

COMP4221: Introduction to Natural Language Processing

Instructor: *Victor Junqiu WEI*
Email: victorwei@cse.ust.hk
Telephone: (852) 2358 8329
Office: 2538 (lift 25/26)

Course Description:

This course provides an introduction to the tasks, models and algorithms of natural language processing (NLP) including the fundamental NLP tasks and models (such as language model, text classification, word vectors, sequence labelling, etc.), and several typical NLP applications (such as question answering, dialogue systems and machine translation, etc.), and the linguistic structure annotation (such as constituency parsing and dependency parsing). The course also introduces some advanced topics such as pre-trained language models and their prompt tuning and instruct tuning methods to give the students more perspectives of the NLP industry.

Course Outcomes:

On successful completion of this course, the students should:

- Explain the principles behind fundamental NLP algorithms.
- Have a deep understanding of several typical NLP applications and linguistic structure annotating algorithms.
- Implement the fundamental NLP algorithms for solving real problems.
- Get a basic understanding of the techniques in the frontline of NLP industries.

Course Prerequisites:

- Computer science: object-oriented programming and data structures, design, and analysis of algorithms.
- Mathematics: multivariable calculus, linear algebra, probability, stochastic process, and statistics.
- Students are expected to have probability, linear algebra, algorithm design and machine learning background. It is suggested to take an introductory algorithm and an introductory machine learning course before taking this course.

Course Topics:

Topics
Text Processing (Regular Expressions, Tokenization, Alignments, Edit Distance)
N-Gram Language Model
Text Classification
Word Representation Learning
Sequence Labelling for Parts of Speech and Named Entities
Neural Networks and Neural Language Models
RNN and Transformer Models

Pretrained Language Models (Basic)
Sequence-to-Sequence Models and Machine Translation
Question Answering and Information Retrieval
Chatbots and Dialogue Systems
Constituency Parsing
Dependency Parsing
Pretrained Language Models (Advanced)
Prompting and Instruct Tuning

Assessments:

In general, the earned grade in the course will be based on the calculated total points according to the following schedule:

Activity or Task	Weight
Assignments	30%
Course Project	20%
Midterm Exam	20%
Final Exam	30%
Total	100%

Textbooks:

- Daniel Jurafsky & James H. Martin. *Speech and Language Processing*. Third Edition draft. Draft of January 12, 2022.
- Lee, Raymond ST. *Natural Language Processing: A Textbook with Python Implementation*. Springer Nature, 2023.
- Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.
- Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000.