Ling 4400: Introduction to Natural Language Processing

Course Number: LING-4400 (Fall, 2024)
Time: Monday and Wednesday, 14:00 - 15:15
Location: Maguire, 103

Instructor: Ethan Wilcox
Office: Poulton 248 (Office Hours: Wednesday, 3:30 - 4:30)
Email: ethan.wilcox@georgetown.edu
Website: https://wilcoxeg.github.io/

Teaching Assistant: Lauren Levine
Office Hours: Monday, 12:30 - 1:30, Lauinger 2nd floor
Email: lel76@georgetown.edu

Summary: This course will introduce students to the basics of Natural Language Processing (NLP), a field that combines linguistics and computer science to produce applications, such as generative AI, that are profoundly impacting our society. We will cover a range of topics that form the basis of these exciting technological advances and will provide students with a platform for future study and research in this area. We will learn to implement simple representations such as finite-state techniques, n-gram models and basic parsing in the Python programming language. Previous knowledge of Python is not required, but students should be prepared to invest the necessary time and effort to become proficient over the course of the semester. Students who take this course will gain a thorough understanding of the fundamental methods used in natural language processing, along with an ability to assess the strengths and weaknesses of natural language technologies based on these methods.

This course fulfills the college's *Quantitative Reasoning and Data Literacy* (QRDL) requirement as part of the Core Curriculum.

Learning Outcomes:

- Students will learn how to write code in Python
- Students will learn how to debug Python code using a debugger
- Students will become familiar with the theoretical fundamentals of programming
- Students will become familiar with the fundamental techniques used in natural language processing, including string manipulation, finite state techniques, language modeling, and vector space models
- Students will gain practice thinking critically about the behavior, abilities, and limitations of current NLP tools, including large language models

Course Schedule

Week	Module	Readings, and Assignments Out & Due
Aug 26, 2024	Introduction	8/28 Assignment 1 (installing Python)

(1 session, Aug 28th)		
Sep 2, 2024	Intro to Python, String Manipulation	9/4 Assignment 1 due 9/4 Assignment 2 (palindrome checker)
Sep 9, 2024	Tokenization, Regular Expressions	Jurafsky & Martin, Chapter 2.1 - 2.7
Sep 16, 2024	Regular Expressions, Finite State Automata	9/16 Assignment 2 due 9/16 Assignment 3 (phone scraper)
Sep 23, 2024	Finite State Automata, Finite State Transducers	9/23 Assignment 3 due 9/23 Assignment 4 (regular expressions)
Sep 30, 2024	Introduction to Probability, Midterm Review	9/30 Assignment 4 due
Oct 7, 2024	**Monday, October 7th: Midterm Exam (in class)** Language Modeling	10/9 Assignment 5 (language modeling) Jurafsky & Martin, Chapter 3.1 - 3.5
Oct 14, 2024 (1 session, Oct. 16th)	Advanced Language Modeling	Jurafsky & Martin, Chapter 3.5 - 3.8
Oct 21, 2024	Intro to Neural Network Language Models	10/23 Assignment 5 due
Oct 28, 2024	Markov Models	10/28 Assignment 6 (part of speech tagging)
Nov 4, 2024	Named Entity Recognition, Introduction to Context-Free Grammars	11/4 Assignment 6 due 11/4 Assignment 7 (named entity recognition)
Nov 11, 2024	Context-Free Grammars and Parsing	
Nov 18, 2024	Intro to Vector Space Modeling	11/18 Assignment 7 due 11/20 Assignment 8 (vector space models)
Nov 25, 2024 (1 session; Nov 25th)	Topic Modeling	
Dec 2, 2024	Classifiers, Wrap-up	12/4 Assignment 8 due
Dec 9, 2024 (1 session; Dec 10)	Final Review	
Dec 18, 2024	**Final Exam (12:30 - 2:30 pm)**	

Prerequisites: There are no prerequisites for this course. In particular, we will assume that students do not have any prior experience with programming. If you *do* have some limited prior experience with

coding, but not with NLP, this course is still likely appropriate for you, however, please speak with Ethan if you have any questions or concerns.

Grade Breakdown

Midterm Exam	30%
Final Exam	30%
Homework Assignments	30%
Participation and Attendance	10%

Participation and Attendance: This course is a challenging, fast-paced introduction to two different topics, computer programming and natural language processing. Given that we will be moving quickly between modules, it is imperative that students attend sessions in person. Students are permitted to miss two course sessions, no questions asked, however, you are expected to make up the material that you have missed. If you need to miss a session due to a planned medical event, a medical emergency, or a family emergency, or an excused religious absence, please send an email to Ethan (and cc Lauren) as soon as you can. For more information regarding attendance and excuses, please see the academic standards section of the Undergraduate Student Bulletin (https://bulletin.georgetown.edu/regulations/standards/)

- Class time will involve personal and group coding sessions, so **please bring your laptop to class!**

Assignments: Over the course of the semester, students will complete eight programming assignments. Assignments will be released shortly after class on the specified day. Assignments are due before the start of class on their specified due date.

- **Submitting Assignments:** Homework submissions will happen through Canvas.
- **Late Work Policy:** Students are expected to submit assignments on time. You are allowed two 24-hour extensions over the course of the semester, no questions asked. Please send Ethan and Lauren an email when you wish to use an extension, so we can keep track. If a personal or medical situation means that you are at risk of submitting multiple late assignments, please contact Ethan and Lauren.
- **Group Work:** The code that you submit as part of homework assignments *must be written by you*. However, working in groups on assignments can be extremely helpful. If you work with a group to come up with *pseudocode* (a high-level plan for how your program will run), or if you work with a group to understand and debug a piece of non-functioning code, that's OK. But you need to implement the pseudocode, or fix the bug yourself!

For more information on the University's honor code, please see https://honorcouncil.georgetown.edu/.

Readings: Most readings from the course will be derived from Jurafsky and Martin, 3rd edition, which can be accessed online here: https://web.stanford.edu/~jurafsky/slp3/. Readings not from this book will be posted, online, in Canvas.

Midterm and Final: The midterm and final exams will not require live coding on a computer, or remembering specific formulas. Rather, they will involve commenting on or correcting code, explaining

why and how code does (or doesn't) work, as well as explaining concepts we've covered in class. The TA will lead practice sessions in advance of each exam.

Use of AI Assistants: Use of AI assistants, such as ChatGPT, Bard, or Claude, as well as coding plug-ins like Copilot, are not explicitly forbidden. These tools are increasingly part of our world, and are based on the technologies we will learn about in this class. Indeed, one of the learning outcomes is to encourage critical thinking about the abilities and limitations of such systems.

With that in mind, a warning: While LLM-based tools can be extremely effective, they suffer from two major drawbacks. First, they often make mistakes. Second, they reduce the amount *you* have to think about and understand the code you run (that's the point!). If you are an advanced coder or an expert debugger, then you will be able to spot the mistakes that these systems make and fix them. However, for beginners, it's quite possible that correcting a buggy LLM-generated algorithm will take more time than writing the algorithm yourself. Furthermore, every time you rely on an LLM to generate something for you, that's one less opportunity you have to solidify the concepts we learn in lectures and to practice coding as a skill. Importantly, you won't be able to use LLMs during the midterm or the final. If you use these systems extensively in the beginning you'll have a much harder time preparing for the exams.

Accommodations: If you have a recognized accommodation through ARC, please contact Ethan and Laura. More information about accommodations and support, including student-athlete support, can be found on the ARC website (https://academicsupport.georgetown.edu/)