CSCI 384

Computational Linguistics

Fall 2023 MWF 12:55–2:05 MEY 131

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

Thomas VanDrunen

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

Office: MEY 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.)

Purpose of the course. In addition to the stated topic of *computational linguistics*, this course plays a distinctive role in the CSCI program in that in that it is thoroughly 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 approach 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 Python.
3. Identify the elements of widely-used computational-linguistic 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 society and 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.

Course outline. The course topics are eclectic in that some are based on their linguistic or mathematical background, some by their algorithms, and some by their applications. They are broadly categorized into traditional and machine-learning methods, which constitute two halves of the semester. See the course website for a schedule.

I. Traditional methods

   A. Introduction (2 days)
      - Python
      - NLTK and other libraries
      - NLP terminology

   B. Regular expressions (2 days)
      - Definition, rules, and limitations
      - Regex-based chatbots

   C. Edit distance (1 day)

   D. Information theory (2 days)
      - Concepts and terminology
      - Noisy channel model
      - Compression
      - Character-based language models

   E. Statistical language models (4 days)
      - n-grams and other language statistics
      - Lexical language models
      - Smoothing and interpolation
      - Applications and evaluation

   F. Part-of-speech tagging and hidden Markov models (3 days)
      - Lexical categories
      - Hidden Markov model definition
      - Viterbi algorithm

   G. Parsing (3 days)
      - Context-free grammars
      - CKY parsing algorithm

   H. Lexical semantics (2 days)
      - Word senses and relationships
      - WordNet

II. Machine-learning methods

   A. Basic training in machine learning (2 days)
      - Concepts and tasks
      - Bag-of-words model

   B. Naive Bayes classification and sentiment analysis (3 days)
      - Bayes’s theorem
      - Sentiment analysis

   C. Stylometry and authorship attribution (3 days)
      - Concepts and problem statement
      - Techniques and practice

   D. Vector semantics and word embeddings (3 days)
      - The idea of word vectors
      - Word2vec and other approaches

   E. Neural nets and related models (3 days)
      - Neural net basics
      - Recurrent neural nets
      - Long short-term memory

   F. Machine translation (2 days)

   G. Large language models (3 days)

Course procedures

How we do this course. Most topics (two to four class periods of material) include a reading from the textbook, a couple days in class working out the concepts, a day in lab experimenting with the an application of the topic, a programming assignment, and a quiz or two to be taken through Canvas. Occasionally there will be a reading from a supplemental source or an additional, non-programming assignment (such as a response to a reading).

The programming assignments will come rapidly at the beginning of the semester and gradually taper off in frequency (but become harder). For the last few topics we'll focus on comprehending the readings rather than coding up the algorithms.

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 libraries we will use are `nltk`, `numpy`, and `sklearn`.

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.

Programming assignments. Much of the outside work in this course, especially in the first half of the semester, will go into the programming assignments. The expected programming assignments are a Python warm-up, solving regular expression problems, implementing edit distance, implementing the Huffman encoding, implementing an n-gram language model with Good-Turing smoothing and linear interpolation, implementing a POS-tagger using HMM, implementing a CKY parser, implementing a naïve Bayes classifier, implementing word2vec embedding, and solving an authorship attribution problem. Several of these involve using dynamic programming.

Tests. There will be a midterm and a final. The midterm is expected to be held in class on Wed, Oct 9 (subject to change). Our final exam block is Tuesday, Dec 12, 10:30am–12:30 pm. The final exam will be "mostly-non-cumulative."

Grading. The graded elements of this course are in three categories: *The midterm*, *the final*, *programming assignments*, and *other*. The "other" category comprises non-programming assignments, labs, and quizzes.

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

For students who have met the minimum requirements, their *semester score* is the geometric mean of their scores in these four categories.

$$\sqrt[4]{Midterm \cdot Final \cdot ProgrammingAssignments \cdot Other}$$

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 midterm and, after that, upon request.

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

Policies etc

ACADEMIC INTEGRITY. Collaboration among students enrolled in the course is permitted on all programming assignments. Certain other assignments also explicitly allow collaboration. Obtaining solutions to programming 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. Students should not use Copilot, ChatGPT, or similar tools for programming or written assignments. For quizzes that are administered through Canvas, students are *permitted but not encouraged* to reference their notes and official course materials; students should not confer with classmates or others on quizzes. 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 programming assignments, which may be divided in whole-number units among the assignments (for example, one assignment one day late and another assignment two days late). Beyond that, late assignments will not normally be accepted. Please inform the instructor when you are using a late day on the day that the assignment is due, or earlier. Quizzes and other assignments will not normally be accepted late. Assignments will typically be due at 11:59pm on their due date. Quizzes will typically be due at 11:59pm on a Tuesday or Thursday.

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 Tuesday, Dec 12, 10:30am–12:30 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.

ACCOMMODATIONS. 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; typically I have appointments available throughout the day on Tuesdays and in the morning on Thursdays.

SENDING PYTHON BY EMAIL. Be aware that the college's email system *silently rejects* any email with a `.py` attachment. This means that if you send me a Python file by email, I will not know that you tried to send it, and you will not be alerted that it didn't go through. To send me Python code by email, such as to ask for help on an assignment, please either paste the code directly into the body of the message (preferred) or change the extension on the file (e.g., `.txt`) before attaching the file. Please *do not* send a screenshot.

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, since our text book is available only in electronic form, on certain days I will allow you to use a device in class *only for referring to the textbook*. Apart from that, no electronic device should be visible any time during class. 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.** Note that this policy also applies to lab sessions. **Students will receive 0 for a lab activity if their phones or other devices are visible any time during a class session in the lab.**

All this, the Lord willing.

College syllabus statements

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

The "Academic Information" website referred to below is found and `https://catalog.wheaton.edu/undergraduate/academic-policies-information/academic-information/`

ACADEMIC INTEGRITY. (See "Integrity of Scholarship" on "Academic Information" website.)
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.

ACCOMMODATIONS. (See "Learning and Accessibility Services" on the "Academic Information" website).
Wheaton College is committed to providing access and inclusion for all persons with disabilities, inside and outside the classroom. Students are encouraged to discuss with their professors if they foresee any disability-related barriers in a course. Students who need accommodations in order to fully access this course's content or any part of the learning experience should connect with Learning and Accessibility Services (LAS) as soon as possible to request accommodations `http://wheaton.edu/las` (Student Services Building -Suite209, `las@wheaton.edu`, phone 630.752.5615). The accommodations process is dynamic, interactive, and completely free and confidential. Do not hesitate to reach out or ask any questions.

BEHAVIOR POLICY. (See "Classroom Demeanor" on the "Academic Information" website).

GENDER-INCLUSIVE LANGUAGE. (See "Gender Inclusive Language" on the "Academic Information" website).
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."

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. Make a one-on-one appointment with a writing consultant here [`https://wheaton.mywconline.com/`].