# W&M DATA 340 4 Natural Language Processing, Spring 2024 Syllabus

Schedule | Email Instructor

## Course overview

This course is designed to introduce students to Natural Language Processing (NLP) and its applications in academic research, data science, and industry. Students will learn how to use natural language processing techniques to gain a deeper understanding of a research question and/or topic.

## Course venue and time

- Integrated Science Center, room 1280. Tuesdays (T) and Thursdays (Th) 5:00 - 6:20
- See course in Blackboard

## Instructor contact and office hours

- email: jmtucker02@wm.edu
- website: https://jamesmtucker.com
- Office hours by appointment only

## Programming language

This course is language agnostic. You can submit your homework and project in whatever programming language you prefer. In class lectures, we will use Python, R, or Mojo for the most part.

## Tools

Students are encouraged to utilize generative AI tools when they are stuck on implementing code or ideating on topics for projects. If you copy code from a generative AI tool, like ChatGPT or GitHub CoPilot, please provide a comment in your programming language's comment syntax that you resourced a tool to solve your problem. If you resourced Stack Overflow or another source, please footnote your sources accordingly.

## Course objectives

- Understand the basics of natural language processing techniques and how they can be used to in data driven decision models
- Learn how to use natural language processing tools and libraries to perform tasks such as text classification, sentiment analysis, and text generation
- Develop the ability and experience to design and implement natural language processing systems for real-world applications
- Explore ethical and social implications of natural language processing and artificial intelligence

## Course topics

1. Introduction to natural language processing
2. Data set creation and documentation
3. Text preprocessing and cleaning
4. Text classification and sentiment analysis
5. Neural Networks, Transformers, Large Language Models
6. Ethical and social implications of natural language processing and artificial intelligence

## Textbook

### Required

- Jurafsky, Dan and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Online: https://web.stanford.edu/~jurafsky/slp3/. PDF

### Recommended

- Arcila Calderon, Carlos, et al. *Computational Analysis of Communication.* United Kingdom, Wiley, 2022. Google Books
- Tunstall, Lewis, Leandro von Werra, and Thomas Wolf. *Natural Language Processing with Transformers* O'Reilly Media, 2022. Google Books

## Assignments

### Descriptions

There are a possible 165 points to earn in this course.

All assignments are posted GitHub Repo and links are provided in Blackboard.

- Understanding NLP (10pts) - A conversation with ChatGPT about Linguistics, NLP, and Data Science
- Course reading (15pts) - Please read the required readings listed in the course syllabus. Lecture topics will parallel the reading and the readings will help you comprehend the subject matter in greater detail.
- NLP Problem sets (100pts) - these assignments are designed to reinforce the ideas discussed in lecture or the assigned reading. Whereas the student can expect to solve the problem sets from the topics covered in lecture, it is encouraged that the student explore additional solutions and creative thinking.
  - Problem set 0: Exploratory Data Analysis in a NLP context - Distributions and Features
  - Problem set 1: Author prediction - Federalist Papers
  - Problem set 2: Semantic Search Engine - king is to queen as man is to ?
  - Problem set 3: Clustering news articles - Retrieval augmented generation (RAG)
  - Problem set 4: Fine-tuning a LLM on prompt engineering
- Final Project (40pts) - Demonstrate your mastery of NLP concepts with a topic of your choosing

### Late time currency

You get 4 days of late time. This means that if you cannot turn in an assignment on time, you can use your late time currency to extend a deadline by however many days are needed, not extending beyond 4. You get four days credit. Use the time wisely. If you need more than four days time, you can earn an additional day by completing a coding challenge:

1. Sparse matrix multiplication
2. Probability of disease
3. Distances
4. Regression tree
5. K-means

## Schedule

| date | day | topic | assignment | reading | academic calendar |
|---|---|---|---|---|---|
| 2024-01-25 | Th. | Intro & syllabus | | | |

| date | day | topic | assignment | reading | academic calendar |
|---|---|---|---|---|---|
| 2024-01-30 | Tu. | NLP, Data science, and GenAI | [ChatGPT: P] | Lockhart | |
| 2024-02-01 | Th. | Vectorized computation, Data structures | | van Atteveldt et al. 5 6, 7 | |
| 2024-02-06 | Tu. | Statistics and language | [ChatGPT: D] | Manning and Schütze | |
| 2024-02-08 | Th. | Documents as bags of words | | Jurafsky & Martin 3, Grimmer 3 | |
| 2024-02-13 | Tu. | N-gram language models | [Problem Set 0: P] | Jurafsky & Martin 4, 5 | |
| 2024-02-15 | Th. | N-gram langauge models | | Jurafsky & Martin 4, 5 | |
| 2024-02-20 | Tu. | Distributional semantics (Word2Vec, Doc2Vec) | [Problem Set 0: D] | Jurafsky & Martin 6 | |
| 2024-02-22 | Th. | Training a Word2Vec/Doc2Vec Model | | Jurafsky & Martin 6, Gensim | |
| 2024-02-27 | Tu. | Documents as sequences | [Problem Set 1: P] | Jurafsky & Martin 7 | |
| 2024-02-29 | Th. | Data modeling, cleaning, optimizing | | Jurafsky & Martin 2 | |
| 2024-03-05 | Tu. | Neural networks architecture | [Problem Set 1: D] | Jurafsky & Martin 7 | |
| 2024-03-07 | Th. | LLMs, Language, & Semantics | | | |
| 2024-03-12 | Tu. | No class | | | Spring Break |
| 2024-03-14 | Th. | No class | | | Spring Break |
| 2024-03-19 | Tu. | Recurrent neural networks (RNNs) | [Problem Set 2: P] | Jurafsky & Martin 9 | |
| 2024-03-21 | Th. | Distrib. semantics contextual embeddings | | Jurafsky & Martin 10 | |
| 2024-03-26 | Tu. | Retrieval augmented generation | [Problem Set 2: D] | Jurafsky & Martin 14 | |
| 2024-03-28 | Th. | Clustering and visualizing embeddings | | | |
| 2024-04-02 | Tu. | Transformer neural network | [Problem Set 3: P] | Jurafsky & Martin 10 | |
| 2024-04-04 | Th. | Attention mechanism | | Jurafsky & Martin 10 | |
| 2024-04-09 | Tu. | Fine-tuning pretrained models | [Problem Set 3: D] | Jurafsky & Martin 11, Howard | |
| 2024-04-11 | Th. | Fine-tuning pretrained models | | Jurafsky & Martin 11, Howard | |
| 2024-04-16 | Tu. | Quantifying sentiment as a feature | [Problem Set 4: P] | | |
| 2024-04-18 | Th. | Quantifying Named Entities | | | |
| 2024-04-23 | Tu. | LLMs and knowledge graphs | [Problem Set 4: D] | | |
| 2024-04-25 | Th. | LLMs and causal reasoning | | | |

| date | day | topic | assignment | reading | academic calendar |
|---|---|---|---|---|---|
| 2024-04-30 | Tu. | Debugging NLP models, ethics, data | | | |
| 2024-05-02 | Th. | Open discussion & project highlights | | | |
| 2024-05-07 | Tu. | | | | Final exam periods |
| 2024-05-09 | Th. | | | | Final exam periods |
| 2024-05-14 | Tu. | | | | Final exam periods |

## Course policies

Please read and take notice of the following:

### Grade scale

| | Mark | | Mark |
|---|---|---|---|
| 93 - 100 | A | 73 - 76 | C |
| 90 - 92 | A- | 70 - 72 | C- |
| 87 - 89 | B+ | 67 - 69 | D+ |
| 83 - 86 | B | 63 - 66 | D |
| 80 - 82 | B- | 60 - 62 | D- |
| 77 - 79 | C+ | 00 - 59 | F |

### Grading appeals

To appeal a grade, schedule a meeting to discuss it with me.

### Communications

Course announcements will be posted in Blackboard. Please check Blackboard regularly for announcements.

The course readings, data sets, and code are available on the course GitHub repo.

### Absences

If you are absent please email me and let me know or send me a text message. Course work is due as detailed in the course schedule. Late work is penalized 2% of the earned mark for every day it is late. If you are absent on a day that an assignment or project milestone is due, please make sure to turn it in early. If you are ill, please communicate with me regarding an extension.

### Mental Well-Being

William & Mary recognizes that students juggle different responsibilities and can face challenges that make learning difficult. There are many resources available at W&M to help students navigate emotional/psychological, physical/medical, material/accessibility concerns, including:

- The W&M Counseling Center at (757) 221-3620. Services are free and confidential.
- The W&M Health Center at (757) 221-4386.
- For additional support or resources & questions, Contact the Dean of Students at 757-221-2510.

**Important dates**

| date | academic event |
|---|---|
| 2024-01-23 | Add/drop period begins at 1:00 pm |
| 2024-01-24 | First day of classes Non-degree seeking registration begins |
| 2024-02-02 | Last day to add/drop Deadline to file a minor declaration form for May or August graduates |
| 2024-02-03 | Withdrawal period begins |
| 2024-03-04 | Midterm grading begins |
| 2024-03-09 | Spring Break |
| 2024-03-10 | Spring Break |
| 2024-03-11 | Spring Break |
| 2024-03-12 | Spring Break |
| 2024-03-13 | Spring Break |
| 2024-03-14 | Spring Break |
| 2024-03-15 | Spring Break |
| 2024-03-16 | Spring Break |
| 2024-03-17 | Spring Break |
| 2024-03-18 | Classes resume after Spring Break |
| 2024-03-24 | Midterm grading ends at 11:59 p.m. |
| 2024-03-25 | Last day to withdraw from a full-term course |
| 2024-05-03 | Last day of classes |
| 2024-05-04 | Reading periods |
| 2024-05-05 | Reading periods |
| 2024-05-11 | Reading periods |
| 2024-05-12 | Reading periods |
| 2024-05-06 | Final exam periods |
| 2024-05-07 | Final exam periods |
| 2024-05-08 | Final exam periods |

| date | academic event |
|---|---|
| 2024-05-09 | Final exam periods |
| 2024-05-10 | Final exam periods |
| 2024-05-13 | Final exam periods |
| 2024-05-14 | Final exam periods |
| 2024-05-16 | Spring grades due by 9 a.m. for graduating students |
| 2024-05-16 | Spring Degree Conferral and Commencement Ceremony |
| 2024-05-17 | Spring Degree Conferral and Commencement Ceremony |
| 2024-05-18 | Spring Degree Conferral and Commencement Ceremony |
| 2024-05-21 | Spring grades due by 9 a.m. for continuing students |

**W&M honor code**

Students are expected to conduct themselves according to the Honor Code.