AIML 428
Text Mining and NLP

Xiaoying Sharon Gao

Computer Science

Victoria University of Wellington

Menu

- Admin Messages
- Course Organisation
- Introduction to Text Mining
- Introduction to Natural Language Processing

Admin messages

- Lectures are recorded
  - Available on Nuku
  - Only voice and slides are recorded
  - If there are privacy and security issues
    - Must email me before next lecture
    - Private conversation before 11:00 or after 11:50
- Nuku: engagement
  - Quiz, survey, introduce yourself, submit one line of text.
- If you are a distance student enrolled in 2020-2022
  - If you have to do it remotely, email me your name, ID, location, enrolment document

Lectures and our Web site

- Lecturer:
  - Xiaoying Sharon Gao
  - CO 339
  - Ph: 04 463 5978
  - Email: xgao@ecs.vuw.ac.nz
- Lectures:
  - Two lectures each week, Monday , Thursday, 11:00-11:50, AM105
- Course material
  - http://ecs.wgtn.ac.nz/Courses/AIML428_2024T1/
  - Read course outline
  - Assignment page will be updated

Assessment

- Join discussions, peer review: 5%
- Give a presentation: 15%,
  - More details next lecture
    - Choose a topic: new technology or application
    - Sign up for Thursday in Week 2, 4, 6, 8, 10, 12
- Complete a Project:
  - Baseline code: 5%, Friday week 4
  - Full code: 10%, Friday week 6
  - Project report: 15%, Friday week 8
  - Marked in-person in week 7 (need to sign up)
- Write a paper review: 15%, Friday week 10

Test

- Assessment period
- 35%
- Close book, on paper, in person

Topics may cover

- Text representation
- Text classification
- Document clustering
- Opinion mining (Sentiment analysis)
- Information extraction
- Information retrieval, Web page ranking
- Personalized search
- Query expansion
- Recommender systems
- Large language models, transformers
- Machine Translation

The Course Objectives

- Achieve an understanding of the basic problems and basic principles in a variety of related research areas such as text classification and information retrieval.
- Achieve practical experience of building text classification systems.
- Develop skills at reading, understanding, and giving presentations on papers from the research literature.
- Develop skills for further research, including academic writing

Your Background

- Python
  - COMP132, COMP309
- Machine learning:
  - COMP307
  - KNN, NB, SVM, CNN, RNN

The big picture of the course

- Computer Science
  - Artificial Intelligence
    - Machine Leaning and Data Mining
      - Text mining
    - Natural Language Processing

Text Mining

- Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text.

- Typical text mining tasks include
  - text categorization,
  - text clustering,
  - concept/entity extraction,
  - production of granular taxonomies,
  - sentiment analysis,
  - document summarization,
  - entity relation modelling (i.e., learning relations between named entities).

Text mining process

Text Mining Applications

- Risk management
- Knowledge management
- Cybercrime prevention
- Customer care service
- Fraud detection through claims investigation
- Contextual Advertising
- Business intelligence
- Content enrichment
- Spam filtering
- Social media data analysis

Widely used in Industry, Business, Bank, Insurance, Healthcare, Medical, Biology, Marketing, Media, Finance

Natural Language Processing

- Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data.
- Many applications:
  - Machine translation
  - Speech to text
    - Voice is very different: adult vs Child, female vs male
  - Text to speech, Text to image
  - Generating text
  - Question answering
  - Image caption generation
  - Extract information
  - Categorize, organize documents

Why NLP?

- language vs intelligence:
  - The Turing test
- Two main AI areas: Vision and Language
  - It is exciting to see the two areas are merging

- NLP has many challenges
  - Ambiguity: Word Sense, Sentence structure, Pronoun Resolution
  - Knowledge rich: Domain knowledge, Common sense
  - Still hard: Sarcasm, metaphors
  - Incomplete, keep growing and changing
  - Many open research questions, many opportunities

AI History

- Started in 1950s
- Two winters: 1970s, 1980s
  - Bottle neck: common sense
- From 1990s:
  - The Web, large amount of data, super computers

  - Neural Networks
  - Word embeddings, Attention is all you need!
  - Transformers
  - ChatGPT, large language models
  - Sora, Gemini, many more to come

- https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/