USC Viterbi
School of Engineering

CSCI 499: Natural Language Processing

Units: 4.0
Fall 2024 — MonWed — 10:00-11:50AM

Location: TBD

Instructor: Jesse Thomason

Office: RTH 402
Office Hours: TBD
Contact Info: jessetho@usc.edu

Catalogue Description
Natural language processing; language modeling; deep neural networks for language processing; language technologies and their socioeconomic and societal impacts

Course Description
Natural Language Processing (NLP) is an area of computing research and practice that aims to enable machines to reason over human text and speech. High-profile technologies like ChatGPT brought NLP to the forefront of public discussion both inside and outside academia. But what underpins such technologies? This course will explore how natural language can serve as an interaction medium between users and machines with a focus on the history and development of language models (LMs). Students will become familiar with concepts and methods in NLP like distributional semantics, and see how those concepts feed into the architectural design of modern LMs trained using deep learning, and will get hands-on experience with building and evaluating small-scale LMs. The class will also explore details and variants of the real-world consequences of deploying large-scale LMs and NLP technologies more generally, such as the ethics and harms associated with them.

Learning Objectives
By the end of this course, students will be able to:
- **O1**: Apply key pieces of modern natural language processing pipelines, such as recurrent and Transformer-based sequence-to-sequence models.
- **O2**: Explain concepts underlying natural language processing in their own words.
- **O3**: Identify structures, conventions, and algorithmic details underpinning natural language processing technologies.
- **O4**: Design and carry out a research project that aims to answer a question in natural language processing.

Recommended Preparation
Fluency with Python programming at the level of ITP 216. Knowledge of algorithms and computing at the level of CSCI 270. Knowledge of artificial intelligence at the level of CSCI 360 or knowledge of machine learning at the level of CSCI 467.

Course Notes
Lecture notes will be made available online after each class.

Technological Proficiency and Hardware/Software Required
Students must have access to a computer with the ability to install and control the Python and Pytorch distributions on that machine.

Required Readings and Supplementary Materials
The required reading includes chapters from the textbooks below. All reading material is freely and publicly available online. Class-specific readings are mentioned under the Course Schedule.
- Jurafsky and Martin. “Speech and Language Processing.” 3rd Ed. (J&M) This textbook contains chapters on the fundamentals of natural language processing and is freely and publicly available online through the above link.

Optional Readings and Supplementary Materials
Supplementary Material includes these textbooks (freely and publicly available online):
- Eisenstein. “Natural Language Processing.” This textbook contains an overview of machine learning approaches for NLP and is freely and publicly available online through the above link. ISBN: 9780262042840

- Goldberg. "Neural Network Methods for Natural Language Processing." This textbook provides a deep learning perspective towards NLP and is freely and publicly available online through the above link. ISBN-10: 1627052984

Description of Assignments and How They Will Be Assessed

Homework Assignments (3 assignments; 30% of final grade):
Through coding homework assignments, students will apply key pieces of modern natural language processing pipelines, such as recurrent and Transformer-based sequence-to-sequence models (Learning Objective **O1**). Homework coding assignments will involve implementing core concepts using frameworks like PyTorch, and then training and executing corresponding machine learning models on real natural language processing data. These coding assignments will be graded based on code correctness in terms of producing expected output, as well as through a written report documenting the code design choices and results of the training and evaluation experiments associated with the assignment.

Paper review (10% of final grade):
Students will write a research paper review to explain concepts underlying natural language processing in their own words (Learning Objective **O2**). The course explores topics through a series of assigned readings in the form of research papers and book chapters. Students will select one reading option and submit a two-page summary of that reading. Reviews will be assessed based on answering a small set of questions, to be released clearly and correctly at the time of the paper assignment. In most cases, each question will warrant at minimum a paragraph to answer.

In-Class Quizzes (15% of final grade):
In-class quizzes will evaluate students' ability to identify structures, conventions, and algorithmic details underpinning natural language processing technologies (Learning Objective **O3**). Quizzes will occur in a subset of class periods and will involve questions about recently covered topics. Students are expected to turn in completed quiz sheets in class. Students who have OSAS extensions may turn in the scanned quiz sheet via email to the instructor by the end of the day. Quizzes will be announced at least one week in advance of taking place during class.

Semester Project (40% of final grade):
Students will design and carry out a research project that aims to answer a question in natural language processing (Learning Objective **O4**). Students will work individually or in small teams (e.g., 2-3). The focus of the class project can be research-focused or application-focused. A research-focused project will develop models and analyze data of an existing problem in NLP, or formulate a new problem altogether. An application-focused project will train (possibly only fine-tuning) and deploy NLP models to new application areas, while not necessarily developing any novel research question to be answered. Students will leverage tools, concepts, and techniques presented in the class. The project involves identifying a communication or exploration need that language could resolve, data sources available to inform the problem and method, and the techniques needed to approach it. The grading distribution for deliverables of the course project, as well as expectations for each deliverable, are detailed below. The deliverables include a project proposal (1-2 pages single spaced), a mid-project report (4-8 pages single spaced), a final presentation (timed, with time for questions), and a final report (~8 pages single space for the main document, up to 15 with appendix/figures). The final report will be due on the day of the University-scheduled final examination. Reports will be graded based on clarity and completeness. The project is a total of 40% of the final grade with the following breakdown:

Project Component	**% of Final Course Grade**
Proposal	5%
Midterm Report	10%

Final Presentation	10%
Final	15%
Total % Of Course Grade:	**40%**

Details about each project deliverable follow:

- The *project proposal* (about 2 pages) should outline the type of project (research-focused or application-focused), and then answer the following questions clearly in a sentence and/or a few paragraphs each, as appropriate: What are you trying to do? Articulate your objectives using absolutely no jargon. How is it done today, and what are the limits of current practice? What is new in your approach, and why do you think it will be successful? Who cares? If you are successful, what difference will it make? What are the risks? What could go wrong, and how will you pivot early on if that happens? How much will it cost? That is, what resources will you need in terms of time and computation? Are these reasonable for a semester, and what access do you have? Identify two milestones along the way to your finished project.
- The *project midterm report* (about 5 pages) should cover these questions in detail: What are you trying to do? Articulate your objectives using absolutely no jargon. Who cares? If you are successful, what difference will it make? Additionally, it should cover related work in detail, as well as document the proposed method and what is new in that approach, as well as how you plan to conduct experiments, what hypotheses these experiments test, and how you will evaluate the quality of the results. If you have preliminary results to report, they should be discussed here. Additionally, you should revisit your milestones from the proposal, document your progress with respect to these, and propose updated milestones for the remainder of the semester.
- The *project final presentation* will be given by your team in class and graded on clarity, completeness, and presentation quality. Each member of the group should present for about equal time. Your presentation should convey: the main takeaway you have reached or aim to reach with your project; a brief motivation which can include some background; the key insight that overcomes the problems presented in the motivation with what was done before this project; the technical details behind the insight; and an overview of what remains to be done before the final report.
- The *project final report* (about 8 pages) should answer all the following questions in detail, and is expected to include revised content from the midterm report based on feedback your team received at that time.
- What were you trying to do? Articulate your objectives using absolutely no jargon. Who cares? If you were successful, what difference would it make? How is it done today, and what are the limits of current practice? What is new in your approach, and why did you think it would be successful? How did you conduct experiments? What hypotheses did those experiments test? How did you evaluate the quality of the results? What are the results of your experiments? What did you discover? Were you able to support your hypotheses? What are the main takeaways of your project? What would you do next if you wanted to keep working in this space? What new questions can you formulate, given the work this semester?

Participation

Class participation (5% of final grade):

To encourage students to explain concepts underlying natural language processing in their own words (Learning Objective **O2**), we will evaluate each student's engagements in course discussions during class and through online platforms like Piazza. Points are earned by asking insightful questions during lecture and project presentations, volunteering to present group work results during lecture activities, providing details and constructive answers to online questions on platforms like Piazza, and otherwise contributing to productive class discussions.

Grading Breakdown of All Course Assessments

Assessment Tool	% of Grade
Homework Assignments (x3)	30%
Paper review	10%
In-class quizzes	15%
Final Project	40%
Class participation	5%
TOTAL	100%

Assignment Submission Policy

Assignments may be turned in until 11:59pm on the due date, after which they are considered 1 day late for every additional 12am-11:59pm period that passes. Brightspace assignments will be created through which to upload assignment material.

Course-Specific Policies

The course will employ a Late Day Token system that enables some flexibility with homework and project deliverables that are not in-class presentations. Note that, regardless of the use of Late Day Tokens, any assignment turned in 10 days late or more will receive an automatic zero.

The course will allow for a budget of 5 Late Day Tokens per student. These tokens can be expended on homework assignments, the paper review, and project deliverables (NOT quizzes or presentations) to extend the deadline, one day at a time, for a student without incurring a late penalty. These tokens should be used with no justification or explanation for taking the late time required (i.e., you do not need to explain your reason). Going over budget (e.g., turning things in late with no Late Day Tokens to expend) will incur grade penalties of 5% per day late. To ensure reasonable grading turnarounds and discussions of solutions, any assignment turned in 10 days late or more will receive an automatic zero regardless of the use of Late Day Tokens.

For project teams, Late Day Token expenditures are on a per-student basis (i.e., if a team of 2 turns in their midterm report one day late, a member expending a Late Day Token will receive a 0% late penalty, while a member not expending a Late Day Token will receive a 5% late penalty).

There are no refunds for late days: unused late days cannot be converted into credit of any kind.

Academic Integrity

Unless otherwise noted, this course will follow the expectations for academic integrity as stated in the USC Student Handbook. The general USC guidelines on Academic Integrity and Course Content Distribution are provided in the subsequent "Statement on Academic Conduct and Support Systems" section.

For this class, unless specifically designated as a 'group project,' all assignments are expected to be completed individually.

Violations of academic integrity will be taken seriously and result in a formal report to the Office of Academic Integrity as well as a reduction of the associated assignment grade, potentially to zero.

Please ask the instructor or TA(s) if you are unsure about what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

You may not record this class without the express permission of the instructor and all other students in the class. Distribution of any notes, recordings, exams, or other materials from a university class or lectures —

other than for individual or class group study — is prohibited without the express permission of the instructor.

Use of Generative AI in this Course

Generative AI is not permitted: Since creating, analytical, and critical thinking skills are part of the learning outcomes of this course, all assignments should be prepared by the student working individually or in groups as described on each assignment. Students may not have another person or entity complete any portion of the assignment. Developing strong competencies in these areas will prepare you for a competitive workplace. Therefore, using AI-generated tools is prohibited in this course, will be identified as plagiarism, and will be reported to the Office of Academic Integrity.

Additionally, note that generative AI tools are trained, often without appropriate license, on text and images from folks whose intellectual property you do not own and may not license. The text and images that are generated by such tools are inherently plagiarized content; by the end of this course, we hope that you have a clear understanding of why that is.

Course Evaluations

The course will follow the standard protocol for end-of-semester course evaluations. Time will be taken out of the final or penultimate week of class to enable students to complete these evaluations during class time under proctoring by a student volunteer. The instructor and TAs will not be present during the completion of these evaluations during class time, in accordance with University policy.

Course Schedule

All deliverables are due at 11:59pm PT of the due date supplied here. Due dates will correspond to the last day of the week on which the class is held.

	Topics/Daily Activities	Readings/Preparation	Deliverables
Week 1	Introduction and Course Overview Text Processing: Tokenization, Syntax, Semantics	J&M Chap 1,4	Homework 1 *released*
Week 2	n-Gram Language Models	J&M Chap 3	
Week 3	Optimization and Neural Networks	J&M Chap 5 & 7	Homework 1 *due*
Week 4	Distributional Semantics and Word Embeddings	J&M Chap 6	Project proposal *due* Homework 2 *released*
Week 5	Recurrent Neural Networks	J&M Chap 9	
Week 6	Sequence-to-Sequence Modeling	J&M Chap 8 & 9	Homework 2 *due* Homework 3 *released*
Week 7	Transformer Language Models	J&M Chap 10	
Week 8	Masked Language Modeling and Finetuning	J&M Chap 11	Homework 3 *due*
Week 9	Large Language Models (LLMs) Prompting, Instruction tuning, In-context learning	J&M Chap 12	
Week 10	Language Processing and Society	Hovy & Spruit, 2016	Paper review *due*
Week 11	Project Discussions - Flipped Classroom		Project midterm report *due*
Week 12	Multimodal and Multimodal Language Models	Bloom-BigScience Clip - Radford et al., 2021	
Week 13	Cutting Edge Topics in NLP		
Week 14	Student Project Presentations		Project Presentations *due*
Week 15	Student Project Presentations		Project Presentations *due*
FINAL	**Project Final Report**		Project Final Report *due* on the university-scheduled date of the final exam

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Course Content Distribution and Synchronous Session Recordings Policies
USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each

course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)

A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.