INFO/CS 4300: Language and Information, Spring 2016

Course Syllabus

- **Time and place** TuTh 2:55pm-4:10pm, Hollister Hall B14
- **Instructor:** Prof. Cristian Danescu-Niculescu-Mizil
- **Head PhD TA:** Jack Hessel
- **PhD TA:** Xanda Schofield
- **Course Piazza page:** piazza.com/cornell/spring2016/cs4300info4300
- **Course homepage:** www.cs.cornell.edu/Courses/cs4300/2016sp/
- **Summary:** How can we make sense of the vast amounts of information available online, and how do we relate it 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), text classification and social analysis. The coursework will include small projects that play on the interaction between knowledge and social factors.
- **Prerequisites:**
- INFO 2950 (or CS 2800 and a linear algebra course)
- INFO 3300 (or a machine learning course)
- Good Python programming skills and familiarity with IPython Notebooks, of which we'll make extensive use.
- **Related courses offered this semester at Cornell:**
- CS 5740 Natural Language Processing
- CS 4786/5786 Machine Learning for Data Science
- CS 4850: Mathematical Foundations for the Information Age

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 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, as for most lectures there will be no lecture slides. If you must miss a class, please email the instructor beforehand to provide an explanation. Late submissions will not be accepted, save for 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/quizzes [40%];
- midterm [20%];
- open ended final project [30%];

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), up to a maximum of 2%.

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 Th	Intro, Dimensions of information systems, Conversational behavior
2	Types and tokens, Document similarity
3	Vector space models, TF-IDF weighting
4 Th	Indexing, Boolean search
5	Evaluation of IR systems
6	Ranked retrieval
7	Relevance feedback
8	Midterm, Text classification
9	Feature design, Feature selection, Re-ranking
10	Spring Break
11	Project proposals
12	Social features, Page Rank
13	Hubs and authorities Spectral analysis
14	Opinion mining, Trust, Deception
15	Final project presentations
16 Tu	TBD