# COMP 7045
# *Natural Language Processing & Large Language Model*

## Instructors:
## **Dr. Jing MA**

## Teaching Assistant:
## **Fu Rao**

# Outline

- General information

- Course contents

- Expected learning outcomes

- Assessment methods

# Outline

- **General information**
- Course contents
- Expected learning outcomes
- Assessment methods

# Timetable

- **Time of our classes**
  - 13 Lectures:
    - Jan. 12 – Apr. 20
    - Mon 18:30 – 21:20

  - 2 Public Holidays
    - Feb. 16, Apr. 6

- **Classroom of our classes**
  - SCT502 (HSH Campus)

- **Locations of our labs**
  - RRS638, RRS735

| 2026 | Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|------|-----|-----|-----|-----|-----|-----|-----|
| **January** | | | | | 1 | 2 | 3 |
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| | 25 | 26 | 27 | 28 | 29 | 30 | 31 |

1   The first day of January

| **February** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------|-----|-----|-----|-----|-----|-----|-----|
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| | 22 | 23 | 24 | 25 | 26 | 27 | 28 |

17-19 Lunar New Year holidays

| **March** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------|-----|-----|-----|-----|-----|-----|-----|
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| | 29 | 30 | 31 | | | | |

| **April** | | | | 1 | 2 | 3 | 4 |
|------|-----|-----|-----|-----|-----|-----|-----|
| | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| | 26 | 27 | 28 | 29 | 30 | | |

3,4,7 Easter holidays
6   The day following Ching Ming Festival

| | |
|---|---|
| 9-10 | Exchange Student Orientation |
| 10 | First Semester Ends |
| 12 | Second Semester Begins |
| 12 | Second Semester Classes Begin |
| 12 | First Day to Add/Drop Courses (Second Semester) |
| 14 | Continuing Education School Board Meeting |
| 16-22 | First Semester Make-up Examinations |
| 19 | Transdisciplinary Undergraduate Programme Board Meeting |
| 24 | Last Day to Add/Drop Courses (Second Semester) |
| 26 | Senate Meeting |
| 4 | Creative Arts School Board Meeting |
| 5 | Arts and Social Sciences Faculty Board Meeting |
| 27 | Communication School Board Meeting |
| 1-7 | Christian Emphasis Week |
| 3 | Business School Board Meeting |
| 3 | Science Faculty Board Meeting |
| 4 | Trimester III Tuition Payment Due Date |
| 9-21 | Trimester II Examinations |
| 16 | Senate Meeting |
| 23 | Trimester III Classes Begin |
| 23 | First Day to Add/Drop Courses (Trimester III) |
| 24 | Council Meeting |
| 1 | Creative Arts School Board Meeting |
| 1 | Last Day to Add/Drop Courses (Trimester III) |
| 16 | Arts and Social Sciences Faculty Board Meeting |
| 17 | Chinese Medicine School Board Meeting |
| 22 | Baccalaureate Service |
| 22 | Continuing Education School Board Meeting |
| 24 | Communication School Board Meeting |
| 25 | Last Day of Classes (Second Semester) |
| 28 | Business School Board Meeting |

# Teaching Assistant

- **Mr. Fu Rao**
  - PhD student in Department of Computer Science
- Course duty
  - Assisting in marking course assessments
  - Project Support
  - Tutorial Support (Q&A)
- Contact information
  - Email: csraofu@comp.hkbu.edu.hk
  - Office: DLB 625 6/F
  - Office Hour: (Fri) 2pm-4pm

# Reference Books

- **SPEECH and LANGUAGE PROCESSING**
  - [Daniel Jurafsky](#) and [James H. Martin](#), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition with Language Models, 3rd edition. Prentice Hall. 2025.
  - [https://web.stanford.edu/~jurafsky/slp3/](https://web.stanford.edu/~jurafsky/slp3/)
  - Raschka S. Build a Large Language Model (From Scratch). Simon and Schuster. 2024
  - Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. The MIT Press. 1999

- **Recommended Reading**
  - Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. 1st edition, O'Reilly Media; 2009. ([http://www.nltk.org/book/](http://www.nltk.org/book/))
  - Lane H and Dyshel M. Natural Language Processing in Action. Simon and Schuster. 2025.
  - Jason Brownlee. Deep Learning for Natural Language Processing. Machine Learning Mastery, 2018.

- **We will release the lecture notes, tutorial materials (every week)**
  - Slides cannot replace the textbook.
  - They are at most a guideline.
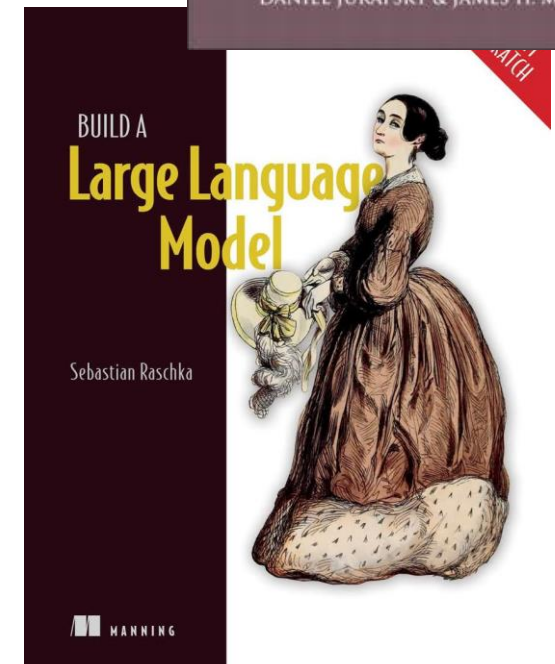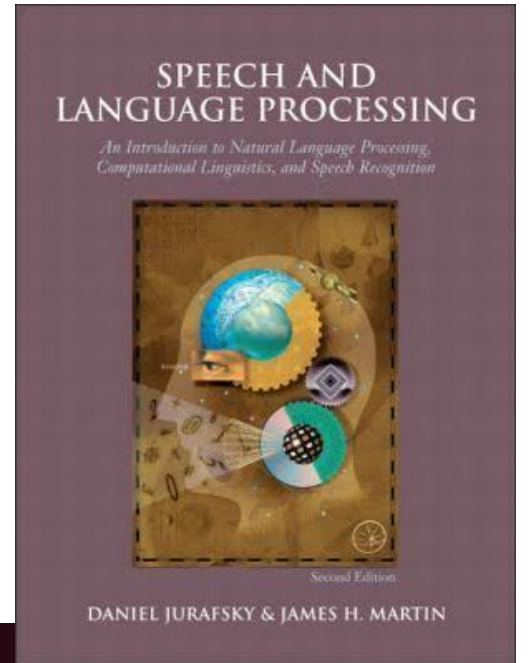
# Outline

- General information
- **Course contents**
- Expected learning outcomes
- Assessment methods

# Course Contents

- This course aims at teaching students the state-of-the-art NLP data analytics, including <span style="color:red">techniques, software, applications</span>, and <span style="color:red">perspectives</span> with textual data.
- The class will cover, but not be limited to, the following *topics*:
    - Text preprocessing such as segmentation, stemming, etc;
    - Statistical Language Models;
    - Syntactic Analysis such as Tagging and Parsing;
    - Word Embedding such as word2vec, GloVe, ELMo, etc;
    - Deep learning model to solve some NLP tasks;
    - Attention and Transformer;
    - Neural Language Models
    - Large Language Models
    - NLP tasks and applications;
- The assignments and project may involve various application domains

# Tutorial Subjects

Grasp basic knowledge about Python and Pytorch/Tensorflow
We will use Google Colab to train & test some models
Some local platforms such as Anaconda

The content cover the following topics:
• <span style="color:red">Textural processing</span>: regular expression, text segmentation, tagging, stemming, etc
• Create a statistical language model, generate text.
• Use a space vector model to represent query and document, information retrieval system.
• Word embedding and visualization tools
• Train toy deep learning models: RNN and CNN
• Train a transformer-based model for classification task
• LLM fine-tuning, a toy LLM for text summarization

# Course Expectations

- Understand what is Natural Language Processing (NLP) and what are the main characteristics of NLP.

- Get familiar with the infrastructure and techniques of NLP, which are associated with real-world data and applications;

- Introduce some the current key issues and the evolution of NLP, and apply software tools for NLP analytics technology

- Get to know the research challenges in the area of NLP and machine learning.

- Gain insights into Large Language Models, train and use LLMs to solve practical problems

# Class Arrangements (may be adjusted as needed)

| Week | Date | Course Topic |
|------|------|--------------|
| 1 | Jan 12 | Course Information & Introduction to NLP |
| 2 | Jan 19 | Text preprocessing |
| 3 | Jan 26 | Statistical Language Models |
| 4 | Feb 2 | Syntactic Analysis |
| 5 | Feb 9 | Lab1 : Hugging Face & Colab, Text Preprocessing, Train a Language Model for Text Generation |
| 6 | Feb 23 | Quiz/In-class Exercise, Word Embedding |
| 7 | Mar 2 | Deep Neural Networks, Attention and Transformer |
| 8 | Mar 9 | Neural Language Models |
| 9 | Mar 16 | Lab2: Train a Text Classification Model, Word Embedding and Visualization, Bert Model and Train your Neural Network |
| 10 | Mar 23 | Large Language Model |
| 11 | Mar 30 | Quiz, Lab3: Prompt Engineering, Train A toy LLM |
| 12 | Apr 13 | NLP Applications, course review |
| 13 | Apr 20 | Mini-Project Presentation |

# Outline

- General information
- Course contents
- **Expected learning outcomes**
- Assessment methods

# Pre-Requisites

> - Excitement about language!
> - Willingness to learn

| Mandatory | Desirable |
|---|---|
| - Basics of probability calculus<br>- English grammar<br>- Data Structures  Algorithms<br>- Python programming | - NLP<br>- Machine Learning<br>- Deep learning |

## This course will NOT cover:

- How to do basic Coding

- How to understand the mathematical mechanisms of NLP models, machine learning and deep learning.

- Generative models for modalities other than text

# Learning Outcomes

- **Knowledge**

  - Describe the fundamental concepts and methodologies of natural language processing and large Language Models

  - Explain the advantages and limitations of methods developed for different scenarios

  - Identify relevant language processing techniques to meet real-world needs

- **Professional Skills**

  - Apply specific methods and techniques in a number of natural language applications

  - Evaluate the solutions designed to technical problems

- **Attitude**

  - Engage in life-long independent learning for problem-solving and advance state-of-the art solutions

# Outline

- General information

- Course contents

- Expected learning outcomes

- **Assessment methods**

# Assessment Methods

- **Evaluation Objective**
  - Spread evaluation over the whole course, not just one exam or one report
- **Continuous Assessment (50%)**
  - **(Take-home / In-class) Assignment(s) + Quizz(es) + Course mini-project**
  - To assess students' mastery of the concepts, techniques and their applications
- **Examination (50%)**
  - Final examination
  - To measure the extent to which the students have reached all of the learning outcomes
- **Important Notices:**
  - **Plagiarism:** Students who plagiarized and who were plagiarized will be given **zero mark**.
  - **Final Exam:** In order to pass this course, students should attain **at least 30% of the final examination mark.**

# More about Projects and Assignments

- **Individual/Group project.**
  - You need to do an oral presentation and submit a report.
  - Choose a topic, literature review, implementation, result analysis
  - Can use any programming languages & third-party libraries
  - Detailed schedule will be announced later
    - Report submission (Tentative deadline: End of week-13)
    - Report presentation ~10 minutes
    - will be graded by the instructor, TA , and other students.

- **Assignments & quiz**
  - release before recess week