

CS674/INFO 630: Advanced Language Technologies, Fall 2007
8/23/07: Course description and policies

All satisfied with their seats? O.K. No talking, no smoking, no knitting, no newspaper reading, no sleeping, and for God's sake take notes.

– Nabokov, *Lectures on Literature*

Instructor Lillian Lee. Office hours: usually Thursdays 3-4 in Upson 4152, or by appointment; for contact info and updates, see <http://www.cs.cornell.edu/home/llee> .

Lecture time and place TR 1:25-2:40, Hollister 312.

Course homepage <http://www.cs.cornell.edu/courses/cs674/2007fa> . Along with course handouts and lecture guides, additional references and other auxiliary information will be posted.

You may also find the Spring 2006 webpage, <http://www.cs.cornell.edu/courses/cs630/2006sp/> , useful (and, most likely, similar).

Prerequisites Knowledge of elementary computer science, probability, and linear algebra. Neither CS/INFO 430 nor CS/COGST/LING 474 are prerequisites.

Philosophy (principles and caveats), Fall 2007 version The overall aim of this course is to cover fundamental ideas underlying research in information retrieval (IR) and natural language processing (NLP). Thus, one goal is that students should understand the development of prior influential models and algorithms to such a degree as to be able to create new techniques — as a result, in this course, there will be particular emphasis on the mathematical derivations behind classic approaches. A second goal is to encourage the development of a skill that is crucial to the research enterprise: the ability to formulate interesting yet “feasible” questions. This is one motivation behind the lecture-guide course-work requirement, described further below.

CS 674/INFO 630 does not exist in a vacuum. Cornell offers many excellent courses that are directly relevant to modern human-language technologies. I have decided to minimize overlap with these classes. It is true that this poses the risk of giving the mistaken impression that IR and NLP are somehow disjoint from machine learning, network analysis, human-computer interaction, linguistics, and other important areas. Let me therefore stress that students interested in research in IR and NLP should take courses in the areas just mentioned. It is merely the fact that they probably will do so or have already done so that has triggered the decision to de-emphasize these subject areas in this particular course.

Tentative syllabus Three fundamental paradigms in information retrieval: the vector-space model; the (Robertson-Spärck Jones) probabilistic retrieval paradigm; the language-modeling approach. Relevance feedback, explicit and implicit. Latent Semantic Indexing (LSI). Feature-based context-free grammars (CFGs). Tree adjoining grammars (TAGs). Parsing. The Expectation-Maximization (EM) algorithm. Maximum-entropy modeling.

(OVER)

Course work The approximate proportion of the course grade is indicated in brackets.

- Exams [30%]: In-class midterm, Thursday October 18 [15%]; final exam Thursday, December 13, 2-4:30pm [15%].
- Participation [10%]. Participation can occur outside of the regular class period (e.g., technical conversation in office hours or via email).
- Lecture guides [60%]. It is often said that the best way to learn something is to teach it. Also, a crucial part of graduate education is developing the ability to formulate (and answer) good questions.

To that end, for each lecture, one or more student groups will be responsible for writing up a *lecture guide*. A *satisfactory* guide would consist of:

1. roughly textbook-quality “scribe notes” (edited and organized transcriptions) for the lecture; and
2. an original, non-trivial “finger-exercise” problem that tests one or more basic concepts, together with a worked solution to that problem.

For *maximum* credit, the guide should also include a “deeper” question and solution (partial solutions may be acceptable if the question is sufficiently “research-y”). Extra credit may be awarded for extra effort, such as including brief summaries of recent related results or posing additional interesting problems. Samples can be viewed on the Spring 2006 webpage mentioned above.

The procedure is as follows. Lecture guides should be submitted as (spell-checked and proof-read) hardcopy within a week (except when breaks or exam-preparation times intervene, in which case extensions will be arranged). Feedback — every document can benefit from revision! — will be provided in the form of markup on this hardcopy. An updated version should then be submitted as a pdf file via email within a week (hopefully sooner if the changes to be made are minor) so that the finished lecture guide can be posted to the course webpage.

A check-plus/check/check-minus/zero grade will be assigned to the scribe notes first submitted, and, separately, to the worked problem(s). The quality of the update (e.g., whether it fixes problems pointed out in the feedback) will also be graded as satisfactory/unsatisfactory.

Groups will probably be responsible for a lecture guide about once every two to three weeks, although the exact schedule will depend on how many students enroll. My administrative assistant, Cindy Robinson, will be helping to coordinate scheduling, so please expect (and respond to) email from cindy@cs.cornell.edu.

Academic Integrity Academic and scientific integrity compels one to properly attribute to others any work, ideas, or phrasing that one did not create oneself.

I take violations of the Code of Academic Integrity (<http://www.cuinfo.cornell.edu/Academic/AIC.html>) very seriously, and have assigned failing grades for such violations in the past.