Ling/CS/Info/Ista 439 & 539
Jackson
Spring 2024

Statistical Natural Language Processing

Course Information

This 3-credit course introduces the key concepts underlying statistical natural language processing. Students will learn a variety of techniques for the computational modeling of natural language, including: n-gram models, smoothing, Hidden Markov models, Bayesian inference, expectation maximization, the Viterbi algorithm, the Inside-Outside algorithm for probabilistic context-free grammars, and higher-order language models. *(from the course catalog)*

Course objectives

In this course, we will …

- cover machine learning basics and text classification algorithms, such as …
  - naive Bayes
  - logistic regression
- explore a range of important natural language processing (NLP) topics, such as …
  - word representations (ex. embeddings)
  - sequence labeling (part of speech tagging, shallow parsing/chunking, etc.)
  - structured prediction (chart-based parsing, transition-based dependency parsing, etc.)

Learning outcomes

Students will be able to…

- carry out a variety of natural language processing (NLP) tasks[1]
- compare techniques for word and document representations[1]
- implement a subset of the algorithms and architectures covered in this class[2]
- understand an NLP tool or approach well enough to explain it to others.[2]

[1] Relates to Linguistics Department's UG Program Outcome 1.
[2] Relates to Linguistics Department's HLT Program Outcomes 1, 2, & 3.

HLT learning outcomes addressed in this course

1. Students will demonstrate programming skills for the workplace.
2. Students can use fundamental algorithms & concepts of Natural Language Processing.
3. Students can use tools & packages typically used in Natural Language Processing.

Prerequisites

- Programming competency (at the level of ISTA 130 or higher)

Instructor

name	Eric Jackson
email	ejackson1@arizona.edu
hours	Mondays 2:00pm–4:00pm (Arizona time, UTC-7) in person (COMM 114A) and online via Zoom at `https://arizona.zoom.us/j/89529422793` (passcode 736013), and by appointment.

Requirements

Locations and times

This is an in-person course, but we will make use of online tools for certain learning activities.

Our in-person class sessions will be held on Mondays and Wednesdays (except for university holidays), 11:00am to 12:15pm, in the Electrical and Computer Engineering (ECE) building, room 107. Our first class session will be Wednesday, January 10th, and our last class session will be Wednesday, May 1st, for a total of 30 sessions.

Because I'm teaching both in-person courses and fully online courses this term, my office hours will be offered both in-person and online; regardless of your class modality, you may attend office hours in either format. Office hour times, location, and an online link can be found on the instructor introduction page and on the syllabus.

Please see the course D2L page for important dates and further information.

Schedule of topics

The course includes an introductory unit, to get everyone working in a uniform development environment, followed by four major topics: machine learning basics, word and phrase representation, sequence tagging, and structured prediction. If time permits, we may see one additional topic.

Authoritative due dates will be listed for each assignment in D2L. Check the D2L course calendar to make sure you don't forget or miss a deadline.

Readings

A draft version of the textbook used in this course is available for free on-line.

> *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Jurafsky and Martin, `https://web.stanford.edu/%7Ejurafsky/slp3/`

Additional readings may be available digitally in the course D2L site. In addition to the course textbook and any posted readings, students may find the following resources useful:

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing.* Springer.

Nielsen, Michael. 2015. *Neural Networks and Deep Learning.* Springer. `http://neuralnetworksanddeeplearning.com/`

Bird, Steven, Ewan Klein, and Edward Loper. 2015. *Natural Language Processing with Python.* `https://www.nltk.org/book/`

Géron, Aurélien. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* (2nd ed.) O'Reilly. `http://neuralnetworksanddeeplearning.com/`

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016 *Deep Learning.* MIT press. `https://www.deeplearningbook.org/`

Attendance and participation

Students are expected to actively participate in the course by attending in-person class sessions, reading the assigned readings, completing the assigned learning activities and programming homework, and engaging with the instructor and other students in the course forum. You're all adults, and you're responsible to invest time in your own learning.

If the content of a lecture is not clear, you are expected to send a question to the instructor by email, meet with the instructor in regular office hours or arrange another time to meet, or post a question for clarification on the course forum.

The preferred place to ask questions about the course is on the on the course forum at `https://forum.hlt.arizona.edu/#narrow/stream/47-ling-539-sp2024`, not on D2L. If *you* have a question, it's possible that someone else has a similar question. Having the question and answer on the forum means that everyone benefits from it. The course forum is also where you will also find course announcements.

For emergencies or for personal matters that you don't wish to put in a private post, please email the instructor. If you cannot arrange to speak with me during my regular office hours and are unable to discuss course-related issues over email or in a private message on the course forum, you should contact me by email to set up an individual appointment via Zoom.

For planning purposes, please note that my working hours are generally M–F from 8AM–6PM (MST). Typically, you can expect a response from me within a day.

Assignments and grading

All of the assignments have been designed to aid your learning and retention of course material. I expect everyone to attempt all of them, to gain the most from the course. The due date for each assignment will be posted with the assignment in D2L. All times will be given in Arizona time (Mountain Standard, GMT-7). Accepting late work would mean that I cannot give timely feedback to the rest of the class on the issues in that assignment. **Except for university-approved reasons listed below, late work will not be accepted.**

Review and Mastery activities are an opportunity for you to practice applying the new concepts we learn in class, in a context where it's perfectly fine to not get things right. These online activities provide immediate feedback to you, to help you see where you might need to review class notes or readings, or ask for clarification.

Graded programming assignments will be given via GitHub Classroom, accessed through the course website (D2L). These programming assignments will be completed in a uniform development environment using Python in a Jupyter Notebook, and graded using NBGrader. Test cases will be provided to indicate where your solutions have problems. ***Note: working on the Jupyter Notebook outside of this development environment may introduce errors in the NBGrader grading, and is not recommended***

For undergraduates enrolled in 439 only, your overall course grade will be calculated based on this weighting of assignments:

type	number	total
programming assignments	4	70%
review and mastery	11	30%

Additional work for those enrolled in 539:
A **private Kaggle competition** will provide an opportunity to apply the techniques learned in class to a set of real-world data. For this assignment, students will complete (a) a blog post describing your approach to the problem, (b) a GitHub repository with the source code for your solution, and (c) a submission to the class competition on Kaggle.com.

Your overall course grade will be calculated based on this weighting of assignments:

type	number	total
programming assignments	4	65%
review and mastery	11	15%
class Kaggle competition	1	20%

Course grades will be calculated based on the following percentages:

Grade	**Point Range**		
A	90	–	100
B	80	–	89
C	70	–	79
D	60	–	69
E	0	–	59

Technology

To complete your programming assignments, we recommend that you use a laptop or desktop with $\geq$ 8GB of RAM. All assignments and tutorials will be presented using a uniform Linux-based development environment which students will learn to configure during the first unit of class. To complete your assignments, you will need . . .

- A Linux desktop environment such as Ubuntu 22.04 (can be installed as a virtual machine)
- A GitHub account
- Docker (installed on your course-specific development environment)
- A modern web browser (Firefox or Chrome/Chromium)

Student Work Policy and Collaboration

Students are encouraged to discuss problems and general approaches for solutions, but everyone must turn in their own work. You may not submit assignments that are substantially the same as your classmates. Changing variable names is not sufficient to justify submitting work which is taken wholly from some other source, whether that is a classmate or a website. You may not submit work which was simply copied and pasted from generative AI. The purpose of assignments is to get **you** to think about the issues, and having ChatGPT craft an answer for you does not bring any benefit **to you** in terms of your own learning.

Assignments that seem suspiciously similar, those that seem to be **plagiarized**, or those that seem to have been produced using generative AI, will be forwarded to the Dean of Students office in accordance with the Code of Academic Integrity (linked below). Please be a responsible adult and don't run the risk of losing credit for an assignment by copying, by allowing others to copy from you, *or* by having ChatGPT do your assignment for you.

Generative AI is a tool, just like a calculator is a tool for doing math, or a bicycle is a tool for transportation. In some contexts, being able to use a calculator is an important skill—while in other contexts, like when you're taking a math test to see whether you know basic math facts, solely using a calculator short-changes your education. A bicycle is a tool that allows us to get from one place to another faster and more efficiently than running—but if you're going to be tested in your time for a 5k run, it won't help you to train for running solely by riding a bicycle. We will probably all need to know how to use generative language models for tasks at some point, but having one write your homework or forum posts for this class is not appropriate. Put in the thinking yourself, and reap the mental benefit for yourself.

University boilerplate

All of the following items are required by the university to be included on syllabi. If you find something here that is surprising or unexpected, please bring it up with me as soon as possible.

By way of a brief summary:

Disabilities If you have a disability that affects how you will need to do the work in this class, please let me know *within the first week of class*.

Academic Code of Conduct Cheating and plagiarism are not remotely acceptable in any way. You are responsible for knowing whether your own behavior qualifies as plagiarism, and whether your use of AI is inappropriate. Disruptive behavior in class—which here includes audio, video, or text on any of our course websites or by email—is not acceptable. Please be respectful of others.

Sensitive Material This is a university and you are adults. It is possible that we may touch on topics that some students could find sensitive during the semester. Given the focus of this course, this seems unlikely, but I alert you nonetheless.

Health & Wellbeing

The university has a specific site for COVID information: `http://covid19.arizona.edu`. If you are experiencing personal or financial challenges from any health-related issue, let me know as soon as you can if we need to make accommodations, and please stay safe.

The semester ahead may come with ups and downs in both physical and mental health, but there are lots of ways to support yourself. Eat well, get regular exercise, and don't neglect things like self-care, talking with friends and family, or getting a fresh perspective from a supportive group. Stress is a normal part of life and may even motivate you sometimes, but chronic or overwhelming stress can affect your physical and mental health and wellbeing. Pay attention to your personal signs that you're overly stressed, like changes in your mood, appetite, sleep, behavior, or new physical symptoms (aches, pains, etc.) that interfere with school and daily life. If you notice these signs or have questions about helpful resources, I welcome you to talk with me. You can also visit `caps.arizona.edu/mental-health` for mental health tools and resources.

Mental Health & Wellness Resources

- **Health & Wellness:** Campus Health provides quality medical, mental health, and wellness services for students. Visit `health.arizona.edu` or call 520-621-9202 (520-570-7898 for help after hours)
- **Mental Health:** Campus Health's Counseling & Psych Services offers a range of mental health support tools and services like self-care strategies, peer support, groups and workshops, and professional mental health services. Visit `caps.arizona.edu/mental-health` or call CAPS 24/7 at 520-621-3334 to learn more.
- **Crisis Support:**

  Suicide & Crisis Lifeline: call 988 Crisis Text Line: text TALK to 741-741 Visit `preventsuicide.arizona.edu` for more suicide prevention tips and resources

Absence and Class Participation Policy

Attendance in an all-online course is not evaluated like attendance in an in-person course. For this course, attendance will be represented by active reading, completion, and participation in online course activities, including loading/viewing materials and completing activities posted on D2L, OpenClass, our course forum, and any other related websites.

The UA's policy concerning Class Attendance, Participation, and Administrative Drops is available at: http://catalog.arizona.edu/policy/class-attendance-participation-and-administrative-drop

The UA policy regarding absences is that any sincerely held religious belief, observance or practice will be accommodated where reasonable, http://policy.arizona.edu/human-resources/religious-accommodation-policy.

Absences pre-approved by the UA Dean of Students (or Dean Designee) will be honored. See: https://deanofstudents.arizona.edu/absences

Classroom Behavior Policy

To foster a positive learning environment, students and instructors have a shared responsibility. We want a safe, welcoming, and inclusive environment where all of us feel comfortable with each other and where we can challenge ourselves to succeed. To that end, our focus is on the tasks at hand and not on extraneous activities.

Students are asked to refrain from disruptive conversations with others in the course, including on asynchronous course platforms. Students observed engaging in disruptive activity will be asked to cease this behavior. Those who continue inappropriate behavior will be removed from that venue and may be reported to the Dean of Students.

Threatening Behavior Policy

The UA Threatening Behavior by Students Policy prohibits threats of physical harm to any member of the University community, including to oneself. See http://policy.arizona.edu/education-and-student-affairs/threatening-behavior-students.

Accessibility and Accommodations

At the University of Arizona, we strive to make learning experiences as accessible as possible. If you anticipate or experience barriers based on disability or pregnancy, please contact the Disability Resource Center (520-621-3268, https://drc.arizona.edu/) to establish reasonable accommodations.

Code of Academic Integrity

Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of independent effort unless otherwise instructed. **If you use a code snippet that you came up with from discussions with a classmate, that you found online, or even that you got from a large language model, it's important to cite where it came from, whether that source was Sally Classmate, GitHub.com, stackexchange.com, or ChatGPT.**

Students are expected to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See: `http://deanofstudents.arizona.edu/academic-integrity/students/academic-integrity`.

The UA Library provides a helpful learning module for students to understand and avoid plagiarism: `https://libguides.library.arizona.edu/info-strategies/plagiarism`

The UA Library also has resources to guide you to appropriate and safe use of AI and large language models: `https://libguides.library.arizona.edu/students-chatgpt/integrity`

UA Nondiscrimination and Anti-harassment Policy

The University is committed to creating and maintaining an environment free of discrimination; see

`http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy`

Subject to Change Statement

Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.