DATA 470D3 — Natural Language Processing

Professor: Stephen Davies
Fall semester 2025

Class: TR 3:30pm in Farmer 054

Final exam: Thursday, Dec 11th, 3:30–6pm

Office Hours (James Farmer Hall 044):	
Mon	1–4pm
Tue	12–2pm

`http://stephendavies.org/nlp`

Much of the data we work with in Data Science can be considered **structured** data. It's rigidly unambiguous, and conforms wholly to some "schema" that defines the various parts it contains and what they mean. Most structured data is pretty easy to work with, and that's because its authors made it so. After all, the main point of collecting information in (say) a tabular format is so that it can be consulted and queried easily.

In stark contrast to this is **unstructured data**, which is any electronic material that does *not* conform to a schema. Images and sound files are in this category, as are videos, but the most common is **text** expressed in a **natural language** like Spanish, Chinese, or English. Some experts estimate that over 80% of the data used by any organization consists of unstructured text. What a treasure trove, if we can make use of it!

The field of NLP, and this class, will be focused on extracting meaning from text despite its lack of structure. We'll learn the linguistic concepts and the statistical tools necessary to approach this maddeningly-inconsistent domain, and write programs that can derive insight from something as formal as a journal article or as informal as a tweet.

As even your grandparents know, NLP applications have positively exploded in reach and popularity over the last several years, including the debut of ChatGPT to the public in November 2022. I consider this quite possibly the most important technology of our lifetime – perhaps ever – and breaking it down to first principles will be a highlight of this semester. Without being too egocentric, let me suggest there's a decent chance that this will be the most important college course you ever take.

Course Objectives

- To survey the field of Natural Language Processing (NLP), so that you know what tasks it can currently accomplish and what is still on the horizon.
- To experiment with pre-neural NLP algorithms, in order to appreciate their capabilities and limitations.
- To examine in detail a modern neural architecture and illuminate how it achieves its magic.
- To give you the opportunity to write real programs running on real text data sets, and to see what kinds of practical issues are involved.
- To confront some of the major ethical concerns that NLP involves, including privacy and consent; bias and fairness; transparency and accountability; and intellectual property and ownership.

Student Learning Outcomes

After completing this course, students will be able to...

- ...identify the different tasks associated with the field of Natural Language Processing (NLP), and to explain their purposes, assumptions, promises, and limits.
- ...create custom NLP models in PyTorch for small corpora, and evaluate their effectiveness.
- ...obtain (*e.g.*, from HuggingFace) and fine-tune existing NLP models for small corpora, and evaluate their effectiveness.
- ...make use of some key NLP-related Python libraries and apply them to larger corpora.
- ...both:
  - comprehend the statistical theory underpinning important NLP algorithms, and
  - apply that theory and implement those algorithms concretely in a programming language (Python).

- ...confront and make informed decisions regarding some of the major ethical concerns that NLP involves, including privacy and consent; bias and fairness; transparency and accountability; and intellectual property and ownership.

Rules of the game

1. There are absolutely, positively, NO stupid questions!! Your job is not to already know everything before you start the course. Your job is to try hard to learn, and part of that involves asking questions. I'm a nice guy, and I will not ever belittle you, snub you, or make fun of you; and if anyone else does so I will personally break both of their arms.
2. This class will be interactive. When I point at you in class, say your first name, and be prepared to try and answer questions. (Don't worry if you don't know all the answers.)
3. The books we'll be using (my *Quick Steep Climb* text, and Jurafsky & Martin's *Speech and Language Processing*) are mandatory, and you will be required to actually read the sections I assign. Luckily for you, J&M is (1) an extremely well-written book, and (2) all freely available online thanks to Dan and Jim. *QSC* is also free and open source.
4. The reading checks, quizzes and the final exam will cover both (1) lecture material from class, and (2) the assigned readings. Much of what's in #2 will make an appearance in #1, but this is not guaranteed.
5. Don't skip class. Just don't. It's bad form. I work hard to prepare for class, to make it compelling and relevant. It hurts my feelings when you don't come. Plus you miss out on important stuff, and you'll end up falling behind if you skip lecture. So come every time. Come happy, fresh, excited, ready to think and to participate.
6. **Absolutely no laptops, cell phones, or other devices during class.** I've had students claim that they take notes on their laptop during lecture, but even if it's true, those things are way too big a distraction to you and your fellow students to make it worth it. Just stay tuned in, because I move fast.
7. For any Zoom class that may take place this semester, **you must have your webcam on during the entirety of the lecture.** If you don't have a working Webcam, buy one immediately.

Books

- **QSC:** *A Quick, Steep Climb Up Linear Algebra*, version 1.1.0: Davies, Stephen. 2021. Available online at `http://stephendavies.org/quick.pdf`.
- **J&M**: *Speech and Language Processing*, 3rd Edition Draft, Jan 12 2025 release: Jurafsky, Daniel and Martin, James. 2025. Available online at `https://web.stanford.edu/~jurafsky/slp3/`.

I'll be surgically assigning specific sections of QSC to build the background you need to work with vectors, matrices, and tensors in PyTorch. If you're a Computer Science major and have already taken CPSC 284, this will be review for you. But I'm warning you: this stuff needs to be *sharp* in your mind. Do actually plan on spending time reviewing.

J&M is still unfinished (Jim told me via email that the work on the 3rd Edition is proceeding "at a glacial pace") and tbh probably will always be. But it's hands-down one of the best technical books I've ever read. The authors are giants in the field – legends, even – and they know NLP as well as anyone on the planet ever has. Enjoy their freely available book!

The Honor Code and this course

For this course:

- The linear algebra reading checks will take place in class, at the very start of the class period the first three weeks. They are closed-book, closed-notes, and timed (at about 10 minutes).
- The Canvas quizzes are open-book and open-notes, but they must be taken alone, in a quiet place, without any form of contact with anyone, and *without any use of any website* other than the book's website and the class website. *No* ChatGPT or other form of AI is allowed. These quizzes will be timed generously at anywhere from 30-60 minutes.
- **You must not use any part of any other student's Python code.** You may not directly copy any part of anyone else's program. That being said, it is okay to work alongside classmates in a study group, and you are allowed to

talk through homework assignments, ask others how they solved a particular problem, and so forth. **If you do this, you must name all the people you worked with on your homework submission.** If you fail to give this information clearly on your homework submission, it is an Honor Code violation.

- You *are* permitted to use AI (such as ChatGPT) while working on the homeworks, but **in your submission you *must* include link(s) to all the chat(s) you created while working on it.** If you fail to give this information clearly in your homework submission, it is an Honor Code violation.

Late policy

No late work will be accepted this semester. Get your stuff in on time!

Grading

Grading this semester will be based on "experience points" (XP). As you complete activities, you will earn XP towards your final total. XP can never be lost, only gained, but you have to earn what you get (*i.e.*, you don't "start off with a 100%" and lose points based on mistakes you make).

There will be opportunities to earn XP throughout the course. Some of these will be spontaneous as the mood strikes me. Some you can earn by completing in-class activities. Some may be in response to impressive things I see you do as the semester progresses. The following opportunities are *guaranteed* to be available to you:

Guaranteed XP opportunities:

Activity	Possible XP
Five linear algebra reading checks	5 each
Six open-book, open-note, timed Canvas quizzes	25 each
Five homework assignments	40 each
2-page paper: ethical considerations of LLMs	50
Final exam (comprehensive)	150
Various and sundry others	varies

Grading levels

Here are the levels you may achieve, together with the grade awarded (if any) and the points necessary to reach!

Level	**Total XP**	Semester grade
William Shakespeare	**550**	A+
Homer	**500**	A
Fyodor Dostoevsky	**450**	A–
Charlotte Brönte	**400**	B+
Leo Tolstoy	**375**	B
Sophocles	**350**	B–
Jane Austen	**330**	C+
Toni Morrison	**310**	C
Emily Dickinson	**290**	C–
Geoffrey Chaucer	**270**	D+
Charles Dickens	**250**	D
George Eliot	**230**	D–
Isaac Asimov	**210**	
C.S. Lewis	**190**	
George Orwell	**170**	
Virginia Woolf	**150**	
Ursula Le Guin	**130**	
J.R.R. Tolkien	**110**	
Mary Shelley	**90**	
Mark Twain	**75**	
Lewis Carroll	**60**	
Agatha Christie	**45**	
Sir Arthur Conan Doyle	**30**	
E.L. James	**20**	
Edward Bulwer-Lytton	**10**	
Zerna Addis Sharp	**0**	

Submitting homeworks

Rules for submitting homeworks (whether in Python or in English) will be given when the homework is assigned. For programs, you'll be emailing me your program code as an attachment, and using a specific subject line to distinguish it from my hordes of other email. Meeting the deadline is a matter of sending your email before time expires. For responsive readings, **you must submit a hardcopy of your paper – nothing electronic will serve as an adequate substitute.** I'll probably collect these in class.

Also, most of my homeworks are due at “midnight.” Here’s what “midnight” means: if a homework is due “at midnight on Thursday,” then it is due after all of Thursday has elapsed, and the clock strikes twelve. (In other words, this is good news: you have all Thursday to work on it.)

Basis for determining mid-semester reports

For midterm progress reports, I look mostly at (a) whether you’ve been turning assignments in (and on time), and (b) quiz scores. If either or both of these categories are lacking, it’s a sign of danger, and I will give you a “U” for your mid-semester grade. Please don’t hesitate at all to come talk to me about this so we can figure out how you can do better in the course.

Use of Artificial Intelligence (AI) technologies

AI is **not** permitted on any quiz or exam. Doing so will be considered a violation of course policy and as such, the student may be referred to the UMW Honor Council for a violation of academic integrity.

You *are* permitted to use AI (such as ChatGPT) while working on the homeworks, but **in your submission you *must* include link(s) to all the chat(s) you created while working on it.** Failure to do so will be considered a violation of course policy and as such, the student may be referred to the UMW Honor Council for a violation of academic integrity.

Similarly, you *are* permitted to use AI (such as ChatGPT) while composing your responsive readings, but **in your submission you *must* include link(s) to all the chat(s) you created while working on it.** Failure to do so will be considered a violation of course policy and as such, the student may be referred to the UMW Honor Council for a violation of academic integrity.

Guidelines for class participation

I believe that students learn best when they participate wholeheartedly in all aspects of the learning process. Hence while your grade will not be partially determined by any “class participation score” *per se*, it is very much to your advantage, and very much recommended, that you come to lecture every single class period, and participate fully in it.

Title IX Statement

UMW faculty are committed to supporting students and upholding the University’s *Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence.* Under Title IX and this Policy, discrimination based upon sex or gender is prohibited. If you experience an incident of sex or gender based discrimination, we encourage you to report it. **While you may talk to me, understand that as a “Responsible Employee” of the University, I must report to UMW’s Title IX Coordinator what you share.** If you wish to speak to someone confidentially, please contact the confidential resources below. They can connect you with support services and help you explore your options. You may also seek assistance from UMW’s Title IX Coordinator; their contact information can be found below. Please visit `http://diversity.umw.edu/title-ix/` to view *UMW’s Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence* and to find further information on support and resources.

Resources
Ruth Davison, Ph.D.
Title IX Coordinator
Lee Hall, Room 401
540-654-5656
`rdavison@umw.edu`

Confidential Resources
On-Campus
Talley Center for Counseling Services
Lee Hall 106, 540-654-1053

Student Health Center
Lee Hall 112, 540-654-1040

Off-Campus
Empowerhouse (24-hr hotline)
540-373-9373

RCASA (24-hr hotline)
540-371-1666

Recording Policy

Classroom activities in this course may be recorded by students enrolled in the course for the personal, educational use of that student or for all students presently enrolled in the class only, and may not be further copied, distributed, published or otherwise used for any other purpose without the express written consent of the course instructor. All students are advised that classroom activities may be taped by students for this purpose. Distribution or sale of class recordings is prohibited without the written permission of the instructor and other students who are recorded. Distribution without permission is a violation of copyright law. This policy is consistent with UMW's *Policy on Recording Class and Distribution of Course Materials.*

Accessibility statement

The Office of Disability Resources has been designated by the university as the primary office to guide, counsel, and assist students with disabilities. If you receive services through the Office of Disability Resources and require accommodations for this class, please provide me a copy of your accommodation letter via email or during a meeting. I encourage you to follow-up with me about your accommodations and needs within this class. I will hold any information you share with me in the strictest confidence unless you give me permission to do otherwise.

If you have not made contact with the Office of Disability Resources and have reasonable accommodation needs, their office is located in Seacobeck 005, phone number is (540) 654-1266 and email is `odr@umw.edu`. The office will require appropriate documentation of disability.

Basic needs security

Learning effectively and engaging wholly in class is dependent upon our basic security and having our fundamental needs met: having a safe place to sleep at night, regular access to nutritious food, and some assurance of safety. If you have difficulty affording groceries or accessing sufficient food to eat every day, or if you lack a safe and stable place to live, please contact Chris Porter, Assistant Dean of Students, at `cjporter@umw.edu`. Additionally, the Gwen Hale Resource Center is a free resource on campus, providing food, toiletries and clothing to any member of our community. It is open Monday, Tuesday and Friday from 1pm-6pm, on the 5th floor (floor A for Attic) of Lee Hall, or `resource@umw.edu`. Finally, you are always welcome to talk with me about needs, if you are comfortable doing so. This will enable me to provide any resources I may possess.

How to reach me

Come to office hours, see me after class, or e-mail me (`stephen@umw.edu`).

How to reach you

I will be communicating with you outside of class time via e-mail, so make sure to check your UMW e-mail every day! I will also post announcements to the course website, so be sure to subscribe to its RSS feed and check it in your feed reader at least once a day!

Calendar

The official calendar for the course, complete with assignment due dates, tests, *etc.*, will be maintained on the course website at `http://stephendavies.org/nlp`. **In any way that the website conflicts with the tentative calendar below, the website is to be considered correct, and the tentative calendar below out of date.**

Week	Topics	Due
1	Regex's, corpora, words and tokens, tokenization	RC's #1&2
2	N-gram models, BoW models, text classification	RC's #3&4
3	PyTorch tensor calculation, logistic regression	RC #5, HW #0
4	Calc primer; stochastic gradient descent	HW #1
5	Vector semantics and embeddings	Quiz 1
6	Neural models, in theory and in PyTorch	HW #2
7	Training neural models: backprop and autodiff	Quiz 2
8	*(fall break)* The attention mechanism	HW #3
9	The transformer architecture	Quiz 3
10	Contextual and positional embeddings	Quiz 4
11	The HuggingFace code base, hub, and community	HW #4
12	Training, sampling, and evaluating LLMs	Quiz 5
13	Machine translation and summarization	Quiz 6
14	Question answering and RAG *(Thanksgiving)*	HW #5
15	Dialogue systems and future trends in NLP	Paper
Finals		Final exam