DSAA 3051: Introduction to Natural Language Processing and Knowledge Engineering

Fall 2026, Course Credits: 3

Instructor: Jiaheng WEI Teaching Assistant: Mingjie XU, Yuhan PU

Time: MoWe 12:00PM - 13:20PM, 01/09/2026 - 07/12/2026
Location: Rm 102, E1

August 2026

1 Introduction

Course Description This course provides a comprehensive overview of modern approaches for (1) natural language processing and (2) knowledge graphs. Students will explore text processing, linguistic preprocessing, and foundational language models, progressing from n-grams to word embeddings and neural language models. The curriculum covers semantic and discourse analysis, information extraction, and the integration of knowledge bases. Key topics include knowledge-intensive tasks, and the construction and application of knowledge graphs, along with ontologies and semantic web technologies like RDF, OWL, and SPARQL. The course also delves into large language models, emphasizing transformer architectures and ethical considerations of NLP and knowledge systems such as bias, fairness, and privacy. Through lectures, discussions, and practical exercises, students will gain critical insights into the development, maintenance, and application of NLP and knowledge systems, preparing them to address real-world challenges in these fields.

Rationale for Introducing this course Natural Language Processing (NLP) and Knowledge Engineering (KE) are foundational pillars of modern artificial intelligence, driving innovations in language understanding and generation, knowledge representation, and intelligent systems. This course is designed to equip students with the theoretical and practical skills necessary to navigate the intersection of these fields, addressing the growing demand for expertise in areas such as semantic analysis, knowledge graphs, and large language models.

The course aims to provide a comprehensive overview of NLP and KE, from foundational techniques (e.g., tokenization, n-gram models) to cutting-edge advancements (e.g., knowledge graphs, large language models). By integrating linguistic theory, machine learning, and knowledge representation, students will learn to build systems that can process human language, extract structured knowledge, and reason over complex data. This skill set is critical for applications ranging from chatbots and semantic search engines to knowledge-driven systems.

By taking this course, students are expected to build a solid foundation of concepts of natural language processing and knowledge engineering and learn to apply different methods for solving real-world problems.

2 Course Information

Time & Location:

- Lecture Time: MoWe 12:00 PM – 13:20 PM, 01/09/2026 – 07/12/2026
- Lecture Location: Rm 102, E1
- Lab Time: We 13:30 PM – 14:20 PM, 01/09/2026 – 07/12/2026
- Lab Location: Rm 102, E1

Lecturer:

- Lecturer: Dr. Jiaheng WEI;
- Assistant Professor: DSA Thrust, Info Hub, HKUST-GZ;
- Research Interests: Trustworthy Machine Learning and Large Language Models;
- Email: jiahengwei@hkust-gz.edu.cn;
- Website: https://sites.google.com/ucsc.edu/jiahengwei;
- Office: E3-305;
- Office hours: We 14:30 PM -15:30 PM

You are welcome to come to office hours with specific questions about the material, or just to chat about things you find interesting and want to explore further.

TAs:

- Mingjie XU: mxu188@connect.hkust-gz.edu.cn;
- Yuhan PU: ypu159@connect.hkust-gz.edu.cn;

Each TA supervises ≈10-15 students, regarding question/assignment/exam concerns.

3 Planned Topics (Tentative)

1. Introduction to NLP and Knowledge Engineering (Week 1)
   Course structure and expectations; Historical context and evolution of NLP and Knowledge Engineering.
2. Text Processing and Linguistic Preprocessing (Week 2)
   Tokenization, stemming, and lemmatization; Introduction to semantics: word senses and meanings.
3. N-Gram Models and Basics of Language Modeling (Week 3)
   N-Gram language models and their applications; Limitations of statistical language models.
4. Word Embeddings and Semantic Representation (Week 4)
   Word2Vec, GloVe, and fastText; Semantic similarity and word sense disambiguation.
5. Neural Language Modeling (Week 5)
   From word embeddings to contextual embeddings; Neural sequence labeling.
6. Information Extraction and Knowledge Bases (Week 6)
   Relation extraction and entity linking; Introduction to knowledge bases and their integration with NLP.
7. Midterm Exam (Week 7)
8. Ontologies and Semantic Web Technologies (Week 8)
   Basics of ontologies and their role in knowledge engineering; Introduction to RDF, OWL, and SPARQL
9. Knowledge Graphs Construction, Maintainence and Applications (Week 9)
   Techniques for constructing and maintaining knowledge graphs; Knowledge graphs for downstream applications.
10. Applications of Knowledge Engineering in NLP (Week 10)
    Knowledge-intensive tasks: entity linking and semantic search; Integration of KG and KB with NLP systems.

11. Large Language Models (Week 11)

    Transformer architecture to the generative paradigm; Pre-training, supervised fine-tuning and RLHF.

12. Safety and Security: Bias, Fairness and Privacy (Week 12)

    Additional Considerations of NLP systems; Safeguard the NLP system.

13. Summary and Discussions (Week 13)

14. Final Exam (Week 14)

4 Learning Outcomes and Alignment

4.1 Key Course Intended Learning Outcomes.

Upon completion of this course, students are expected to be able to do the following:

Course ILOs	Nature of the learning outcomes
1. Explain the fundamental concepts of natural language processing.	A
2. Analyze fundamental tasks in NLP.	A
3. Use some hands-on skills for solving NLP problems.	A
4. Differentiate some classic and state-of-the-art techniques.	A
5. Communicate effectively to convey scientific knowledge.	A, B
6. Develop a broad interest in NLP and Knowledge Engineering.	A, C

- A: Knowledge/Content Related; B: Academic; C: Others.

4.2 Contribution of Learning Outcomes to Programs of Study.

Program of study 1: BSc in Data Science and Big Data Technology Program ILOs	To be achieved through these course ILOs
1. Explain data problems arising in different areas of science, technology, and the society. (Knowledge)	CILO-1, CILO-3, CILO-5
2. Model the problem with learned mathematics theories and models (Execution)	CILO-1, CILO-2, CILO-3, CILO-4
3. Apply different mathematical tools to model data problems in application areas. (Execution)	CILO-1, CILO-3, CILO-4
4. Design and implement efficient algorithms to model the data and solve the problem. (Execution)	CILO-1, CILO-2, CILO-3, CILO-4
5. Evaluate information and make independent judgements through constructing and inferencing with appropriate data models. (Judgement)	CILO-1, CILO-2, CILO-4
6. Communicate effectively about data science to both laymen and experts. (Communication)	CILO-5
7. Demonstrate self-direction in tackling and solving problems and act autonomously in planning and implementing tasks. (Autonomy)	CILO-1, CILO-6
8. Use a global perspective in conjunction with data analytic techniques to address issues of importance in science, technology, and the society. (International outlook)	CILO-2, CILO-4, CILO-6

5 Grading & Assessment

Assessment Task & Proportion of Final Grade (%)	**Course ILOs this task is to assess**
Course Participation [10%]	CILO-1, CILO-2, CILO-3, CILO-4, CILO-5, CILO-6
Written Assignments [30%]	CILO-1, CILO-2, CILO-3, CILO-4, CILO-5, CILO-6
Midterm test [20%]	CILO-1, CILO-2, CILO-3
Final exam [40%]	CILO-4, CILO-5, CILO-6