INFO/CS 4300: Language and Information, Spring 2019

Course Syllabus

- **PhD TAs:** Xilun Chen, JiHyun Jeong
- **Graduate TAs:** Harrison Unruh, Charles Yu
- **Undergrad TAs listed on Piaza**
- **Office hours schedule listed on Piazza** (Resources -> Staff)
- **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.

Academic Integrity

We will strictly follow Cornell University's policies on academic integrity as outlined in the Academic Integrity Handbook.

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.

Late submissions and attendance

Attendance is mandatory. We use a teaching method where taking your own notes in class is part of the learning process. As such, for most lectures no slides

or lecture notes will be provided. If you must miss a class, please email the instructor beforehand to provide an explanation. Late submissions will not be accepted, save for (documented) major medical or family events.

Electronic device policy

Notes for this class should be taken on paper. Use of electronic devices such as laptops and tablets will not be permitted during class (with the exception of specific activities). We are not plain evil, we are just following extensive research on the negative effects of in-class laptop use on learning.

Grading (subject to change)

Grades will be based on:

- participation (in-class or on Piazza) [10%];
- assignments/homeworks/in-class quizzes [40%];
- midterm [20%];
- open ended final project [30%];

No auditing is allowed.

Participation is distinct from attendance (which is mandatory). You can gain participating points by making meaningful contributions in class or on Piazza (e.g., answering the instructor's questions in class, answering other people's questions on Piazza). Given the size of the class not everybody will be able to participate in class, so Piazza participation is key: you can get full participation score by only participating on Piazza. What is considered "meaningful contribution" is at the discretion of the instructor.

Grade statistics will not be released. You final grade will reflect your command and understanding of the course material, and the effort you have put in the assignments and in your final project, and will not be influenced by the grades of your colleagues.

Midterm (subject to change)

The midterm will be administered in-class, likely in mid March (exact date will be announced later). Since this will be an in-class midterm there will be no makeup, so plan to attend.

SONA Credits

You can get extra credits for participating in experiments and research studies through Science Research Participation System. You will receive 0.5 extra credit for each 30 minute study (or equivalent). Note that you can not receive more than 0.5 for this course.

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 outline

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 Th	Ranked retrieval
7	Relevance feedback
8	Midterm
9	Text classification, rundown of textual features
10	Practical unsupervised text classification
11	Spring Break
12	Social features, Page Rank
13	Hubs and authorities Spectral analysis
14	Opinion mining, Trust, Deception
15	Final project presentations
16 Tu	TBD