ECE8813 Special Topic on Statistical Natural Language Processing (C.-H. Lee)

Course Description:

With the availability of a large collection of text corpora in electronic forms in libraries and on the web, statistical language processing is becoming an important tool for understanding words and their interactions. Many signal processing tools are applicable to language analysis. Language engineering also emerges as a new area for research and applications. In this course, foundations of statistical natural language processing will be covered, and many applications will be addressed. Plenty of text examples will be used for hands-on homework exercises. An individual course project is the main outcome of this special topic course.

Intended Audience: students interested in learning about language engineering

Course Outline:

- Introduction
- Mathematical Foundations
- Linguistics Essentials
- Corpus-Based Linguistic Analysis
- Word Collocations
- Statistical *n*-Gram
- Word Sense Disambiguation
- Markov Models
- Part-of-Speech Tagging
- Probabilistic Parsing
- Clustering
- Information Retrieval
- Text Categorization
- Statistical Alignment and Machine Translation

Grading Policy: Students are graded based on the following:

- **Homework** (25%)
- **Project (**35%**): a project presentation and report are required**
- **Examinations** (40%)
    - Midterm (15%):
    - Final (25%):

Prerequisites: ECE3075 or equivalent (ISYE 3770 or MATH 3770 or CEE 3770)

Estimated Student Size: 12-15 (with potential attendees from CoC)

Main Text: C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 2001 (ISBN: 0262133601)

Supplemental Reading: will be furnished during the semester

Students with disabilities

Georgia Tech offers accommodations to students with disabilities. If you need a classroom accommodation, please make an appointment with the ADAPTS office (see http://www.adapts.gatech.edu).

Georgia Tech Academic Honor Code

This course will follow the policies and guidelines set forth by the Georgia Tech Honor Code (www.honor.gatech.edu). In particular, we will observe the following policies.

Plagiarism: Plagiarizing is defined by Webster's as "to steal and pass off (the ideas or words of another) as one's own : use (another's production) without crediting the source."

If caught plagiarizing, you will be dealt with according to the GT Academic Honor Code.

Homework: Late homework will not be accepted. You are allowed (and encouraged) to work together with other students on homework, as long as you write up and turn in your own solutions. You are also allowed (and encouraged) to ask me questions, although you should try to think about the problems before asking. I strongly encourage you to work on extra problems from the book on your own.

Quizzes and Exams: Cheating off of another person's test or quiz is unethical and unacceptable. Cheating off of anyone else's work is a direct violation of the GT Academic Honor Code, and will be dealt with accordingly.

Finally, for any questions involving these or any other Academic Honor Code issues, please consult me, my teaching assistants, or www.honor.gatech.edu.