CSCI 384

Computational Linguistics

Fall 2017 MWF 12:55–2:05 SCI 131

http://cs.wheaton.edu/~tvandrun/cs384

Thomas VanDrunen

☎630-752-5692 📱630-639-2255 ✉Thomas.VanDrunen@wheaton.edu

Office: SCI 163 Office hours: MWF 3:30–4:30pm; Th 9:00–10:30am, 11-11:30 and 1:15–3:15pm.

Contents

Catalog description. An exploration of big ideas in computational linguistics, natural language processing, and/or language technologies. Language models, n-grams, information theory and entropy, and semantics. Applications of computational linguistics such as part-of-speech tagging, authorship attribution, automatic translation, and sentiment analysis. Prerequisite: CSCI 345 (non-majors without the prerequisite may enroll with departmental approval).

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. At the end of this course students should

- demonstrate understanding 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. . .);
- demonstrate a basic competence in some of the tools and methods of the field (in particular, the NLTK library for Python and the Stylo package for R);
- demonstrate understanding of the algorithms and mathematics powering the language technologies in the world around us; and
- describe and assess how language technologies are affecting popular culture, business and marketing, scholarship in the humanities (including biblical scholarship), our understanding of human language, and the ministry of the gospel.

Some additional, incidental students will learn are

- Basic Python programming.
- Intermediate algorithms using dynamic programming.
- Lexical categories, grammars, and other ideas from linguistics.
- Fundamental concepts of information theory.
- Basic concepts of probability and statistical inference.
- Introductory R programming.
- Fundamental concepts in machine learning.

This course will support and reinforce all of the computer science program outcomes.

OUTLINE.

The following is a *conceptual* outline of the course to explain the relationships among the various topics and themes. It is not a *temporal* outline; background, core, and applied topics will be pursued in parallel. For the sequence of coverage and schedule, see the course website.

I. Background

   A. Linguistic

      i. Phonetic
      ii. Lexical
      iii. Syntactic
      iv. Semantic
      v. Rhetorical

   B. Statistical

      i. Basic probability
      ii. Bayes's theorem
      iii. Distributions
      iv. Hidden Markov models

   C. Computational

      i. Regular expressions
      ii. Finite-state automata
      iii. Dynamic programming
      iv. Machine learning

II. Core

   A. Words

      i. Types, tokens, n-grams
      ii. Heads and tails, hapax legomena, function words, stop words
      iii. Rank vs frequency; Zipf's law
      iv. Lexical semantics

   B. Language models

      i. Maximum likelihood, relative frequency
      ii. Perplexity
      iii. Smoothing

   C. Information theory

      i. Goals and basic theory
      ii. Entropy
      iii. Linguistic insights from entropy

III. Applications

   A. Standard

      i. POS tagging
      ii. Spelling correction
      iii. Parsing

   B. Analytical

      i. Sentiment analysis
      ii. Stylometry and authorship attribution
      iii. Information retrieval

   C. Synthetic

      i. Machine translation
      ii. Text generation
      iii. Discourse agents

Course procedures

HOW WE DO THIS COURSE. For most class periods, students will prepare by reading a portion the text book and/or supplemental material to read, and/or complete a short exercise. Often class will begin with a quiz to enforce the reading and to help students assess their progress. Classtime will be a mixture of lecture and group work, especially coding problems done in the lab. Most of the outside work in this course will go into the projects (about six). The final exam will assess students' mastery and retention of the formal side of the topic stream.

Note in particular that in this class this semester I am not enforcing readings by having students submit summaries but instead using quizzes.

PROJECTS. Expected topics and seasons for the projects: Project 1 (regular expressions and NLTK), early September; project 2 (language models, evaluated intrinsicly), mid September; project 3 (edit distance and extrinsic evaluation of language models), late September, early October; project 4 (Hidden Markov Models and POS tagging), mid/late October; project 5 (parsing), early/mid November; project 6 (stylometry), late November, early December. Subject to change.

GRADING. There will be a take-home midterm near the middle of the semester (exact dates to be announced later). The final exam will be taken during our scheduled exam block, Thursday, Dec 14, 8:00-10:00 am.

instrument	*weight*
Small pieces	20
Projects (about 6)	45 (total)
Midtern (take home)	15
Final exam	20

"Small pieces" includes quizzes, short assignments, and in-class exercises (which might be completed outside class), etc.

Policies etc

ACADEMIC INTEGRITY. Collaboration among students enrolled in the course is permitted on all projects and assignments. 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. The take-home midterm must be done completely independently, using only official materials for the course as reference. Any violations will be handled though the college's disciplinary process.

LATE ASSIGNMENTS. Students are allowed three late days for projects, which may be distributed among the project (for example, one project one day late and another project two days late). Beyond that, late projects will not normally be accepted.

Fine print: "Late days" may be used only in whole number amounts and are interpreted as calendar days, regardless of whether school is in session. Projects will be due at midnight between their posted due-date and the following day.

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

EXAMINATIONS. Students are expected to take all tests, quizzes, and exams as scheduled. In the case where a test must be missed because of legitimate travel or other activities, a student should notify the instructor no later than one week ahead of time and request an alternate time to take the test. In the case of illness or other emergency preventing a student from taking a test as scheduled, the student should notify the instructor as soon as possible, and the instructor will make a reasonable accommodation for the student. The instructor is under no obligation to give any credit to students for tests to which they fail to show up without prior arrangement or notification in non-emergency situations. The final exam is Thursday, Dec 14, 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 statement:* 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. If you have documented need for accomodations, **please talk to the instructor** about what accomodations you find most useful. 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. The college requires the following statement to be included on all syllabi: *For academic discourse, spoken and written, the faculty expects students to use gender inclusive language for human beings.*

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 uninterruptible, 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 keep laptops and all electronic devices put away and silenced during class. ***Text in class and DIE.***