CSCI 394

Seminar: Computational Linguistics

Fall 2013 Th 8:30–10:20 SCI 131

http://cs.wheaton.edu/~tvandrun/cs394

Thomas VanDrunen ☎630-752-5692 630-639-2255 ✉Thomas.VanDrunen@wheaton.edu
Office: SCI 163 Office hours: MWThF 3:15–4:45; Th 11:00–12:00.

Contents

CATALOG DESCRIPTION. An exploration of big ideas in computational linguistics, natrual language processing, and/or language technologies. To be tailored to the backgrounds and interests of those enrolled.

TEXTBOOK. Daniel Jurafsky and James Martin, *Speech and Language Processing,* second edition, Prentice Hall, 2009.

(Also recommended:) Steven Bird et al, *Natural Language Processing with Python,* O'Reilly, 2009. (A copy will be available in the computer science student lounge and also at http://nltk.org/book/)

OBJECTIVES. Since this is an experimental seminar course, the objectives are elastic. Generally, at the end of this course students should

- be aware of the principle definitions, goals, methods, contributions, and applications of the field of computational linguistics (and/or natural language processing, language technologies, statistical language modeling. . .);
- have some experience using some of the tools and methods of the field;
- have an increased understanding of the algorithms and mathematics powering the language technologies in the world around us; and
- have learned something about programming/algorithms, linguistics, or statistics (hopefully all three).

OUTLINE.

I. Foundations
   A. General introduction
   B. Regular expressions and automata
   C. Linguistics background
   D. The noisy channel model
   E. Statistics background

II. Words
   A. Types and tokens
   B. Unigrams
   C. Smoothing
   D. Higher n-grams

III. Information theory

   A. Definitions and concepts
   B. The entropy of English
   C. Perplexity of language models

IV. Beyond words

   A. Categorizations
   B. Hidden Markov Models
   C. Part-of-speech tagging
   D. Lexical semantics
   E. Parsing and generative grammars

V. Applications

   Possible topics: Discourse analysis, authorship attribution, usage trends, machine translation, discourse agents, information retrieval...

For a schedule, see the course website.

Course procedures

How we do this course. For each class period, students will have a portion of the text book and/or supplemental material to read ahead of time (this may be divided into "read carefully" and "skim" portions). Most of classtime will be in a lecture format, with occasional group work on problems (including coding problems). Most of the outside work in this course will go into four projects. Students will be lightly examined for cummulative retention of the concepts in a final exam to be held during this course's exam block.

There may also be quizzes and short assignments to enforce and exercise the concepts in class and to give a feel for what the exam will be like. (These would be rolled into the "Participation" portion of the grading scheme below.)

Projects. Expected topics and estimated assigned dates / due dates for projects: Project 1 (regular expressions), Sept 4–Sept 25; project 2 (n-grams), Sept 25–Oct 30; project 3 (HMMs), Oct 30–Nov 14; project 4 (application), Nov 14–Dec 12. Subject to change.

Grading. The final exam block is Thursday, Dec 19, 8:00–10:00 am.

instrument	*weight*
Participation	10
Projects (4)	60 (total)
Final exam (light)	30

The course is not meant to be grade-competitive. Assignments and grading will take into consideration the disparity of backgrounds among the students (possibly by having multiple project tracks). Those who participate and learn something should get a good grade.

Policies etc

Instructor's Two-Hour Pledge. I am committed to keeping the work load in this course proportional to its two-hour credit designation. Please alert me if you perceive the work load is becoming more than half of what would be expected for a four-hour course.

For comparison, if this were a four-hour course, the workload and weighting of the grades would probably look like this:

instrument	*weight*
Participation	10
Close-ended projects (4)	35 (total)
Midterm	15
Term project	25
Final exam	15

Academic Integrity. Collaboration among students enrolled in the course is permitted on all projects and assignments (in some cases, collaboration among students with complementary skills in fact may be necessary). Obtaining solutions to projects through electronic or other media is strictly prohibited. Any ideas obtained through electronic, print, or other media must be cited as they would in a research paper. Any violations will be handled though the college's disciplinary process.

Assignments. Late projects will not normally be accepted. A student may renegotiate a deadline no later that 24 hours before the original deadline (except in cases of truly unforseen circumstances); good reason must be cited for such an extension, and a point deduction may be incurred.

Attendance. Students are expected to attend all class periods. It is courtesy to inform the instructor when a class must be missed.

Examinations. The final exam is Thursday, Dec 19, at 8:00 AM. I do not allow students to take finals early (which is also the college's policy), so make appropriate travel arrangements.

Special needs. *Institutional boilerplate:* Wheaton College is committed to providing reasonable accommodations for students with disabilities. Any student with a documented disability needing academic adjustments is requested to contact the Academic and Disability Services Office as early in the semester as possible. Please call 630.752.5941 or send an e-mail to `jennifer.nicodem@wheaton.edu` for further information.

My own statement: Whenever possible, classroom activities and testing procedures will be adjusted to respond to requests for accommodation by students with disabilities who have documented their situation with the registrar and who have arranged to have the documentation forwarded to the course instructor. Computer Science students who need special adjustments made to computer hardware or software in order to facilitate their participation must also document their needs with the registrar in advance before any accommodation will be attempted.

Gender-inclusive language. For academic discourse, spoken and written, the faculty expects students to use gender inclusive language for human beings. This is not actually relevant for this course, but it's an official college policy that the syllabus contain some sort of notice like this.

Office hours. I try to keep a balance: Stop by anytime, but prefer my scheduled office hours. Any time my door is closed, it means I'm doing something uninterruptable, such as making an important phone call. Do not bother knocking; instead, come back in a few minutes or send me an email.

Dress and deportment. Please dress in a way that shows you take class seriously—more like a job than a slumber party. (If you need to wear athletic clothes because of activities before or after class, that's ok, but try to make yourself as professional-looking as possible.) If you must eat during class (for schedule or health reasons), please let the instructor know ahead of time; we will talk about how to minimize the distraction.

Electronic devices. Please talk to me before using a laptop or other electronic device for note-taking. You will need to convince me that it truly aides your comprehension. No student has convinced me yet, but if you're the first one then I will give you a stern warning against doing anything else besides note-taking. Trying out programming concepts on your own during class time (for example) is not productive because it takes you away from class discussion. You cannot multi-task as well as you think you can. Moreover, please make sure other electronic devices are silenced and put away. ***Text in class and DIE.***