COMP 7103A Data Mining

2023-24 Semester 1

About the course

- Instructor:
  Prof. Ben Kao (kao@cs.hku.hk, CYC307)
- Tutor:
  Kevin Lam (yklam2@cs.hku.hk, CYC319)
  Kevin Wu (thwu@connect.hku.hk, CYC430)
- Forum, course announcements, lecture (voice) recording: Moodle

Basic Information

- Lectures and tutorials
  - Wednesday, 0930-1230, MW-T2
- Consultation Hours: (by appointment)
  - Ben: Friday 1600-1800
  - Kevin Lam: Wednesday 1400-1600
  - Kevin Wu: Tuesday 1300-1500

Course Objectives (1)

- Big data analysis
  - Web data
  - Sensor data
  - Scientific and medical data
  - Document data
  - Social Network data
  - Customer data
  - Financial data
  - Computing system data
  - Location data
  - Image/video data
  - Knowledge bases
  - …

data collected by google search:

1. IP
2. query string
3. returned links
4. date/time
5. click-through
6. query traces, behaviours

Course Objectives (2)

- Modeling
  - We analyze data for *pattern discovery*, *model construction*, and *prediction*.

Course Objectives (3)

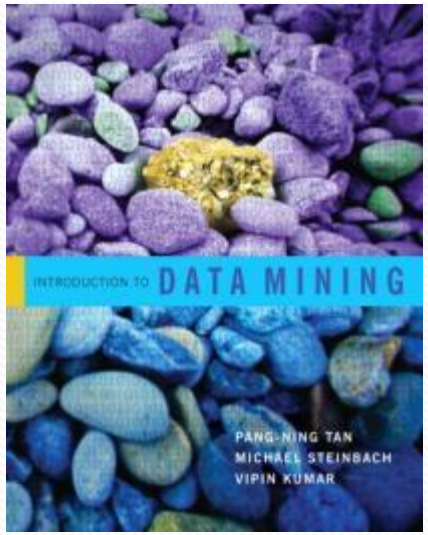
- After taking this course, you should be able to
  - learn data mining methods for analyzing large scale data of different domains
  - apply these methods given some available data
  - preprocess data in an appropriate way in order to apply the desired data mining task
  - post-process data mining results and understand their values

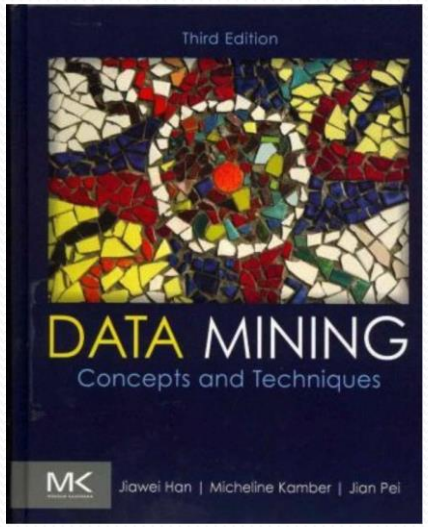
Assessment (Tentative)

- 2 Assignments (40%)
  - Assignments may contain written and/or programming tasks.
  - We will use Weka (http://www.cs.waikato.ac.nz/ml/weka/), an open-source data mining tool.
  - You will preprocess data and then feed them to Weka and/or obtain results from Weka and interpret them.
  - Additional (ungraded) exercises.
- Midterm exam (10% in-class)
  - Tentative date: 8th meeting (Nov 1st)
- Final examination (50%)

Textbook

- Course material based on the textbook
  - *Introduction to data mining*,
    P.-N. Tan, M. Steinbach, V. Kumar,
    Addison Wesley
- Another good reference book:
  - *Data Mining: Concepts and Techniques*,
    Han, Kamber, Pei,
    Morgan Kaufmann





Other Resources

- The Weka open source software:
  - http://www.cs.waikato.ac.nz/~ml/weka/
  - http://weka.pentaho.com/
- Data mining community website
  - http://www.kdnuggets.com/
- Data mining research conferences
  - ACM *SIGKDD*, IEEE *ICDM*, ACM *CIKM*

Topics (Tentative)

Textbook chapters

Week	Topics	References
1-3	Introduction, Data (Types, Preprocessing, Similarity Measures, OLAP)	Chapters 1-2
4-6	Classification	Chapters 3-4
7-8	Association Analysis	Chapters 5-6
8	Midterm	
9-10	Cluster Analysis	Chapter 7-8
10	Research*	Papers

* Will not appear in exam

Tutorials

Week	Tutorial
3	Weka: Data Preparation
4	Weka: Classification
5	Data Mining in Python I
6	Data Mining in Python II
7	Pre-midterm review and exercises/ Assignment 1 review and feedback
9	Recommender Systems
10	Midterm solution Weka: Association rule mining & clustering

Plagiarism

- Departmental guideline:
  - http://intranet.cs.hku.hk/csintranet/contents/general/shared/plagiar.jsp
- First Attempt: Students who admit committing plagiarism for the first time shall be warned in writing and receive a zero mark for the component concerned. For those who do not confess, the case would be referred to the Program Director for consideration.
- Second Attempt: If students commit plagiarism more than once during the course of studies, the case shall be referred to the Program Director for consideration. The Program Director will investigate the case and consider referring it to the University Disciplinary Committee, which may impose any of the following penalties: a published reprimand, suspension of study for a period of time, fine, or expulsion from the University.

Notes

- This course does not require prior knowledge in data mining
- Background needed:
  - linear algebra
  - statistics
  - probability theory
  - algorithms
  - programming