# Course Syllabus

Jump to Today

**Tu and Th, 2:55-4:10pm, Klarman KG70** (the room with the piano!)

## Course Staff

| | |
|---|---|
| Instructor | Prof. Cristian Danescu-Niculescu-Mizil (http://www.cs.cornell.edu/~cristian) |
| Grad TAs | Jonathan Chang, Tushaar Gangavarapu, Willy Jiang, Kate Liang |
| Undergrad TAs | Ashwin Ajit, Daniel Botros, Pun Chaixanien, Nicholas Chernogor, Matthew DeSousa, Nicole Ding, Kai Horstmann, Julie Jeong, Kassandra Jordan, Tanisha Kore, Logan Kraver, Neha Kulshreshtha, Sabrina Li, ZhenShu Luan, Sue Ni, Marco Palomino, Harsh Patel, Samhita Raman, Kelly Ryoo, Neha Sunkara, Aaishi Uppuluri, Ria Vora, Ruby Wang, Nadia Wong, Yichen Yao, Sean Zhang |

## Office Hours

| | |
|---|---|
| TA OH | https://queueme.in/course/CS4300-sp-24 (https://queueme.in/course/CS4300-sp-24) (includes OH schedule)<br>*Note: OH schedule to be finalized by the end of the first week; OH already scheduled on QMI will run as expected.* |
| Instructor OH | https://dnm-office-hours.youcanbook.me/ (https://dnm-office-hours.youcanbook.me/) (by appointment only). |

## Relevant Links (access restricted to enrolled students)

- **Course website (Canvas)**: https://canvas.cornell.edu/courses/62833/ (https://canvas.cornell.edu/courses/62833/)
- **CMS**: https://cmsx.cs.cornell.edu/web/auth/?action=course&courseid=1454 (https://cmsx.cs.cornell.edu/web/auth/?action=course&courseid=1454)
- **Ed Stem**: https://edstem.org/us/courses/53550/discussion/ (https://edstem.org/us/courses/53550/discussion/)
- **Gradescope**: https://www.gradescope.com/courses/709539 (https://www.gradescope.com/courses/709539)
- **Course Email**: info-cs-4300@cornell.edu (mailto:info-cs-4300@cornell.edu) (emails go to grad TAs only)

### Summary

How to make sense of the vast amounts of information available online, and how to relate it and to the social context in which it appears? This course introduces basic tools for retrieving and analyzing unstructured textual information from the web and social media. Applications include information retrieval (with human feedback), sentiment analysis and social analysis of text. The coursework will include programming projects that play on the interaction between knowledge and social factors.

### Prerequisites

- **Linear algebra and discrete math**: INFO 2950 or (MATH 2940 and CS 2800)
- **Programming proficiency**: CS 2110 or equivalent and good Python skills

### Learning outcomes

- Numerically represent text documents using vector space models, and compute operations on them such as assessing their similarity.
- Implement algorithms to efficiently organize and search through collections of documents.
- Extract textual and social features from documents and use them to build classification models.
- Evaluate the performance of search, retrieval, and classification models along standard metrics.
- Design, build, and implement a scalable web application that uses concepts learned in class to perform tasks of interest to a broad audience.

---

### Course Policies

**Attendance is mandatory.** We follow a teaching method where you take your own notes as a part of the learning process. As such, *for most lectures, no slides or notes will be provided.* If you must miss a class, please be sure to email us (at: **info-cs-4300@cornell.edu (mailto:info-cs-4300@cornell.edu)** ) in advance, with an explanation.

**Late submission policy.** Unless otherwise stated, all submissions are due 11.59pm on the due date, and you are encouraged to submit early versions, since resubmissions up to the deadline are counted without penalty. We understand that you may be bombarded with multiple deadlines falling on *that* dreadful day; to this end, we offer *at most one "slip day" per assignment*, and three slip days in total over the semester. You may use slip days *only* for individual (non-team) assignments (and not project milestones, etc.).

If you need additional accommodation, ask in writing (addressed to: **info-cs-4300@cornell.edu (mailto:info-cs-4300@cornell.edu)** ) before the deadline, with rationale and a plan for when you will be able to submit the work. We reserve the right to disregard explanations sent after the deadline.

**Use of electronic devices is not allowed.** Notes for this class should be taken on paper or on tablets with a digital pen, and use of electronic devices (laptops, tablets with keyboards, phones) is not

permitted during class time (with some exceptions for specific activities). We're not plain evil; we're just following the findings of extensive research on the negative effects of in-class laptop use on learning.

**Participation.** (Not to be confused with attendance.) We value meaningful contributions ***in class or on Ed*** (e.g., answering the instructor's questions in class, answering other people's questions on Ed). Given the size of the class, not everybody will be able to participate in class, so Ed participation is key: you can get a full participation score by only participating in Ed. There will also be in-class activities that will count toward participation. You can also receive extra credit for participation by making substantial and meaningful contributions during lecture time. What is considered a "meaningful contribution" is at the discretion of the instructor.

**Academic integrity.** We will strictly follow Cornell University's policies on academic integrity as outlined in the **Academic Integrity Handbook (https://provost.cornell.edu/_files/faculty-resources/essential-guide-academic-integrity.pdf)** .

Any work submitted by a student in this course for academic credit will be the student's own work. For this course, collaboration is allowed only when it is made explicit in the assignment or project description. In case of doubt, contact the instructor.

All assignments may be subject to submission for textual/coding similarity review to plagiarism detection services.

All course materials are intellectual property belonging to the author. Students are not permitted to buy, sell or distribute any course materials without the express permission of the instructor. Such unauthorized behavior constitutes academic misconduct.

---

### Accommodations

**Students with Disabilities:** Your access to this course is important to us. Please request your 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 both you and the instructors. Please follow up with a Grad TA to discuss the necessary logistics of your accommodations.
- If you are approved for exam accommodations, please consult with a GRAD TA 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 a grad TA or SDS right away.
- If you need immediate accommodation, please speak with the instructors after class or send an email message to us, a Grad TA and SDS at **sds_cu@cornell.edu (mailto:sds_cu@cornell.edu)** .

If you have or think you may have, a disability, please contact Student Disability Services for a confidential discussion: **sds_cu@cornell.edu (mailto:sds_cu@cornell.edu)** or visit **sds.cornell.edu (http://sds.cornell.edu)** to learn more.

**Mental health and well-being:** Your health and well-being 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. Remember, your mental health and emotional well-being are just as important as your physical health. 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: **https://mentalhealth.cornell.edu/get-support/support-students (https://mentalhealth.cornell.edu/get-support/support-students)** . 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 us, your college student services office, your resident advisor, or Cornel Health for support.

---

### Grading

Grades will be based on:

- participation in ED [5%], participation in class for extra credit;
- midterm [20%];
- open-ended final project [30%];
- assignments/in-class quizzes [45%];

There will also be several opportunities for extra credit in all of these categories. Extra credit will be added directly to the final course grade.

Grade statistics will not be released, except for the midterm where an interpretation of your score will be provided. Your final grade will reflect your command and understanding of the course material, and the effort you have put into the assignments and your final project, and will not be influenced by the grades of your colleagues. Grades will be assigned according to the Cornell University Grading System. A+'s are rare and reserved for students with exceptional performance and participation; what counts as exceptional is at the discretion of the instructor.

Students are referred to the Cornell University Grading System for details. Information can be found at **https://courses.cornell.edu/content.php?catoid=31&navoid=7933#Grading_System (https://courses.cornell.edu/content.php?catoid=31&navoid=7933#Grading_System)** and **https://cpb-us-e1.wpmucdn.com/blogs.cornell.edu/dist/3/6798/files/2016/01/CUgrading-1qbumf0.pdf (https://cpb-us-e1.wpmucdn.com/blogs.cornell.edu/dist/3/6798/files/2016/01/CUgrading-1qbumf0.pdf)**

---

### Prelim (and related logistics)

The midterm will be administered in mid-March during class time (the exact dates will be announced later). Attendance is mandatory and no make-up exams will be administered (save for exceptional cases involving documented medical emergencies).

---

**Schedule**

The schedule and list of topics will be in flux. Here is a tentative outline:

| Week | Content |
|---|---|
| 1 | Intro, Dimensions of information systems, Conversational behavior |
| 2 | Types and tokens, Document similarity |
| 3 | Vector space models, TF-IDF weighting |
| 4 | Indexing, Boolean search |
| 5 | Evaluation of IR systems |
| 6 | Ranked retrieval |
| 7 | Relevance feedback |
| 8 | Midterm |
| 9 | Text classification, rundown of textual features |
| 10 | Practical unsupervised text classification |
| 11 | Large language models in IR |

| | |
|---|---|
| 12 | Hubs and authorities Spectral analysis |
| 13 | Opinion mining, Trust, Deception |
| 14 | Project Presentations |
| 15 | TBD |

**Recommended Textbooks**

- Manning, Raghavan, and Schutze. 2008. Introduction to Information Retrieval. Cambridge University Press.
- Jurafsky and Martin. 2009. Speech and Language Processing (2nd Edition). Pearson.

# Course Summary:

| Date | Details | Due |
|---|---|---|
| Fri Jan 26, 2024 | **A0: Getting (to know) the Kardashians (https://canvas.cornell.edu/courses/62833/assignments/627605)** | due by 11:59pm |
| Tue Jan 30, 2024 | **A1: Keeping up with social information (https://canvas.cornell.edu/courses/62833/assignments/628340)** | due by 11:59pm |