LING 531
Jackson
Fall 2024

Human Language Technology II

Course information

This intermediate-level course is a continuation of LING 529 and covers the basics of information retrieval, focusing on both search and classification.

Course objectives

This course presents fundamental considerations and concepts for text-based search in the context of a simple keyword search function with Boolean connectives. We'll then refine our methods for effective search—with the goal of returning the *best* results, satisfying a user's actual *information need*—by exploring ways to represent similarity and to use term weighting. We'll round out or view of search by thinking about how to determine *how well* our search application is meeting this goal. In the second half of the course, we'll use some of these concepts that we developed in the context of search to explore methods for document classification, comparing both statistical approaches and vectorspace approaches.

Learning outcomes

Successful students in this course will...

1. use Python and tools commonly used in commercial applications to perform simple but realistic information retrieval tasks.
2. parse and prepare a corpus for search and classification tasks.
3. create an index for searching a corpus.
4. implement similarity and evaluation measures for searching a corpus.
5. implement classification models using statistical and vectorspace methods.
6. compare and evaluate these search and classification methods and the results they yield.

Learning outcomes #1 and #2 relate to Linguistics HLT program outcomes #1 and #3. Learning outcomes #3-6 relate to Linguistics HLT program outcome #2.

HLT learning outcomes addressed in this course

By completion of the HLT program, students will be able to:

1. **write, debug, and document readable and efficient code** in programming languages commonly used to develop, implement, and evaluate Natural Language Processing models, as demonstrated through course projects and a professional internship.
2. **select and apply appropriate algorithms and core concepts** in Natural Language Processing to perform common tasks and solve realistic problems, as demonstrated through course projects and a professional internship.
3. **utilize common tools and libraries** used in NLP by integrating them into course projects and real-world applications or workflows, as demonstrated through course projects and a professional internship.

Prerequisites

Ling 529 or equivalent. Basic to intermediate ability to program in Python is assumed.

Instructor

name	Eric Jackson
email	ejackson1@arizona.edu
office hours	Mondays 10:00am–12:00pm (Arizona time, UTC-7) and by appointment, in-person in Communications 114A, online via Zoom at https://arizona.zoom.us/j/84420158691 (passcode 074337)
appointments	`https://calendly.com/meet_with_eric/`

Students should ask all course-related questions in the course forum, 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 ejackson1@arizona.edu.

For planning purposes, please note that I respond to emails and questions posted on the forum **Monday through Friday from 9am to 5pm MST**. I am not available for course-related interaction on weekends. Typically, you can expect a response from me within one working day.

Topics and schedule

The schedule below should be interpreted as follows:

- For readings, you should complete the indicated reading *before* the week begins. New materials go live in D2L on Monday mornings.
- Review & Mastery activities and forum posts (from the week's material) are due the following Monday morning at 9am (Arizona time). Exceptionally, the final forum

post becomes available Monday and is due on that Thursday (ie, you won't have the weekend to work on it).

- Programming assignments are all due Tuesday at noon (Arizona time) in the following week.
- Each topic has associated videos and Jupyter Notebooks; you should go through these for the week indicated.

The first week, last week, and Thanksgiving week (11/28 & 11/29 are holidays) are all treated as **half weeks** where there are reduced expectations for workload. An asterisk * indicates these short weeks and short assignments, "R&M" is a *Review & Mastery* activity, "Disc" is a *forum discussion post*, and "HW" is a programming assignment, to be explained below.

Week	Dates	Topic	Reading	Notebook	Due
1	10/17-10/20	Overview*	IR ch.1	#1	R&M 1, Disc 1
2	10/21-10/27	Indexing	IR ch.2	#2,3	HW 1*, R&M 2, Disc 2
3	10/28-11/3	Similarity	IR ch.3	#4	HW 2, R&M 3, Disc 3
4	11/4-11/10	Weighting	IR chs.4,6	#5	HW 3, R&M 4, Disc 4
5	11/11-11/17	Evaluating	IR chs.7,8	#6,7	HW 4, R&M 5, Disc 5
6	11/18-11/24	Classifying	IR ch.13	#8	HW 5, R&M 6, Disc 6
7	11/25-12/1	Naive Bayes*	IR ch.14	#9	HW 6, R&M 7, Disc 7
8	12/2-12/8	Rocchio/kNN	none	#10	HW 7*, R&M 8, Disc 8
9	12/9-12/12	Review*	none	#11	HW 8, Disc 9

Readings

Draft versions of the course textbooks are available for free online. Assigned readings come from the first; the second is recommended for reference and should be familiar from HLT 1.

- *Introduction to Information Retrieval*, Manning, Raghavan, and Schütze, `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`
- *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Jurafsky and Martin, `https://web.stanford.edu/%7Ejurafsky/slp3/`

Required assignments and course grading

	Assessment type	Description
60%	Programming assignments *(6 regular @ 8% each, 2 small @ 6% each)*	Simple programming exercises in Jupyter Notebooks designed to deepen your understanding of concepts and techniques covered in each unit. Test cases will be provided to help you refine your solutions.
30%	Review & Mastery activities *(8 assignments, graded only for completion)*	Low-risk assessment consisting of a guided review and questions each week. These are designed to help you **retain and master** material covered in each unit.
10%	Forum interaction *(9 posts graded for completion)*	You will be given an opportunity to choose a topic to write on, with no length requirement, based on the week's reading and topics. You are encouraged to read and respond to other students' posts, as well.

Due dates and times for all assignments will be listed with each assignment in D2L. Submission of your programming assignment will involve submitting your Jupyter Notebook file in the appropriate assignment item in D2L (ie, we won't be using GitHub Classroom). For Review & Mastery activities, your progress and times are tracked in the OpenClass.ai system and synchronized in D2L after the activity due date. Forum posts will be made on the course forum (not in D2L) in a distinct channel for each unit, at `https://forum.hlt.arizona.edu/`. I will manually check for posts and update forum post credit in D2L during my normal working hours. Feedback for programming assignments and grades for each item will be posted in D2L. **Except for accommodations that are set out in general university policies (see below), late work will otherwise not be accepted.**

If you are struggling with a topic or an assignment, please contact me for help well before the due date. I want to help you understand and master this content, but I can only do that if you communicate with me about problems that you encounter. Since this is a 3-credit course compressed into half a semester, it's important to quickly address problems of understanding and keep up with assignments.

Technology

You'll need to access to a linux-like environment including Python 3 and Jupyter. Those of you continuing from HLT 1 are free to use your linux virtual machine from that course. If you're working in Linux or in macOS, you shoould already have access to a working Python installation.

If you prefer to work in Windows, you may install Python 3 and Jupyter in that OS. One way to accomplish this is to use Anaconda, a Python distribution that is free and available for all platforms. You're welcome to use another Python distribution besides Anaconda if you prefer, including one that is preinstalled in Ubuntu or macOS, but it may require some work with Google and `pip` to make sure it does everything that our class notebooks require.

`https://www.anaconda.com/products/individual`

I expect you to have access to Python 3 and Jupyter Notebooks either directly in your system (whether Windows or macOS) or indirectly via Docker, Virtualbox, a dual-boot Linux system, etc. If you have problems with this, please talk to me as soon as possible.

Collaboration with students and with AI

The purpose of this course is to train **your** mind, and to do that, you need to **use** your own mind. You will gain the most benefit from the programming assignments in this course if **you** are the one who has come up with all the code, even if this requires a bit of mental struggle on your part to get it right. **Don't be afraid to struggle for a bit.**

Students are encouraged to discuss problems and general approaches for solutions, but everyone must turn in work that is the product of their own mind. You may not submit assignments that are substantially the same as your classmates, including using someone else's code but simply changing the variable or object names.

If you do feel you need outside help, using portions of code you found online or created with Generative AI is acceptable, but it must constitute no more than 25% of your total code. If you obtain code other than writing it yourself, **you must evaluate it critically and cite where it came from**. Generative AI is a useful tool, like a calculator is a useful tool for doing math, but generative AI for programming is like a calculator that is sometimes completely untrustworthy. You need to know how to perform these programming tasks on your own well enough that you can see where some AI-generated code is partially or completely off the mark, or introduces logic errors even if it runs without runtime errors.

The general principle in all such cases is that the majority of the work you turn in must be new and must be your own. Do your own work, and please ask me in advance if you are unsure whether something will be acceptable or not.

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.