CSCI 384

Computational Linguistics

Fall 2021 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: Drop-in: MWF 3:30–4:30pm; Or by appointment through Calendly

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,* third edition draft, available at https://web.stanford.edu/~jurafsky/slp3/. (The second edition was published by Prentice Hall, 2009.)

Steven Bird et al, *Natural Language Processing with Python,* updated edition, available at http://nltk.org/book/. (The original version of the book was published by O'Reilly, 2009).

PURPOSE OF THE COURSE. In addition to the stated topic of *computaitonal linguistics*, this course plays a distinctive role in the CSCI program in that in that it is throughly interdisciplinary, it exposes students to areas of computer science not covered elsewhere in our curriculum, and has tangible applications used everyday. This course is in some ways a companion class to CSCI 381 Machine Learning, though effort has been made to keep the two courses independent of each other. This course presents a balanced approch among the *theory* of statistical language modeling, the *algorithms* of natural language processing, and the *applications* of language technology, but special emphasis is put on understanding the algorithms by implementing them from scratch.

GOALS AND OBJECTIVE. The goals of this course are that students will

1. Articulate the goals and use the terminology of computational linguistics and related fields.
2. Complete the laboratory exercises and projects to demonstrate a basic competence in the tools and methods of the field, such as the NLTK library for Pythoon.
3. Identify the elements of widely-used computational-linguisitic algorithms.

The objective of the course is that students will be able to

1. Explain how language technologies are derived from linguistics, statistics, and computation.
2. Implement natural language processing algorithms correctly and use them for interesting applications.
3. Discuss how language technologies are affecting popular culture, scholarship in the humanities our understanding of human language, and the ministry of the gospel.

Other, incidental topics that students will learn include 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; and fundamental concepts in machine learning.

In addition to these, together we have the general objective of seeing computational linguistics as a way of knowing God's world and a tool for doing good, to God's glory.

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 and Schoology.

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. Rank vs frequency; Zipf's law
     - iii. 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 or 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. The midterm and final exam will assess students' mastery and retention of the formal side of the topic stream.

Implementation platform. Code examples, labs, and programming projects will be done using Python 3. Students without prior experience in Python are responsible for learning the basics of Python on their own. Resources for learning Python can be found on the course website. We will make extensive use of certain libraries, (especially in lab, less so in projects). The main library we will use is `nltk`.

Laboratory activities. Collaborative in-class lab assignments will constitute a major portion of students' experience in this course. We will use Jupyter notebooks as our programming environment. Students will be penalized for lab activities that are missed and not made up.

Projects. Most of the outside work in this course will go into the projects. We expect between five and eight projects, including

1. Regular expressions
2. Language models
3. Edit distance
4. Parsing
5. Hidden markov models and POS tagging

Electronic course organization. Course material (assignments, quizzes, slides, videos, etc) can be found on Schoology; additionally, the course schedule and some of the material can be seen through a course website I have made which presents the course as in a calendar format. Unless otherwise noted, assignments are to be submitted through Schoology.

I use the "Gradebook" feature on Schoology only to communicate scores on individual assignments and tests. I do **not** use the Schoology gradebook for my official record-keeping for scores, for calculating semester scores, or determining letter grades. Please **ignore** any grade estimate that Schoology gives you for this course.

Grading. The graded elements of this course are in three categories: *Assignments*, comprising smaller pieces like readings, in-class activities, quizzes, and short assignments; *projects*; and *tests*. There will be a midterm near the middle of the semester (exact date to be announced later). The final exam will be taken during our scheduled exam block, Wednesday, Dec 15, 1:30–3:30 pm.

To **pass** this course (receive a grade of D or better), students must perform competently on each goal by completing at least 75% of the the assignments, achieving at least 50% of the points for projects, and having at least a 50% average on the tests.

For students who have met the minimum requirements, their *semester score* is the geometric mean of their scores in these three categories, with projects counted twice. That is, your semester score is

$$\sqrt[4]{Assignments \cdot Projects^2 \cdot Tests}$$

The geometric mean is used because it is self-normalizing: The individual scores will have different scales, but affect the semester score equally.

Letter grades will be determined by score clustering. An estimation of semester grade will be given after the first test and, after that, upon request.

Policies etc

Academic Integrity. Collaboration among students enrolled in the course is permitted on all projects and assignments, besides quizzes. 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 midterm and final must be done completely independently. (If either the midterm or final is given as a take-home, then they will be accompanied with specific rules about what resources are fair to use.) For quizzes that are administered through Schoology, students are permitted but not encouraged to reference their notes and official course materials. Any violations will be handled though the college's disciplinary process. (See also the College's statement below.)

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. Other assignments will not normally be accepted late.

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 Wednesday, Dec 15, 1:30–3:30 pm. Students are not allowed to take the final exam at a different time (except for urgent reasons, approved by the department chair, as per the college's policy), so make appropriate travel arrangements.

Accomodations. If you have a documented need for accommodations, I will have received a letter on your behalf from the Disability Services Office. But *please talk to me* about what accommodations are most useful to you. In particular, if you desire accommodations for test-taking, talk to me a reasonable amount time in advance (say, at least two class periods) so arrangements can be made. (See also the College's statement below.)

Office hours. My *drop-in* office hours this semester are MWF 3:30–4:30pm. You can make an appointment through Calendly; I'm available most of the day on Thursday and sometimes on other days.

Electronic devices. My intent is for class to be an electronic-device-free zone except when we as a class are doing lab activities. However, our text book is available only in electronic form. Accordingly, please use your device (laptop, tablet, etc), *only* for referring to the textbook at note-taking. If you absolutely need to check your phone for something, please discreetly step out in to the hall. **NO TEXTING OR SOCIAL MEDIA USE IN CLASS.**

All this, the Lord willing.

College syllabus statements

THE COLLEGE REQUIRES THAT THE FOLLOWING STATEMENTS BE INCLUDED IN ALL SYLLABI.

ACADEMIC INTEGRITY. The Wheaton College Community Covenant, which all members of our academic community affirm, states that, "According to the Scriptures, followers of Jesus Christ will. . . be people of integrity whose word can be fully trusted (Psalm 15:4; Matt. 5:33-37)." It is expected that Wheaton College students, faculty and staff understand and subscribe to the ideal of academic integrity and take full personal responsibility and accountability for their work. Wheaton College considers violations of academic integrity a serious offense against the basic meaning of an academic community and against the standards of excellence, integrity, and behavior expected of members of our academic community. Violations of academic integrity break the trust that exists among members of the learning community at Wheaton and degrade the College's educational and research mission.

GENDER-INCLUSIVE LANGUAGE. Please be aware of Wheaton College's policy on inclusive language, "For academic discourse, spoken and written, the faculty expects students to use gender inclusive language for human being."

ACCOMMODATIONS.. Wheaton College is committed to providing reasonable accommodations for students with documented learning differences, physical or mental health conditions that qualify under the ADA. Any student needing academic adjustments is requested to contact the Learning and Accessibility Services Office as early in the semester as possible. Please call 630.752.5615 or e-mail las@wheaton.edu for further information.

COVID-19 SYLLABUS STATEMENT.. In accordance with the Wheaton College Face Covering Policy, CDC-approved face coverings are required while attending class. Failure to comply with wearing a face covering will result in dismissal from the class session and an unexcused absence. Multiple violations can lead to dismissal from the class. Student Health Services will officially communicate when a student must be absent from class due to quarantine or isolation. Remote learning will not be offered this fall, and the student is encouraged to coordinate with the instructor any needed adjustments to tests or deadlines. Learning & Accessibility Services will also provide assistance for students in quarantine if necessary.

TITLE IX AND MANDATORY REPORTING. Wheaton College instructors help create a safe learning environment on our campus. Each instructor in the college has a mandatory reporting responsibility related to their role as a faculty member. Faculty members are required to share information with the College when they learn of conduct that violates our Nondiscrimination Policy or information about a crime that may have occurred on Wheaton College's campus. Confidential resources available to students include Confidential Advisors, the Counseling Center, Student Health Services, and the Chaplain's Office. More information on these resources and College Policies is available at http://www.wheaton.edu/equityandtitleIX.

WRITING CENTER. The Writing Center is a free resource that equips undergraduate and graduate students across the disciplines to develop effective writing skills and processes. This academic year, the Writing Center is offering in-person consultations in our Center in Buswell Library, as well as synchronous video consultations online [https://www.wheaton.edu/media/writing-center/A-Client's-Quick-Guide-to-Online-Writing-Consultations---Updated-08.15.20.pdf] Make a one-on-one appointment with a writing consultant here [https://wheaton.mywconline.com/].