Spring 2024

DS 680-002: Natural Language Processing

Mengnan Du

Natural Language Processing - DS 680 Syllabus
Spring 2024

Instructor Information

Instructor	Email	Office	TA
Mengnan Du	mengnan.du@njit.edu	GITC 4410	Haiyan Zhao (hz54@njit.edu)

Code: DS680
Time: Tuesday/Friday, 8:30AM–9:50AM, 2024 Spring
Location: KUPF 207
Mode: Face-to-Face
Office Hours: Tuesday/Friday, 10:00AM–10:40AM, 10:45AM–11:25AM

General Information

Course Description

This course aims to teach how to process one of the fundamental data sources—natural language—with the help of deep learning techniques. The target of this course is to familiarize students with state-of-the-art language models, wide variety of tasks performed with these models and the fusion of these in deep learning architectures. This course will help students read advanced research papers on complex NLP concepts and theories, while the class project will help them apply NLP techniques to different domains.

Prerequisites/Co-requisites

- Programming proficiency in Python
- Basic knowledge of machine learning and neural networks
- Knowledge of probability, linear algebra, and calculus

Course Learning Outcomes

By the end of the course, students will be able to:

- Identify the main neural network architectures used in NLP, including RNNs, LSTMs, Transformers, and Large Language Models (LLMs).
- Explain how these models process and represent language.
- Recognize NLP problems suitable for different neural network approaches.
- Apply pretrained language models like BERT and RoBERTa to NLP tasks through finetuning.
- Build neural network models for text classification, machine translation, question answering, and other applications.
- Evaluate NLP models in terms of performance, fairness, explainability, and robustness.
- Adapt state-of-the-art NLP techniques to new datasets and applications.
- Stay current with the latest NLP research and models.

Coursework, Assessment and Related Outcomes

- **Assignments** (30%): There will be five assignments with both written and programming parts. Each homework is centered around an application and will also deepen your understanding of the theoretical concepts.
    - Each assignment worth 6% of the total grade.
    - Each assignment will have a 2-week deadline for completion.
    - The first assignment will be distributed in Week 2 on Tuesday, January 23.
    - Please refer to the course schedule below for specific due dates.
- **Midterm exam** (15%): The course has a midterm exam that will test your knowledge and problem-solving skills on all material up to and including lecture on March 8 (Week 8, Friday). We will arrange the midterm at our classroom KUPF 207.
- **Final exam** (25%): A final exam will test knowledge and problem-solving skills on all course material.
- **Final project** (30%): The project is expected to be finished in a group. The final project offers you a chance to apply your newly acquired skills towards an in-depth application. The project will be distributed on Friday, January 19. You are required to turn in a project proposal (due on March 1), give a project presentation and complete a paper written in the style of a conference (e.g., ACL) submission (due on April 30).

Course Materials

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed.).
- Jacob Eisenstein. Natural Language Processing
- Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing
- Reference Materials: Recent research papers from conferences such as ACL, EMNLP, NAACL, etc.

Course Schedule (tentative)

Week	Date	Topics	Readings	Assignments
Week1	Tue (1/16)	Introduction to NLP	1. Advances in natural language processing 2. Jacob Eisenstein Ch1.1	
	Fri (1/19)	Text Processing	1. J&M Ch. 2 2. Python's NLTK Package	Project out
Week2	Tue (1/23)	N-gram Language Models	J&M Ch. 3	A1 out
	Fri (1/26)	Text Classification	J&M Ch. 4	
Week3	Tue (1/30)	Naive Bayes	J&M Ch. 4	
	Fri (2/2)	Logistic Regression	J&M Ch. 5	
Week4	Tue (2/6)	Vector Semantics and Embeddings 1	J&M Ch. 6	A1 due A2 out
	Fri (2/9)	Vector Semantics and Embeddings 2	J&M Ch. 6	
Week5	Tue (2/13)	Neural Networks and Neural Language Models 1	J&M Ch. 7	
	Fri (2/16)	Neural Networks and Neural Language Models 2	J&M Ch. 7	
Week6	Tue (2/20)	Sequence Labeling for Parts of Speech and Named Entities	J&M Ch. 8	A2 due A3 out
	Fri (2/23)	RNNs and LSTMs	J&M Ch. 9	
Week7	Tue (2/27)	Machine Translation 1	J&M Ch. 13	

	Fri (3/1)	Machine Translation 2	J&M Ch. 13	Project proposal due
Week8	Tue (3/5)	Midterm Review, Q&A		
	Fri (3/8)	Midterm Exam		
Week9	Tue (3/12)	Spring break		No Classes Scheduled
	Fri (3/15)	Spring break		No Classes Scheduled
Week10	Tue (3/19)	Transformers and Pretrained Language Models 1	J&M Ch. 10	A3 due A4 out
	Fri (3/22)	Transformers and Pretrained Language Models 2	J&M Ch. 10	
Week11	Tue (3/26)	Fine-Tuning and Masked Language Models	J&M Ch. 11	
	Fri (3/29)	Good Friday		No Classes Scheduled
Week12	Thursday (4/2)	Large language models 1		A4 due A5 out
	Fri (4/5)	Large language models 2		
Week13	Tue (4/9)	Question Answering	J&M Ch. 14	
	Fri (4/12)	Invited Talk		
Week14	Tue (4/16)	Explainability in NLP		
	Fri (4/19)	Fairness in NLP	Papers about fairness in NLP	
Week15	Tue (4/23)	Final Review, Q&A		A5 due
	Fri (4/26)	Project presentation		
Week16	Thu (4/30)	Project presentation		Final project report due

Final Projects

The final project offers you the chance to apply your newly acquired skills towards an in-depth NLP application. Students are required to complete the final project in teams no more than 4 students (ideally 3 students as a team).

There are **two** options for the final project:

- Option 1: reproducing an ACL/NAACL/EMNLP 2020-2023 paper (encouraged);
- Option 2: complete a research project based on topics covered in this class (for this option, you need to discuss your proposal and get prior approval from the instructor).

All the final projects will be completed in teams of no more than 4 students (Find your teammates early!).

Deliverables: The final project is worth 30% of your course grade. Deliverables include:

- **Proposal** (0%): You need to turn in a one-page proposal on March 1. The proposal should outline what you propose to do and a rough plan for how you will pursue the project. We will then provide feedback and guidance on the direction to maximize the project's chance of succeeding. *The proposal is not graded.*
- **Project presentation** (10%): At the end of the semester, we will schedule project presentations for all the projects in the class.
- **Final paper** (20%): You need to complete a final report in the style of a conference submission (we recommend you to use the ACL 2023 template). It should begin with an abstract and introduction, clearly describe the proposed idea or exploration, present technical details, give results, provide analysis and discussion of the results, and cite any sources you used.

Policy and honor code:

- The final projects are required to implement in Python. You can use any deep learning framework such as PyTorch, Tensorflow and Keras.
- You are free to discuss ideas and implementation details with other teams. However, under no circumstances may you look at another team's code, or incorporate their code into your project.

Course Policies

Grade Corrections

Check the grades in course work and report errors promptly. Please try and resolve any issue within one week of the grade notification.

Incomplete

A grade of I (incomplete) is given in rare cases where work cannot be completed during the semester due to documented long-term illness or unexpected absence for other serious reasons. A student needs to be in good standing (i.e., passing the course before the absence) and receives a provisional I if there is no time to make up for the documented lost time; an email with a timeline of what is needed to be done will be sent to the student. Note that an I must always be resolved by the end of the next semester.

Fail of the Course

In the case when a student is unable to attend the class or exams, these must be communicated and documented promptly. In any other case, a student will fail this course and obtain an F if 1) missing three or above classes; 2) missing any exams; 3) not submitting course project final report. No exceptions will be granted.

Academic Integrity

Detailed guidance on academic integrity can be found at: Best Practices document.

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at: NJIT Academic Integrity Code.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university. If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu"

Acknowledgements

Portions of the lecture material are adapted from COS484: Natural Language Processing (undergraduate course) taught by Dnqi Chen, CS224N: Natural Language Processing with Deep Learning from Stanford University, and slides from the textbook Speech and Language Processing (3rd edition).