CS5488: BIG DATA ALGORITHMS AND TECHNIQUES

Dr. Ka-Chun Wong (CityU HK)

Syllabus

Week	Lecture Topics (Science)	Lab Topics (Technology)	Remarks
1	Course Outline; Intro. to Big Data	R Tutorial	
2	Data Computing Cycles; R Examples	Lab1 - R Lab Submission	
3	Data Computing - Text Data	Python Tutorial	
4	Data Computing - Time Series Data	Lab2 - Python Lab Submission	Project Grouping Deadline
5	**Midterm**	Hadoop Tutorial	
6	Parallel Computing - Theory	Lab3 - Hadoop Lab Submission	
7	Parallel Computing - Hadoop	Pig Latin Tutorial	
8	Parallel Computing - Spark	Lab4 - Pig Latin Lab Submission	
9	Parallel Computing - Others	Spark Tutorial	
10	Summary	Lab5 - Spark Lab Submission	
11	(Project Time)	(Project Time)	
12	(Project Time)	(Project Time)	
13	(Project Time)	(Project Time)	Project Presentation File Deadline Project Report Deadline

(Please click "Files" on the left menu for further details.)

	Percentage
Project	40%
In-class	5%
Midterm	**15%**
Exam	40%

Links

https://canvas.cityu.edu.hk

http://www.cityu.edu.hk/arro/ac_calendar.asp

Textbooks

- Data Science and Big Data Analytics
- Hadoop: The Definitive Guide
- Lightning-Fast Big Data Analysis
- MapReduce: a flexible data processing tool
- The Hadoop Distributed File System
- Beginning Apache Pig: Big Data Processing Made Easy

Course Materials

- All the course related content, communication, and grading will be posted on CANVAS
- https://canvas.cityu.edu.hk

- http://houseofbots.com/news-detail/2775-4-to-be-successful-at-data-science-think-batman-not-superman

In-class (5%)

The red colors denote the 5 in-class labs; each lab accounts for 1%.

Week	Lecture Topics (Science)	Lab Topics (Technology)	Remarks
1	Course Outline; Intro. to Big Data	R Tutorial	
2	Data Computing Cycles; R Examples	Lab1 - R Lab Submission	
3	Data Computing - Text Data	Python Tutorial	
4	Data Computing - Time Series Data	Lab2 - Python Lab Submission	Project Grouping Deadline
5	**Midterm**	Hadoop Tutorial	
6	Parallel Computing - Theory	Lab3 - Hadoop Lab Submission	
7	Parallel Computing - Hadoop	Pig Latin Tutorial	
8	Parallel Computing - Spark	Lab4 - Pig Latin Lab Submission	
9	Parallel Computing - Others	Spark Tutorial	
10	Summary	Lab5 - Spark Lab Submission	
11	(Project Time)	(Project Time)	
12	(Project Time)	(Project Time)	
13	(Project Time)	(Project Time)	Project Presentation File Deadline Project Report Deadline

(Please click "Files" on the left menu for further details.)

	Percentage
Project	40%
In-class	5%
Midterm	**15%**
Exam	40%

Midterm (15%)

- Based on the lecture notes and tutorial / lab materials
- HKEAA-approved calculators are allowed.
  - http://www.hkeaa.edu.hk/DocLibrary/IPE/cal/CAL2016.pdf

- Objective:
  - To review the knowledge in the first-half of the course so that students can prepare for the final exam in a rigorous manner.

Final Exam (40%)

- 30% of the final exam mark must be obtained to pass the course. (i.e. 30/100)
- Based on the lecture notes and tutorial / lab materials.
- Will be announced later by the university administration.

- Objectives:
  - To assess the capability of students to
    - Identify data computing problems
    - Review the existing concepts in data computing
    - Review the existing technology in data computing
    - Develop data computing solutions
    - Accelerate data computing solutions by parallel computing
    - Apply data computing solutions with specific case studies

Open Project (40%)

- To be consistent with the CityU discovery-enriched curriculum, each group has to identify an interesting problem and propose a data computing solution to solve the problem with parallel computing elements.
  - Hard Deadline: The Saturday of Week 13.
- **3 to 4 members**
- A project cover sheet template has been provided for you.
- Deliverables:
  - Project Cover Sheet
  - Project Report
  - Supporting Materials
- Please submit your project deliverables on CANVAS https://canvas.cityu.edu.hk
- Late submissions are not graded and will be given 0 mark.

Open Project (40%)

Schedule

Before Week 4: Join a group on Canvas. (Leftovers will be assigned randomly)
Week 4: Grouping confirmed -- no more change of grouping.
Week 6: You are advised to confirm a project topic with your groupmates for work load distribution.
Week 12-13: Project presentation during lecture time. The presentation schedule will be randomly assigned.
Week 13: Submit all final deliverables on Canvas.

Open Project (40%)

Report	
Real World Impact / Creativity	/ 5
Solid Works and Output Amount	/ 20
Technical Depth and Correctness	/ 20
Parallel Computing Elements	/ 20
Use of Written English	/ 5
Presentation	
Technical Presentation Amount	/ 20
Technical Presentation Skills	/ 5
Question and Answer (Q&A)	/ 5
	/ 100

Open Project (40%)

Project Example: (More past projects in CANVAS)

"Parallel Data Computing Solutions to Hong Kong Real Estate Data"

1. Collect the Hong Kong real estate data from several sources.
   - Document the source of the data clearly in the report (e.g. https://data.gov.hk/en/).

2. Preprocess and Visualize the data with histograms, scatterplots, and other diagrams you have learned;
   - Preprocess the data so that you can visualize it.
   - Implement data visualizations so that we know better about the data.

3. Analyze the data and discuss your own findings
   - Perform advanced analysis on the data (e.g. data clustering and association rule mining)
   - Explain the findings, and try to make conjectures about the findings you obtained.

4. Discuss how parallel computing is applied to accelerate the data computing process
   - Describe what kind of parallel computing strategy you have implemented (e.g. parallel for loop)
   - Explain why such a parallel computing strategy has been adopted (e.g. memory hierarchy)

5. Conclusion and Future work.
   - State your conclusions and the related pros / cons.
   - If you have enough time, what you can do? What problems are there to be investigated further?

Open Project (40%)

- Possible Data Sources: (but not limited to)
- (You are encouraged to find your own datasets you are interested in; below are just examples that you can choose.)

- Hong Kong Government Data: https://data.gov.hk/en/
- US Government Data: https://www.data.gov/
- Singapore Government Data: https://data.gov.sg/
- UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/
- Panama Papers Graph Data (i.e. Network): https://github.com/amaboura/panama-papers-dataset-2016
- Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/
- Offshore Leaks Database (i.e. Text Data): https://offshoreleaks.icij.org/
-
- Miscellaneous:
- http://www.kdnuggets.com/2011/02/free-public-datasets.html
- https://r-dir.com/reference/datasets.html
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- http://www.datasciencecentral.com/page/search?q=data+sets

Open Project (40%)

- Possible Project Ideas: (but not limited to)
  - Analyze factors relating the gaming performance in League of Legends
  - Exploration of Factors Relating to Movie Box Office Performance
  - Historical Buildings in Hong Kong
  - FIFA players' statistics and Professional Football Clubs' Seasonal Performance
  - A visual exploration of aircraft crashes since 1908
  - NBA in Data: An analytical report on Los Angeles Lakers
  - Hong Kong Housing Trend
  - Gastronomy and Ingredients Matching Across the World
  - Exploring of factors relating to League of Legend world championship performance
  - The frequency of earthquakes
  - Homeless, Hong Kong
  - The Relationship among Gender, Education and Employment in Hong Kong
  - Renewable energy in the European Union
  - Flight Networking and On-time Performance Analysis
  - Analysis of Factors Affecting Global Temperature Rise

Open Project (40%)

- Possible Project Ideas: (but not limited to)
  - Secondary School in Hong Kong
  - World University Rankings and Statistics
  - Exploring currency exchange rate
  - Mass Shooting in America
  - An evaluation of workplace environment in Hong Kong
  - Shootings in NBA
  - Exploration of typhoon in Hong Kong in 21st century
  - IMDB Movie Analysis
  - Data mining in conditions and predictions of G20 countries by continent
  - The Analysis of Mandatory Provident Fund (MPF) Schemes
  - Understanding people's reactions to new movies cia Twitter and film review websites
  - Mobile Application (ios and android system) Ranking and the relevant factors on America market
  - Unemployment rate and major indices of US, Germany and Japan
  - Analysis on the 2016 Legislative Council Election

Academic Honesty at City University

- City University is committed to high standards of academic honesty, and students are expected to ‘present their own work, give proper acknowledgement of other’s work, and honestly report findings obtained’.

- http://www.cityu.edu.hk/provost/academic_honesty/rules_on_academic_honesty.htm

- http://www6.cityu.edu.hk/ah/academic_honesty.htm

VeriGuide

- Plagiarism Detection Engine
- http://veriguide1.cse.cuhk.edu.hk/portal/

Contact

- https://canvas.cityu.edu.hk **(primary contact)**

- Unless personal, please post your technical questions on the CANVAS system so that other students can have the equal opportunities to have a look at the answers.

- Dr. Ka-Chun Wong
  - Consultation Time: After Each Lecture
  - kc.w@cityu.edu.hk (urgent contact only)

END