CSC-496/696: Natural Language Processing and Text as Data

Department of Computer Science, American University

Fall 2024

Last Updated: September 12, 2024

Instructor:	Patrick Wu
Email:	`pwu@american.edu`
Instructor's Office:	DMTI 108A
Class Location:	DMTI 114
Class Time:	Tuesday/Friday, 4:05 PM - 5:20 PM
Instructor's Office Hours:	Thursday, 3:00 PM - 5:00 PM or by appointment

Course Information

Course Summary

In recent years, the quantity and quality of text data have dramatically increased, driving a surge in demand for tools to effectively analyze this data and unlocking new approaches to language modeling. This course introduces natural language processing (NLP) and text as data, focusing on practical applications and problem-solving. Using Python, the class covers fundamental NLP techniques, including text preprocessing, text classification, topic modeling, text embeddings, benchmarking and evaluation, transfer learning, and large language models (including generative LLMs). Applications involve working with data such as product reviews, tweets and other social media posts, news articles, and open-ended surveys to answer questions in domains like the social sciences and business.

Prerequisites

CSC-208 (Introduction to Computer Science II) is technically a prerequisite for the course. **This is really a stand-in for basic programming knowledge.** CSC-468 (Artificial Intelligence) is recommended but not required. Knowledge of the basics of machine learning (supervised and unsupervised learning, classification, train/validation/test splitting, etc.), linear algebra, and probability theory is recommended but not required.

Books

You will not need to buy any of the textbooks. The books can be accessed online, through the University Library, or through Canvas. Readings may also come from other online sources. One of the tricky aspects of this class is that the field continues to evolve as the class goes on. Textbooks written even a few years ago may be out of date!

Please have the readings completed before class. It is okay if you do not understand everything in the assigned reading.

The assigned readings for each class are subject to change.

Required Textbooks (abbreviations will be used throughout the course schedule)

- [SLP] *Speech and Language Processing*. Dan Jurafsky and James H. Martin. 3rd edition draft. 2024. Book can be accessed at https://stanford.edu/~jurafsky/slp3/.
- [NLPT] *Natural Language Processing with Transformers*. Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. Book can be accessed through the University Library.
- [TaD] *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. Readings will be provided on Canvas.
- [ISLP] *An Introduction to Statistical Learning with Applications in Python*. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. Book can be accessed at https://www.statlearning.com/.
- [DL] *Deep Learning*. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Book can be accessed at https://www.deeplearningbook.org/.

There are also some recommended books that may be of interest if you want to dive deeper into a subject.

Optional Textbooks and Readings

- *Neural Networks and Deep Learning*. Michael Nielsen. 2019. Book can be accessed at http://neuralnetworksanddeeplearning.com/.
- *Natural Language Processing*. Jacob Eisenstein. 2018. Book can be accessed at https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf.
- *Natural Language Processing with PyTorch*. Delip Rao and Brian McMahan. 2019. Book can be accessed through the University Library.

Programming and Software

We will be using Google Colab in this course. It is a hosted Jupyter Notebook service, meaning that all computing, including GPUs, takes place on the cloud. It requires no setup and can be used on any computer or tablet. The only requirement is a free Google account. Please let me know if you have trouble signing up for a Google account.

Assignments

There will be five assignments due throughout the semester. The date they are assigned will be under "Class Events." The date they are due will be under "Deadlines."

Midterm Exam

There will be an in-class midterm on Tuesday, October 8. Depending on the class schedule, this date might be moved.

Project

The class will conclude with a project. There are two options for the project. The first is a **research project**. The research project may be an application project (for example, answering a substantive question using NLP tools) or a methodological project (for example, expanding a methodology and explaining why it is useful). The second is to **task-driven project**. Here, you are given a dataset (or datasets) and you must achieve a solution or reach a specific level of performance on a task.

You may work in groups of up to two. Research projects with two co-authors must be more in-depth. Task-driven projects with two co-authors must achieve a higher level of performance. A completed project will consist of the following:

- **Proposal**. A project proposal will be due on November 8. This is one-page, single-spaced document that indicates if you or your group are doing a research project or task-driven project. If it is a research project, it must indicate the research question of interest, what data will be used, and how the research question will be answered. If it is a task-driven project, it must outline a preliminary plan on how the task will be accomplished.
- **Code and data pushed to GitHub**. The repository should be public; please talk with me if this is not possible (e.g., data that cannot be shared). Ideally, results should reproduce exactly. If results cannot be reproduced exactly, the report (see next bullet point) must discuss why.
- **A report**. A template for the report will be provided. The template can be used either in LaTeX or Microsoft Word. The report must be at least 3 pages if it is solo-authored and the report must be at least 5 pages if there are two co-authors. This page count does not include references. For either solo-authored or co-authored work, **papers have a maximum length of 8 pages**.
- **A presentation**. Solo-authored projects will require a 5-minute presentation. Co-authored projects will require a 10-minute presentation. A 5-minute Q&A will follow. You can choose to either pre-record your presentation or present in person. However, you must be present in person in order to answer questions from the class.

Both the code and report are due on **Friday, December 13 at 5:00pm ET**. A link to the Github and a copy of the report must be submitted on Canvas.

Breaking down the 30% that the project is worth of the total class grade: 5% project proposal, 5% presentation and Q&A, and 20% code and report.

Attendance and Participation Policy

It is expected that students will come to class, complete readings, and pay attention and participate in class. While you may use laptops, please do not use phones. **Attendance and participation will be assessed through randomly assigned quizzes and labs.** These quizzes and labs will only be graded on good faith completion.

Grading Policy

- **Attendance and Participation**: 10%
- **Assignments**: 40%
- **Midterm**: 20%
- **Project**: 30%

We will use the following base scheme to assign letter grades.

- A: $\geq 92\%$
- A-: $\geq 90\%$
- B+: $\geq 87\%$
- B: $\geq 82\%$
- B-: $\geq 80\%$
- C+: $\geq 77\%$
- C: $\geq 72\%$
- C-: $\geq 70\%$
- D+: $\geq 67\%$
- D: $\geq 62\%$
- D-: $\geq 60\%$
- F: $< 60\%$

Depending on the distribution of grades at the end of the semester, a curve may be applied. However, the final curve will be no stricter than the base scheme. In other words, if you receive a raw score of 85% in the course, you will be guaranteed at least a letter grade of B. You *may* receive a higher grade depending on the overall distribution of raw scores.

Late Policy

Assignments must be submitted on Canvas. Instructions will be included on each assignment. All assignments are due at **11:59pm ET on their assigned due date**.

Each student has a total of **five free late days** to use across all assignments throughout the semester. Each late day allows you to turn in an assignment up to 24 hours past the initial

deadline with no penalty. You may use as many late dates on each assignment as you wish. You do not need to inform me when you are using a late day. Simply turn in the assignment when you are done. I will keep track of how many late days you have used and will note it on the returned assignment.

Once you have exhausted all late days, late work will receive a penalty of 25% per day.

You may not use late days on the project proposal, final project presentation, or the final project's code and report.

For example, suppose that the assignment deadline is Friday at 11:59pm ET. If you turn in the assignment on Saturday at 2:29am ET (three hours after the deadline), you will use one late day. If you turn in the assignment on Sunday at 3:00am ET, you will use two late days.

Office Hours

I will hold office hours on Thursdays from 3-5pm in my office at DMTI 108A. No appointment is necessary. If this time does not work for you, please schedule an appointment using the scheduler here: https://calendly.com/pwu-american/office-hours. Although the scheduler is set up for one-on-one meetings, please feel free to meet me as a group as well. Only one person in the group needs to make the booking.

Academic Integrity Code

Standards of academic conduct are set forth in the university's Academic Integrity Code. By registering for this course, students have acknowledged their awareness of the Academic Integrity Code and they are obliged to become familiar with their rights and responsibilities as defined by the Code. Violations of the Academic Integrity Code will not be treated lightly and disciplinary action will be taken should violations occur. This includes cheating, fabrication, and plagiarism.

Artificial Intelligence (AI) Use Policy

The use of generative AI tools in this course is limited to specific assignments. On assignments, students will be given explicit permission and guidance for using particular tools, such as ChatGPT; all use of such tools should be appropriately acknowledged with citation. Students are responsible for recognizing the limitations of these tools, and are accountable for AI-generated work that produces invented data or sources. Such concerns may constitute violations of the University's Academic Integrity Code.

You are encouraged to use AI tools such as ChatGPT or Claude to review class material. However, please be aware that LLMs are prone to hallucinations, *especially* when it comes to technical materials!

Collaboration Policy

You are encouraged to work in groups and discuss course materials with your classmates. **Work you submit must be your own**. Violations of this policy include writing a solution

based on a classmate's code, looking at solutions online or using generative AI, or submitting copied answers to questions. **You must note on your assignments who you worked with.**

Students with Disabilities

If you wish to receive accommodations for a disability, please notify me with a memo from the Academic Support and Access Center (ASAC). As accommodations are not retroactive, timely notification at the beginning of the semester, if possible, is strongly encouraged. To register with a disability or for questions about disability accommodations contact the ASAC at 202.885.3360 or asac@american.edu.

Academic Support

All students may take advantage of the Academic Support and Access Center (ASAC) for individual academic coaching, the Writing Center, workshops, tutoring, peer tutor referrals, and Supplemental Instruction. The ASAC is located in Butler Pavilion 300. Additional academic support resources available at AU include the Bender Library, the Math Lab (located in Don Meyers Technology and Innovation Building), and the Center for Language Exploration, Acquisition, & Research (CLEAR) in Anderson Hall. A more complete list of campus-wide resources is available in the ASAC.

Acknowledgements

Many aspects of this course were inspired by the following sources: Zois Boukouvalas's DATA-441/641, Roberto Corizzo's CSC-208, Yue Dong's CS 173 at UCR, Chris Manning's CS224N at Stanford, Ambuj Tewari's STATS 607A at the University of Michigan, and Justin Johnson's EECS 498/598 at the University of Michigan.

Schedule

This schedule is tentative and may be adjusted as the class progresses. This syllabus will be updated as needed (see the top of the document to see the date it was last updated).

Part I: Introduction to NLP and Text as Data

Week 1

Tuesday, August 27

Topics	Introduction, syllabus discussion
Readings	None
Class Events	Course survey goes out (due date: Friday, August 30)
Deadlines	None

Friday, August 30

Topics	Python review
Readings	• Overview of Colaboratory Features • "An Informal Introduction to Python" • "More Control Flow Tools"
Class Events	Assignment 1 goes out (due date: Tuesday, September 17)
Deadlines	Course survey

Week 2

Tuesday, September 3

Topics	Python review
Readings	• "Data Structures" • "Brief Tour of the Standard Library" • "NumPy Fundamentals"
Class Events	None
Deadlines	None

Friday, September 6

Topics	Linear algebra review
Readings	• "10 minutes to pandas" • "Introduction to Data Structures"
Class Events	This class will be prerecorded; **we will not be meeting in person**
Deadlines	None

Week 3

Tuesday, September 10

Topics	Text preprocessing: bag of words, n-grams, evaluation metrics
Readings	• TaD Ch. 2 (on Canvas) • TaD Ch. 5 (on Canvas) • SLP Ch. 2
Class Events	None
Deadlines	None

Friday, September 13

Topics	Text classification
Readings	ISLP Ch. 1, 4
Class Events	None
Deadlines	None

Part II: Language Models, Neural Networks, and Embeddings

Week 4

Tuesday, September 17

Topics	Language models: naive Bayes, logistic regression, neural networks
Readings	SLP Ch. 3
Class Events	Assignment 2 goes out (due date: Tuesday, October 1)
Deadlines	Assignment 1 due

Friday, September 20

Topics	Neural networks
Readings	• ISLP Ch. 10 • SLP Ch. 7 • DL Ch. 6
Class Events	None
Deadlines	None

Week 5

Tuesday, September 24

Topics	Word embeddings, part 1: word2vec
Readings	• "The Illustrated Word2vec." Jay Alammar. 2019. • "Word2Vec Tutorial – The Skip-Gram Model." Chris McCormick. 2017. • "Word2Vec Tutorial Part 2 – Negative Sampling." Chris McCormick. 2017.
Class Events	None
Deadlines	None

Friday, September 27

Topics	Word embeddings, part 2: GloVe
Readings	SLP Ch. 6
Class Events	None
Deadlines	None

Week 6

Tuesday, October 1

Topics	Recurrent neural networks
Readings	SLP Ch. 8
Class Events	None
Deadlines	Assignment 2 due

Friday, October 4

Topics	Applications of word embeddings
Readings	Note: you can skim these readings; the idea is to get the main gist of each paper • "Unsupervised word embeddings capture latent knowledge from materials science literature." Vahe Tshitoyan et al. 2019. • "Partisan Associations of Twitter Users Based on Their Self-descriptions and Word Embeddings." Patrick Y. Wu et al. 2019. • "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." Pedro L. Rodriguez and Arthur Spirling. 2022.
Class Events	None
Deadlines	None

Week 7

Tuesday, October 8

Topics	None - **Midterm**
Readings	None
Class Events	**In-class midterm**
Deadlines	None

Friday, October 11

Topics	None - Fall Break
Readings	None
Class Events	**Fall Break - No class meeting**
Deadlines	None

Part III: Transformers and Large Language Models (LLMs)

Week 8

Tuesday, October 15

Topics	Intro to LLMs: Transformers and Attention
Readings	NLPT, Ch. 1
Class Events	Project information goes out (project proposal due on Friday, November 8), Assignment 3 goes out (due date: Tuesday, October 29)
Deadlines	None

Friday, October 18

Topics	Intro to LLMs: The Architecture of Transformers
Readings	SLP, Ch. 10
Class Events	None
Deadlines	None

Week 9

Tuesday, October 22

Topics	Using Hugging Face
Readings	• Hugging Face NLP Course, Ch. 2 • "Fine-tune a pretrained model"
Class Events	None
Deadlines	None

Friday, October 25

Topics	Fine-tuning
Readings	• NLPT, Ch. 2 • SLP, Ch. 10
Class Events	None
Deadlines	None

Week 10

Tuesday, October 29

Topics	Pre-Training and Post-Training
Readings	• NLPT, Ch. 10 • "Aligning Language Models to Follow Instructions." 2022. • "Illustrating Reinforcement Learning from Human Feedback (RLHF)." Nathan Lambert et al. 2022. • "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." Rafael Rafailov et al. 2024. • "Preference Tuning LLMs with Direct Preference Optimization Methods." Kashif Rasul et al. 2024.
Class Events	Assignment 4 goes out (due date: Tuesday, November 12)
Deadlines	Assignment 3 due

Friday, November 1

Topics	Prompt engineering, Part 1
Readings	• SLP, Ch. 12 • "Prompt engineering" • Prompt Engineering Guide, Introduction • Prompt Engineering Guide, Techniques
Class Events	None
Deadlines	None

Week 11

Tuesday, November 5

Topics	None - Election Day
Readings	None
Class Events	**Election Day - No class meeting**
Deadlines	None

Friday, November 8

Topics	Prompt engineering, Part 2
Readings	• SLP, Ch. 14 • "What is RAG (Retrieval-Augmented Generation)?"
Class Events	None
Deadlines	Project proposal due

Week 12

Tuesday, November 12

Topics	Applications of LLMs, Part 1
Readings	Pending
Class Events	Assignment 5 goes out (due: Friday, November 22)
Deadlines	Assignment 4 due

Friday, November 15

Topics	Applications of LLMs, Part 2
Readings	Pending
Class Events	None
Deadlines	None

Week 13

Tuesday, November 19

Topics	Applications of LLMs, Part 3
Readings	Pending
Class Events	None
Deadlines	None

Friday, November 22

Topics	Applications of LLMs, Part 4
Readings	Pending
Class Events	Assignment 5 due
Deadlines	None

Week 14

Tuesday, November 26

Topics	Wrapping up the course - meeting virtually
Readings	None
Class Events	**Because of Thanksgiving Break, we will be meeting virtually**
Deadlines	None

Friday, November 29

Topics	None - Thanksgiving Break
Readings	None
Class Events	**Thanksgiving Break: No class meeting**
Deadlines	None

Week 15

Tuesday, December 3

Topics	Project presentations
Readings	None
Class Events	Final project due on Friday, December 13 **at 5:00pm**
Deadlines	None

Friday, December 6

Topics	Project presentations
Readings	None
Class Events	Final project due on Friday, December 13 **at 5:00pm**
Deadlines	None