{LING/CS/INFO/ISTA} 439/539
Statistical Natural Language Processing

Gus Hahn-Powell

Last Revised March 8, 2024

1 Course Information

From the course catalog:

> This 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, Viterbi, Inside-Outside Algorithm for Probabilistic Context-Free Grammars, and higher-order language models.

NOTE: The main programming language used in the course will be Python (3.8).

For more information, see the course overview page.

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 ...

- familiar with a variety of natural language processing (NLP) tasks[1]
- capable of comparing techniques for word and document representations[1]
- tasked with implementing a small subset of the algorithms/architectures covered in this class[2]
- **539 only**: apply what they have learned to their own research or on a topic suggested by the instructor[2]

Credits: 3 units

Prerequisites

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

Locations and Times

This is an asynchronous online course. We will not be meeting in-person. Please see the course D2L page for important dates and further information.

Instructor

Gus Hahn-Powell
Email: `hahnpowell@arizona.edu`

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

Office hours: *See our course page on D2L*
Appointments:

- `https://parsertongue.org/availability/`

Contact

Students should ask all course-related questions on the `ling-539-online-sp2024` channel/stream in the course forum (`https://forum.hlt.arizona.edu`), where you will also find announcements. For emergencies, or personal matters that you don't wish to put in a private post, please email your instructor at `hahnpowell@arizona.edu`.

For planning purposes, please note that **your instructor responds to emails and posted questions Monday & Friday between 9AM and 5PM (MST).** Typically, you can expect a response within a day.

Schedule

Assessments Please check D2L for assessment details and due dates:

- 439: `https://d2l.arizona.edu`
- 539: `https://d2l.arizona.edu`

Readings & lectures Please check the **Content** section of D2L for unit-specific readings and lectures:

- 439: `https://d2l.arizona.edu`
- 539: `https://d2l.arizona.edu`

Technology

This is a fully online class. As such, you will need a stable internet connection to access course content and submit assignments. To complete your 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 development environment which students will learn to configure during the first week of class (instructions will be provided). To complete your assignments, you will need ...

- A Linux desktop environment such as Ubuntu 20.04 (can be installed as a virtual machine on Windows)
- Git
- A GitHub account
- Docker (installed on your course-specific development environment)
- A modern web browser (Firefox or Chrome/Chromium)

Virtual office hours will use Zoom.

Python-specific Assistance

For students entirely new to Python (but not to programming), this free self-paced course can help bring you up to speed quickly:

- `https://arizona.openclass.ai/invite?code=tEgINUEN5ExbEg`

Other Resources

ResBaz Arizona hosts a variety of events providing researchers opportunities to connect with experienced data scientists and engineers. For more information, see their homepage:

- `https://researchbazaar.arizona.edu/#portfolio`

Readings

The primary text used in this course is freely available (digital-only):

> Dan Jurafsky and James Martin (2019). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed. Upper Saddle River, N.J: Pearson Prentice Hall. ISBN: 9780131873216. URL: `https://web.stanford.edu/%7Ejurafsky/slp3/`

Required readings (papers, chapters, etc.) will be provided by the instructor.

Supplemental Reading

In addition to the course textbook and any posted readings, students may find the following resources useful:

- Yoav Goldberg (2017). *Neural Network Methods for Natural Language Processing*. San Rafael, California: Morgan & Claypool Publishers. ISBN: 1627052984. URL: `https://ebookcentral.proquest.com/lib/uaz/reader.action?docID=4843762`
- Michael A. Nielsen (2015). *Neural Networks and Deep Learning*. Determination Press. URL: `http://neuralnetworksanddeeplearning.com`
- Luis Serrano (2020). *Grokking Machine Learning*. Manning Publications. URL: `https://www.manning.com/books/grokking-machine-learning`
- Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media. URL: `https://www.nltk.org/book/`
- Aurélien Géron (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, Inc. ISBN: 1492032646. URL: `https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/`

- Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. URL: `http://www.deeplearningbook.org`

Electronic versions of all the recommended resources listed above are freely to University of Arizona students.

2 Evaluation

Undergraduates

	Assessment	Description
70%	Programming Assignments	3-4 assignments that involve implementing algorithms covering in the class. Test cases will be provided to help you refine your solutions. *The programming language used will be Python.*
30%	Review & Mastery Assignments	Low-risk assessment consisting of guided reviews and questions designed to assist in *retaining and mastering* material covered in class. We will be using the OpenClass platform for these assignments.

Graduates

	Assessment	Description
65%	Programming Assignments	*see above*
15%	Review & Mastery Assignments	*see above*
20%	Class competition	Class-wide text classification competition. Components include a **a)** a publicly accessible blog post summarizing your approach, **b)** a GitHub repository with the source of your solution, and **c)** a submission to the class competition leaderboard.

Due Dates

Assignment	**When?**	**Where?**
Programming Assignments	$\geq$ 1-2 weeks after release	GitHub Classroom (via D2L $\rightarrow$ **Assignments**)
Review & Mastery Assignments	1-2 weeks after release	OpenClass (via D2L nav bar)
Class competition: *topic* (**grads only**)	$\approx$ 3 weeks after announcement	GitHub Classroom (via D2L $\rightarrow$ **Assignments**)
Class competition: *all other components* (**grads only**)	final day of class*	GitHub Classroom (via D2L $\rightarrow$ **Assignments**)

Dates above are only estimates and are thus subject to change.

*Final day of class is earlier than the final day of the term; see the course D2L page for an exact date.

3 Grading

Grades will be posted to the course's D2L site: lmsurl

- 439: `https://d2l.arizona.edu`
- 539: `https://d2l.arizona.edu`

For assignments involving code, You will be provided with a subset of test cases to help you refine your solution before submitting.

The grading scheme is as follows:

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

Grade Disputes

Disputes about grades on a particular project will be entertained for two weeks from the day the project is due, or 1 day before grades are due, whichever is sooner. These will be resolved by re-grading the entire project. Note that this can result in a lower grade in the event that new mistakes are discovered.

No negotiations about individual students' letter grades will be entertained once final grades are assigned, except as permitted by the policy stated above.

Collaboration Policy

Students are encouraged to discuss problems and general approaches for solutions. However, **implementations and the associated documentation for each assignment must be completed individually**. **Copying another person's work (even if it comes from a website) and/or reusing solutions provided by a generative AI technology such as GPT-4 is not permitted and will be treated as a case of academic dishonesty**.

Late Policy

Projects are due electronically via D2L or GitHub Classroom by the stated deadline. Permission for an extension must be granted by the lab instructor *in advance* of the deadline in order to receive full credit for a late submission. The first request by a given student is likely to be granted; the probability decreases with each subsequent request. No project will be accepted once solutions are posted online.

4 University Policies

Classroom Behavior

Students are expected to behave respectfully toward each other and to the instructor and TAs.

The Arizona Board of Regents Student Code of Conduct is here:

- `https://public.azregents.edu/Policy%20Manual/5-308-Student%20Code%20of%20Conduct.pdf`

ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to oneself:

- `http://policy.arizona.edu/education-and-student-affairs/threatening-behavior-students`

Valuing Inclusion and Diversity

The University of Arizona strives to foster inclusive learning environments in which diversity is recognized and valued. This course strives to create a civil and welcoming environment for everyone, including students of diverse ethnic, cultural, linguistic, national, and familial backgrounds and gender identities, ages, abilities, and veteran status. Students and faculty are responsible for creating an inclusive learning environment through respectful and civil discussion in the classroom and inclusive practices in class and group work.

- We will use the names and pronouns as selected and proposed by each individual in the course in acknowledgment of the intricate nature of our identities and in respect of each other's integrity. Students may request a University wide name change through LGBTQ Affairs at and learn more about name and pronoun use at `https://lgbtq.arizona.edu/transgender-resources`.

- If you anticipate or experience physical or academic barriers based on disability, pregnancy, or parenting status, please contact your instructor as soon as issues are known and contact Disability Resources (520-621-3268) to establish reasonable accommodations. For additional information on Disability Resources and reasonable accommodations, please visit `https://drc.arizona.edu` or `https://drc.arizona.edu/workplace-access/pregnancy-accommodations`.
- If you have questions about particular policies concerning gender equity, sexual harassment, or sexual assault, please contact the Office of Institutional Equity[3] or the Dean of Students Office[4].

Respectful and Careful Communication

All communication in this class adheres to the principles of civil discourse. Civil discourse is guided by mutual respect and appreciation. Diversity of knowledge is an asset to class discussions. In all communication, you are expected to be scholarly, professional, and respectful. Constructive criticism in discussion of course concepts is highly encouraged. Mocking and/or bullying are never allowed. To be critical does not exclude being polite. See:

- UA National Institute of Civil Discourse: `https://nicd.arizona.edu`
- UA Dean of Student's page on cyberbullying: `https://deanofstudents.arizona.edu/safety/cyberbullying`

Student Code of Academic Integrity

Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work must be the product of independent effort unless otherwise instructed. **Note that "independent" should be interpreted to mean without the assistance of generative AI such as GPT-4**. Students are expected to adhere to the UA Code of Academic Integrity as described here:

- `https://deanofstudents.arizona.edu/policies/code-academic-integrity`

Confidentiality of Student Records

`https://registrar.arizona.edu/personal-information/student-information`

On Dropping Classes

If you find yourself thinking about dropping this (or any other) class, first make sure that that's what you really want to do. Chatting with the instructor or your academic advisor may help. If you drop within the first week of the term, there will be no notation on your transcript; it will be as though you'd never enrolled. After the second week, a drop will be recorded on your transcript. You will receive a "WP" (withdrawn passing) only if you were

[3] `https://titleix.arizona.edu`
[4] `https://deanofstudents.arizona.edu`

passing the class at the time of your drop. Toward the end of the term, dropping becomes a challenge, because you need to explain to the instructor and to the dean why you were unable to drop the class during the first half of the term. For drop deadlines specific to this compressed format, please see this calendar:

- `https://www.registrar.arizona.edu/dates-and-deadlines`

Subject to Change Statement

The instructor reserves the right to change with advance notice where appropriate the content of the course. This right does not apply to posted grading and absence policies or University Policies.