wilkens-teaching / info3350-f23 Public

Cornell INFO 3350 Fall 2023

8 stars 7 forks Activity

Star Notifications

Code Pull requests Actions Security Insights

main Go to file

wilkens Note location of final exam info … 9 hours ago 96

View code

Information Science 3350/6350

Text mining for history and literature

Fall 2023

Staff and sections

Instructor: Matthew Wilkens

README.md

Credits: 3 (3350) or 4 (6350)
Mode: In person (only)

Lecture: MW 9:05-9:55am, Statler 265
Sections:
As listed in the Cornell course roster (subject to change).

- F 9:05- 9:55, Upson 202, Kiara
- F 11:15-12:05, Upson 216, Kiara
- F 2:30- 3:20, Statler 441, Breanna

- F 3:35- 4:25, Hollister 362, Breanna
- Grad section: F 9:05-11:00am, Upson 102, Matt

Office hours: See Canvas
Additional resources: See the **Mechanics** section below.

Waitlist and prerequisites

The course is currently full, but the waitlist on Student Center is open. If you are on the waitlist, please do attend lectures and section meetings while you wait for a PIN. It is otherwise almost impossible to catch up if you're admitted at the end of the add period.

Note that you must have completed INFO 2950 (Introduction to Data Science) or equivalent to enroll in the class.

Summary

The class will introduce methods for computer-assisted analysis of historical and literary text collections. It will cover corpus curation, representing text as data, building statistical models from text, and interpreting quantitative results. The class will also reflect on how computational methods fit with existing practices in the humanities, and how we can use models as complements to our own interpretations. Following the course, students will be able to assist faculty in quantitative and computational humanities scholarship.

Description

Broadly speaking, the course covers text mining, content analysis, and basic machine learning, emphasizing approaches with demonstrated value in literary studies and other humanistic fields. Students will learn how to clean and process textual corpora, extract information from unstructured texts, identify relevant textual and extra-textual features, assess document similarity, cluster and classify authors and texts using a variety of machine-learning methods, visualize the outputs of statistical models, and incorporate quantitative evidence into literary and humanistic analysis. The course will also introduce some of the more interesting recently published results in computational and quantitative humanities.

Most of the methods treated in the class are relevant in multiple fields. Students from all majors are welcome. Students with backgrounds in the humanities are especially encouraged to join.

Objectives and learning goals

- Describe and evaluate major existing results in quantitative humanities research

- Recognize, explain, evaluate, and implement standard techniques for the use of text as data
- Create reliable, compelling, data-driven humanities research reports that apply suitable text-mining methods to existing humanities questions
- Identify and analyze historical, ethical, and epistemic limitations of existing and potential textual corpora
- **For graduate students:** Evaluate and adapt text-mining methods to current research problems relevant to the student's work
- **For graduate students:** Analyze, evaluate, and present current research findings in computational text analysis

Mechanics

We will use:

- **GitHub** (right here) to distribute lecture materials, code, and datasets. The current versions of the syllabus (this page) and the schedule are always on GitHub, too. You might want to watch or star this repo to be notified of changes.
- CMS to manage problem sets and other code work and to track grades.
- Canvas to distribute restricted readings and other non-public materials.
- Ed for Q&A.

Links and detailed info about each of these are available via the course Canvas site.

Note that you must generally be logged in through your Cornell account to access non-public resources (everything but GitHub).

Work and grading

Basis of Grade Determination

Grades will be based on five problem sets (45% in sum), a take-home final exam or project (35%), three reading responses (10% in sum), and participation and professionalism (10%). **You must achieve a passing grade in each of these components to pass the course.**

Graduate students (enrolled in 6350) must complete a final project in place of the final exam.

Grading Scale

Grades will be assigned on the following scale:

```
97 - 100% A+
93 -  97% A
```

```
90 -  93% A-
87 -  90% B+
83 -  87% B
80 -  83% B-
77 -  80% C+
73 -  77% C
70 -  73% C-
67 -  70% D+
63 -  67% D
63 -   0% F
```

Participation points are awarded primarily for performance in **section**, with some consideration of lecture attendance and participation.

- 100% of participation points: almost always contributes, raising thoughtful points
- 80% of participation points: frequently contributes, raising thoughtful points
- 60% of participation points: occasionally makes a valuable contribution
- 40% of participation points: rarely makes a valuable contribution
- 0% of participation points: attends lectures and section, but never contributes, or actively interferes with learning

Extra credit

A small amount of extra credit will be awarded for IS/Communications SONA study completion (0.5 course point per SONA credit assigned to this class, up to 1.0 total course point) and for consistent, helpful contributions to Ed discussions (up to 1.0 course point).

Texts and readings

There is no required textbook for the course. Assigned readings will be available online, either through the open web or via Canvas. See the schedule for details.

There are five textbooks that may be useful for students who wish to consult them. **They are not required and most students will not need them.**

- Guttag. *Introduction to Computation and Programming Using Python (3rd ed.)*. Useful for students who need or want a refresher on basic concepts in Python.
- Walsh. *Introduction to Cultural Analytics and Python*. An intro-level, interactive textbook developed at Cornell that covers material similar to the first half of 3350. A good place to start if you're feeling behind on the fundamentals.
- Bengfort, Bilbro, and Ojeda. *Applied Text Analysis with Python*. A *very* applied book intended for working developers who want to learn the standard Python stack for text analysis.

- Jurafsky and Martin. *Speech and Language Processing (3rd ed.)*. A detailed textbook focusing on many of the core topics in natural language processing. Probably more advanced than most students will require, but a great resource for those who want more technical depth. The linked version is the openly available draft of the third edition. The published second edition is also available for sale.
- Jockers and Thalken. *Text Analysis with R for Students of Literature (2nd ed.)*. An ideal textbook for this class, but in R. Co-authored by one of our PhD students in IS!

Schedule

In general, Monday lectures will introduce new technical material. Wednesday sessions will combine technical instruction with discussion of assigned readings from the scholarly literature. Friday sections are smaller and devoted to discussion of readings, to focused work on problem sets, and to follow-up questions about topics previously introduced.

For the detailed (and updating) list of topics and readings, see the course schedule.

Final exam

A **final exam** in the form of a take-home project is due during the finals period. More information is available in the `final_exam` directory.

Undergraduates (enrolled in 3350) may elect to complete a project in lieu of the exam. If you take this route, you may work in a group of up to three students. The expected amount of work on the project will be scaled by the number of group members. Except in unusual circumstances, all group members will receive the same grade.

Graduate students (enrolled in 6350) **must** complete an independent project in place of the final exam.

Policies

Harassment and respect

All students are entitled to respect from course staff and from their fellow students. All staff are entitled to respect from students and from fellow staff members. Violations of this principle, whether large or small, will not be tolerated.

Respect means that your ideas are taken seriously, that you feel welcome in class settings (including in study groups and online fora), and that you are treated as a full, co-equal member of the class. Harassment describes any action, intentional or otherwise, that abridges the respect owed to every member of the class.

If you experience harassment in any form, or if you would like to discuss your experience in the class, please see me in office hours or contact me by email. The university also has reporting and counseling resources available, including those for sexual harassment and for other bias incidents.

Attendance

This is a class of moderate size that will make frequent use of lecture and section time to discuss readings and to debate different approaches to academic inquiry. For this reason, attendance at lecture and section is required.

If you need to miss a lecture or section meeting, please complete the absence form *before the meeting in question*. When you return, consult with a classmate to review the material you missed. Lectures and sections will not be recorded. If you miss a section meeting, consult with your section leader for appropriate steps. In every case, assigned work remains due at the appointed time.

Slip days

Late work is accepted subject to a limit of **five total slip days** for the semester. You may submit any individual assignment up to **three** days (72 hours) late. The slip day policy *does not* apply to the reading responses, which may not be submitted late, since they are tied to in-class activities.

If you expect to miss a deadline or to be absent for an extended period due to truly exceptional circumstances, contact Professor Wilkens as far in advance as possible so that we can discuss potential accommodations.

Regrade requests

If you feel that the graders have made a clear, objective, and significant mistake in assessing your work, you may request a regrade via CMS not later than one week after feedback is released. Regrade requests are typically processed within a week or two of submission. You will be notified of the outcome as soon as it is ready.

Remember that this process exists to correct mistakes. This process does not exist to lobby for points. We want to give grades that accurately represent our assessment of your understanding of course material. Hence, if you are given a lower score than you should have been, due to an obvious grading error, you should absolutely bring it to our attention. However, we must explicitly mention an additional consequence of the importance of grade accuracy: if we notice that you have been assigned more points than you should have been, we are duty-bound to correct such scores downward to the correct value.

Academic integrity

Each student in this course is expected to abide by the Cornell University Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work unless specifically and explicitly permitted otherwise.

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

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. When in doubt, consult with me during office hours.

Generative artificial intelligence

I study generative AI and, honestly, I don't think general-purpose systems like ChatGPT will get you very far on the assignments for this class. But I could be wrong. In any case, AI systems aren't banned. If you use them for any aspect of your work, you must:

- Clearly indicate the content of your submission that was produced by an AI system or with the assistance of such a system;
- Identify the system(s) you used;
- Include in full (perhaps as an appendix) the prompts you used to produce the identified content, including prompts you explored but did not use;
- Write a short statement summarizing your assessment of the system's performance on the task.

Use of generative AI *without* following these steps will constitute an academic integrity violation.

I reserve the right to ban the use of generative AI systems outright at a later point in the semester if they become a net negative for our learning goals, but I hope and expect that that won't be necessary.

Disabilities

Every student's access is important to us. If you have, or think you may have, a disability, please contact Student Disability Services for a confidential discussion: sds_cu@cornell.edu, 607-254-4545, or sds.cornell.edu.

- Please request any accommodation letter early in the semester, or as soon as you become registered with Student Disability Services (SDS), so that we have adequate time to arrange your approved academic accommodations.

- Once SDS approves your accommodation letter, it will be emailed to you and to me. **Please follow up with me to discuss the necessary logistics of your accommodations.**
- If you are approved for exam accommodations, please consult with me at least two weeks before the scheduled exam date to confirm the testing arrangements.
- If you experience any access barriers in this course, such as with printed content, graphics, online materials, or any communication barriers, reach out to me and/or your SDS counselor right away.
- If you need an immediate accommodation, please speak with me after class or send an email message to me and to SDS.

Metal Health and Wellbeing

Your health and wellbeing are important to us. There are services and resources at Cornell designed specifically to bolster undergraduate, graduate, and professional student mental health and well-being. If you or a friend are struggling emotionally or feeling stressed, fatigued, or burned out, there is a continuum of campus resources available to you. Help is also available any time day or night through Cornell's 24/7 phone consultation (607-255-5155). You can also reach out to me, your college student services office, your resident advisor, or Cornell Health for support.

Releases

No releases published

Packages

No packages published

Languages

● **Jupyter Notebook** 99.9% ● **Python** 0.1%