The Chinese University of Hong Kong
Department of Linguistics and Modern Languages
Second Term, 2024-25

Course Code: LING3401 **Title in English:** Linguistics and Information Technology (IT) **Title in Chinese:** **語言學與資訊科技**
Course description: This course bridges linguistics and modern information technology by offering a comprehensive introduction to computational linguistics, with a focus on current developments in Natural Language Processing (NLP) and Large Language Models (LLMs). Students will explore core principles of computational linguistics, modern NLP tools, and their applications in linguistic analysis, gaining practical skills in text mining and corpus analysis, as well as understanding and working with Large Language Models. The curriculum emphasizes hands-on experience with user-friendly tools through interactive tutorials and guided projects, enabling students to analyze linguistic data, apply NLP tools to solve real-world problems, and evaluate the capabilities and limitations of language models. Target students include linguistics majors without programming experience, junior college students interested in the field, and anyone looking to understand the intersection of linguistics and AI.

Learning outcomes

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

- Understand the core concepts of computational linguistics, NLP, and LLMs
- Develop basic skills in using LLMs and NLP tools and technologies
- Apply text mining and corpus analysis techniques
- Gain insights into the development of LLMs and their practical applications
- Appreciate how these tools contribute to language sciences, both in theory and in practice

Course syllabus

Topic	**Contents/fundamental concepts**
Introduction to Computational Linguistics	Overview of computational linguistics and its applications.
Corpus Linguistics and Data	Introduction to linguistic corpora and annotation techniques.
Semantic Processing, Syntax and Parsing	Word embeddings and vector semantics; applications in syntactic analysis.
Introduction to LLMs	Applications in text generation and summarization.
Part-of-Speech Tagging	Part-of-Speech Tagging and related Analysis
Sentiment Analysis and Text Classification	Spam detection, topic categorization, and sentiment analysis.

Course components (Teaching modes and Learning activities)

Teaching Modes and Learning Activities	
On-site face-to-face (hybrid or online modes may be available in extreme conditions)	Percentage of time 100%
Lectures	70% (2 hours/week)
Interactive tutorial	30% (0.75 hour/week)

or Laboratory activities	30% (0.75 hour/week)
or Discussion of case	30% (0.75 hour/week)
Out-of-classroom	Percentage of time 100%
Self study	50% (1 hours/week)
Project work	50% (1 hour/week)

Assessment type, percentage, and rubrics

Assessment type	**Description**	**Percentage**
Mid-term and Final Exams	Mid-term and final exams will assess the didactic information presented in the lectures. Mid-term: 20% Final: 20%	40%
Final Project Paper	Design a logical experiment for language sciences or conduct a literature review on a topic on Computational Linguistics, NLP, or LLM.	30%
Experimental report	Select two demos from the class and write a testing report for each one. Describe the tasks or analyses you performed, the tools you used, and the methods you followed to obtain the results from those tools.	20%
Research presentation	Students are grouped to present a paper or application on a specific technique that interests them.	10%

Required and recommended readings

This is the list of recommended readings, and more readings will be announced in class. All readings are posted on BlackBoard (http://blackboard.cuhk.edu.hk).

1. Fasold, R. W. & Connor-Linton, J., (2014). An introduction to language and linguistics (2nd edition). Cambridge University Press (Chapter 3: The structure of sentences; Chapter 7: Language and the brain; Chapter 14: Computational linguistics).
2. Jurafsky, D., & Martin, J. H. (2024). Speech and language processing (3rd ed., online draft).

   (Chapter 2: Regular Expressions and Text Normalization; Chapter 4: Naive Bayes and Sentiment Classification; Chapter 6: Vector Semantics; Chapter 10: Large Language Models; Chapter 12: Model Alignment and Prompting)
3. Eisenstein, Jacob. Introduction to Natural Language Processing. Cambridge, Massachusetts: The MIT Press, 2019. (Chapter 1: Introduction; Chapter 2: Text Classification; Chapter 6: Language Models)
4. Boyd, J. D. (2020). Python for linguists. Cambridge University Press (A beginner-friendly introduction to using Python for linguistics tasks; useful for tutorials; Chapters 1-3: Basic Concepts; Chapter 5: Text Processing).
5. Bird, S., Klein, E., & Loper, E. (2023). Natural Language Processing with Python (Updated for Python 3 and NLTK 3). O'Reilly Media. (Chapters 1-3: Language Processing and Python; Chapter 5: Categorizing and Tagging Words)

Supplementary Readings and Materials:

Alammar, J. (2023). The Illustrated Transformer (Blog post); URL: https://jalammar.github.io/illustrated-transformer/

Language Models are Few-Shot Learners by Tom B. Brown et al. (2020).

Voyant Tools - Text Analysis Tools (https://voyant-tools.org/)

Feedback for evaluation

Students are welcome to give comments and feedback by sending them in written form to the instructor's email address or talking to the instructor.

Grade Descriptors

Grade	Overall Course
A	Demonstrates exceptional understanding of key concepts in linguistics and information technology, including the ability to: clearly explain foundational computational linguistics concepts, such as tokenization and syntactic parsing; critically analyze the advantages and limitations of NLP tools; effectively apply corpus analysis techniques using tools; evaluate ethical considerations related to NLP, such as bias and fairness in LLMs; conduct a comprehensive and well-structured final project showcasing in-depth knowledge and application of learned methods.
A-	Shows a strong understanding of key concepts with minor weaknesses in one area, such as the ability to describe or apply corpus analysis methods or evaluate the limitations of computational tools.
B	Demonstrates a good understanding of the subject with weaknesses in no more than two major areas. A student may: show solid knowledge of computational linguistics and NLP concepts; provide acceptable but less detailed analysis of linguistic data or ethical issues.
C	Demonstrates an understanding of the course material with noticeable weaknesses in several areas, such as incomplete descriptions of NLP concepts or limited application of tools in assignments.
D	Demonstrates minimal understanding of the course material with significant weaknesses in most key components.
F	Fails to demonstrate sufficient understanding of the core concepts, with critical gaps in knowledge and application across the course content.

Course Schedule

Class/ week	Date	Topics and requirements	Tutorial
Week 1	Jan 08	Introduction to Linguistics & IT (Overview of the course and its importance)	*Brief demo of LLM and NLP applications*
Week 2	Jan 15	Language Data and Text Processing	*Word frequency analysis*
Week 3	Jan 22	Basic Text Analysis Methods	*Basic text mining and corpus statistics*
Week 4	Jan 29	*Chinese New Year*	
Week 5	Feb 05	Understanding Language Models	*Probability in language* *Basic statistical concepts*
Week 6	Feb 12	Word Meaning and Embeddings	*Word similarity analysis*
Week 7	Feb 19	Introduction to Modern NLP Tools	*NLTK and spaCy basics and environment setup*
Week 8	Feb 26	Large Language Models Fundamentals	*Mid-term (1.5 hours)*
Week 9	March 05	*Reading Week*	
Week 10	March 12	Language Generation and Understanding	*Evaluation methods &* *Common applications*

Week 11	March 19	Text Classification and Analysis	*Demo of Text Classification*
Week 12	March 26	Parsing and Structure Analysis	*Demo of Parsing and Syntactic Analysis*
Week 13	April 02	Information Extraction & Sentiment Analysis	*Demo of Sentiment Analysis*
Week 14	April 09	Applications in Language Technology	*Machine translation, Chatbots Text summarization, Search engines*
Week 15	April 16	Ethics and Bias in Language Technology	*Final Exam (1.5 hours)*

Contact details for teacher(s) or TA(s)

Professor/Lecturer/Instructor:	Prof.
Name:	FENG Gangyi
Office Location:	G09 KKB (or 401 4/F HYS)
Office Hours:	Thursday 15:30-17:00 or by appt English, Cantonese (native), Mandarin (native)
Telephone:	3943-3190
Email:	g.feng@cuhk.edu.hk
Teaching Venue:	Lee Shau Kee Building (LSK) 302
Class/Tutorial Time:	Wed 14:30-16:15 (Lecture), 16:30 – 17:15 (Tutorial)
Website:	https://neurolanglab.github.io/index.html
Other information:	Google Scholar: Gangyi Feng (冯刚毅)

Teaching Assistant/Tutor:	TA
Name:	Chen Yige
Office Hours and Location:	By appt
Telephone:	
Email:	yigechen@link.cuhk.edu.hk
Teaching Venue:	
Other information:	

Details of course website

All announcements of the course will be posted on Blackboard (https://blackboard.cuhk.edu.hk).

Academic honesty and plagiarism

Attention is drawn to University policy and regulations on honesty in academic work, and to the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details may be found at http://www.cuhk.edu.hk/policy/academichonesty/.

With each assignment, students will be required to submit a signed declaration that they are aware of these policies, regulations, guidelines and procedures.

- In the case of group projects, all members of the group should be asked to sign the declaration, each of whom is responsible and liable to disciplinary actions, irrespective of whether he/she has signed the declaration and whether he/she has contributed, directly or indirectly, to the problematic contents.

- For assignments in the form of a computer-generated document that is principally text-based and submitted via VeriGuide, the statement, in the form of a receipt, will be issued by the system upon students' uploading of the soft copy of the assignment.
- Students are fully aware that their work may be investigated by AI content detection software to determine originality.
- Students are fully aware of the AI approach(es) adopted in the course. In the case where some AI tools are allowed, students have made proper acknowledgment and citations as suggested by the course teacher.

Assignments without a properly signed declaration will not be graded by teachers.

Only the final version of the assignment should be submitted via VeriGuide.

The submission of a piece of work, or a part of a piece of work, for more than one purpose (e.g. to satisfy the requirements in two different courses) without declaration to this effect shall be regarded as having committed undeclared multiple submissions. It is common and acceptable to reuse a turn of phrase or a sentence or two from one's own work; but wholesale reuse is problematic. In any case, agreement from the course teacher(s) concerned should be obtained prior to the submission of the piece of work.

The copyright of the teaching materials, including lecture notes, assignments and examination questions, etc., produced by staff members/ teachers of The Chinese University of Hong Kong (CUHK) belongs to CUHK. Students may download the teaching materials produced by the staff members/ teachers from the Learning Management Systems, e.g. Blackboard, adopted by CUHK for their own educational use, but shall not distribute/ share/ copy the materials to a third-party without seeking prior permission from the staff members/ teachers concerned.

Use of generative AI tools

The use of AI tools is allowed with explicit acknowledgment and proper citation for assignments.
The use of AI tools is prohibited for mid-term and final exams.