interested in how to "break" things -- most of the work of programming turns out to be debugging, so we will learn to anticipate and recognize problems. The assignments will be designed to be more than you can finish in class. You will complete this work on your own and turn it in the following Monday. These will be graded ✗, ✓-, ✓, ✓+.

We will assign a more "theoretical" reading for Friday. These will be chosen to highlight different perspectives. Friday's class will be an open discussion. We will focus on how readings relate to our own experiences using and applying computational methods. Discussions don't work unless everyone has done the reading and can engage with it. To keep us all honest, you will submit a response of about 250 words through CMS, due Friday morning before class. These will be graded ✗, ✓-, ✓, ✓+.

Laptops

In order to facilitate interactive in-class work, you are allowed to bring a laptop. Work in pairs will be encouraged. If you have a laptop, you will be expected to use it for relevant work. "Multitasking" is a myth. Distractions limit your ability to learn, *and the ability of those near you*. If your laptop is open, expect to show the results of your in-class work, or to have thoughtful questions.

Academic Integrity

We will follow university policies as outlined in the Academic Integrity Handbook. You are encouraged to discuss homework, but each student will complete assignments alone.

Using other people's code is an important part of programming, but for group projects the code should be substantially the work of the group members except for standard libraries. Any code used in projects that was not written by the group members should be placed in separate files and clearly labeled with their source URLs. If you have benefitted from online resources such as examples or StackOverflow answers, list the URLs in comments in your own code, even if you did not directly copy anything.

Project work that relates to your other classes or research is encouraged, but you may not recycle assignments. There must be no doubt that the work you turn in for this class was done for this class.

Students with Disabilities

We will make every effort possible to ensure that the class works for all students. Contact Prof. Mimno if there is anything we should know about. If there is a specific event such as an exam that you are concerned about, please inform us at least two weeks in advance so that we have time to make arrangements.