Xi'an Jiaotong-Liverpool University
西交利物浦大学

School of Advanced Technology

MODULE HANDBOOK

INT303
Big Data Analytics

Jia WANG

Semester 1

2021-2022

SECTION A: Basic Information

❑ Brief Introduction to the Module

INT303 is for Big data Analytics. It is a specialty module because it's direct relations with applications and job perspectives. It is a highly practical module and in the same time it also lays down many research topics for the further study in master level.

The module is structured in four parts to cover four basic demanding related with the applications.

- *Part I, the basics. it covers the concepts, history and related subjects of Big Data Analytics to clarify the doubts and lay the foundations in conceptual level.*
- *Part II, the big data collection and visualization. It introduces the popular tools available for big data collection and visualization.*
- *Part III, the big data platform. It introduces the popular platforms available for big data processing.*
- *Part IV, the data processing methods and algorithms. Typical and classical data analysis methods and algorithms are covered to provide mathematical and engineering solutions to applications.*
- *Part V, the data processing ethics and social issues. It is strongly recommended to understand the technology with its social obligations. Use them wisely and responsibly.*

❑ Key Module Information

Module name: *Big Data Analytics*
Module code: *INT303*
Credit value: *5*
Semester in which the module is taught: *1*
Pre-requisites needed for the module: *None (Prefer: Statistics, Algorithms, Distributed Computing, HPC, Java, R, and Python)*
Programmes on which the module is shared: *ICS, CST, IMS, DMT*

❑ Delivery Schedule

- *See timetable.*
- *Week1-4: We will use Zoom: (if BBB does not work):*
  - *Time: 11:00am to 13:00pm on Thursday.*
  - *ID: 634 821 1678*
  - *Password: int303*
- *Week 5-14:*
  - *Time: 11:00am to 13:00pm on Thursday.*
  - *Location: IBSS Building-BSG02*

Module Leader and Contact Details

Name: Jia WANG

Brief Biography: *Dr. Jia Wang is currently a lecturer in Xi'an Jiaotong-Liverpool University. She received a Ph.D. degree from the Hong Kong Polytechnic University, M.S. degree from KTH Royal Institute of Technology, B.S. degree in communication engineering from Beijing Jiao Tong University. She had visited the University of Southern California as a visiting scholar. Her research interests span the broadly defined areas of graph mining, reinforcement learning, game theory, and multi-agent systems.*

Her publications can be founded via Google Scholar：

https://scholar.google.com/citations?hl=en&user=CcJikxgAAAAJ

Email address: *jia.wang02@xjtlu.edu.cn*

Office telephone number: *+86 (0)512 81889047*

Room number and office hours: *SD537*

Preferred means of contact: *Email*

- **Additional Teaching Staff and Contact Details**

None

SECTION B: What you can expect from the module

- **Educational Aims of the Module**

1. To introduce the environment and the main application domains where Big Data Analytics (BDA) takes place;

2. To introduce general framework and process of BDA;

3. To study technologies and algorithms that support BDA;

4. To study platforms, tools that are currently used in BDA;

5. To gain an understanding of the best practice in BDA.

- **Learning Outcomes**

Upon completing this module, a student will be able to:

A. demonstrate a solid understanding of processes and issues related to Big Data Analytics (BDA);

B. identify applications of BDA that can help improve business operations;

C. determine the appropriate use of technologies, tools, and software packages to support data analysis involving practical scenarios;

D. be proficient with at least one data analytics software package.

- **Assessment Details**

Sequence	Method	Learning outcomes assessed	Duration	Timing	% of final mark	Resit
#1	*Lab1: Data Scraping*	*All*	*See notice*	*S2*	*15%*	*NO*
#2	*Lab2: Big Data Competition*	*All*	*See notice*	*S2*	*15%*	*NO*
#3	*Written Exam*	*All*	*See notice*	*S2*	*70%*	*NO*

- **Methods of Learning and Teaching**

1. Students will be expected to attend two hours of "formal lectures" a week.
2. Students will be expected to attend a two-hour of "online tutorial/Q&A" to answer issues related to the lectures and lab sessions every 4 weeks.
3. Students will be expected to attend in a two-hour "supervised practices" in own computer and expected to spend more times for the practical tasks.
4. Lectures will introduce the academic content and practical skills which are the subject of the module, while computer practices will allow students to practice those skills.
5. In addition, the students will be expected to devote roughly three more hours of unsupervised time for each lecture hour to solving continuous assessment tasks and private study. Private study will provide time for reflection and consideration of lecture material and background reading.
6. Two periodical reports and one final written exam are expected as the assessment of the module. The deadlines will be announced in the lectures.

- **Syllabus & Teaching Plan**

1. Introduction to Big Data Analytics (3 lectures, each lecture last for 2 hours)
 # Lecture 1: What is big data analytics? (Advanced analytic techniques operate on big data sets). Differentiate with related concepts: such as: data mining, data analysis, data visualization, and statistics
 # Lecture 2: What is data and big data?

- *Defining data and its attributes*
- *Data attributes types and data types*
- *Big data and its types*
- *Data sources, the value of the big data*

 # Lecture 3: Big data Gramma

- *Python operations related to big data analysis*

2. Big data collection and visualization. (2 lectures, each lecture last for 2 hours)
 # Lecture 4: Data collection and Data scraping

- *What is Web Service?*

- *Data Scraping*
- *Gathering data from APIs*

 # Lecture 5: Data Virtualization

- *Visualization motivation*
- *Principle of Visualization*
- *Types of Visualization*

3. Systems and software (1 lecture, each lecture last for 2 hours)

 # Lecture 6: Infrastructure that supports Big Data processing

- *Large-scale computing*
- *Distributed file system*
- *MapReduce: Distributed computing programming model*
- *Spark: Extends MapReduce*

4. The data processing methods and algorithms. (7 lectures, each lecture last for 2 hours)

 # Lecture 7: How to tell a good story/ business intelligence analysis

Large-scale computing

- *Know your audience*
- *How to tell a story*

 # Lecture8: Representing Data and Engineering Features

- *Data Features*
- *Feature Selection*

 # Lecture9: Dimensionality Reduction

- *High Dimensionality*
- *Principal Components Analysis*

 # Lecture10: Big Data Analysis Models

- *Types of Machine Learning Algorithms*
- *Popular Algorithms with Hands-On Demo:*
- *Linear Regression*
- *Logistic Regression*
- *K Nearest Neighbor*

 # Lecture11: Bagging

- *Bagging*
- *Out-of-Bag Error*
- *Improving on Bagging*
- *Random Forests*

 # Lecture12: Boosting

- *Boosting*
- *Gradient Boosting*
- *Set-up and intuition*
- *Connection to Gradient Descent*
- *The Algorithm*
- *AdaBoost*

 # Lecture13: Stacking

- *Motivation for Stacking*
- *Subsemble Approach*

5, the data processing ethics and social issues. (1 lecture, each lecture last for 2 hours)

 # Lecture14: Ethics and social issues.

❑ Lab Schedule

Week 2- 4

Student	Time	Day	Venue	Lecturer/Instructor
Group 1	*13:00-15:00*	*Tuesday*	*Online*	*JiaWANG/ TAs*
Group 2	*13:00-15:00*	*Tuesday*	*Online*	*JiaWANG/ TAs*
Group 4	*13:00-15:00*	*Tuesday*	*Online*	*JiaWANG/ TAs*

Week 5- 14

Student	Time	Day	Venue	Lecturer/Instructor
Group 1	*9:00-11:00*	*Wednesday*	*SD546, SD554*	*JiaWANG/ TAs*
Group 2	*11:00-13:00*	*Wednesday*	*SD546, SD554*	*JiaWANG/ TAs*
Group 4	*13:00-15:00*	*Tuesday*	*SD546, SD554*	*JiaWANG/ TAs*

❑ Reading Materials

Mandatory textbook is a required book in either print or electronic format for a module that students are obligated to purchase.

Title	Author	ISBN/Publisher
Foundation of Big Data Analytics: Concepts, Technologies, Methods and Business	***Gangmin Li***	*978-7-03-058148-8/China Science Publisher*

Optional textbook is a book in print that students can choose to purchase or not.

Introduction to Machine Learning with Python: A Guide for Data Scientists	***Andreas C. Müller, Sarah Guido***	*978-1-44-936941-5/ O'Reilly Media, Inc.*

Reference textbook is a book in print that is considered additional or recommended reading by academic staff and is only purchased for Library's collection where it can be offered for loan.

Title	Author	ISBN/Publisher
Ref 1: Big Data Analytics: Turning Big Data into Big Money.	*Frank J. Ohlhorst 2014*	*978-1-118-14759-7/ John Wiley*
Ref 2: Mining of Massive Datasets	*A.Rajaraman and J.D. Ullman*	*978-1-207-01535-7/ Cambridge*
Ref 3: Data Analysis with Open source Tools	*Philipp K. Janert*	*978-7-5641-2674-2/ O'Reilly*
Ref 4: Introduction to Data Mining	*P. N.Tan, M. Steinbach and V. Kumar*	*978-7-111-31670-1/ Pearson*
Ref 5: Big Data Analytics	*Philip Russom*	*TDWI Research, IBM*
Ref 6: Big data: A revolution that will transform how we live, work and think	*Viktor Mayer-Schonberger and Kenneth Cukier*	*978-1-84854-792-6/ John Murray*
Ref 7: Software for Data Analysis: Programming with R (Statistics and Computing)	*John M. Chambers*	*Springer*

SECTION C: Additional Information

- **Attendance**

  Students who are able to be on campus are reminded of the Academic Policy requiring no less than 50% attendance at classes. Failure to observe this requirement may lead to failure or exclusion from retake examinations in the following year.

- **Student Feedback**

  The University is keen to elicit student feedback to make improvements for each module in every session. It is the University policy that the preferred way of achieving this is by means of an Online Module Evaluation Questionnaire Survey. Students will be invited to complete the questionnaire survey for this module at the end of the semester.

  You are strongly advised to read the policies mentioned below very carefully, which will help you better perform in your academic studies. All the policies and regulations related to your academic study can be found in 'Student Academic Services' section under the heading "Policies and Regulations" on **E-bridge**.

- **Plagiarism, Cheating, and Fabrication of Data.**

Offences of this type can result in attendance at a University-level committee and penalties being imposed. You need to be familiar with the rules. Please see the "Academic Integrity Policy" available on e-Bridge in the 'Student Academic Services' section under the heading 'Policies and Regulations'.

- **Rules of submission for assessed coursework**

  The University has detailed rules and procedures governing the submission of assessed coursework. You need to be familiar with them. Details can be found in the "Code of Practice for Assessment" available on e-Bridge in the 'Student Academic Services' section under the heading 'Policies and Regulations'.

- **Late Submission of Assessed Coursework**

  The University attaches penalties to the late submission of assessed coursework. You need to be familiar with the University's rules. Details can be found in the "Code of Practice for Assessment" available on e-Bridge in the 'Student Academic Services' section under the heading 'Policies and Regulations'.

- **Mitigating Circumstances**

  The University is able to take into account mitigating circumstances, such as illness or personal circumstances which may have adversely affected student performance on a module. It is the student's responsibility to keep their Academic Advisor, Programme Director, or Head of Department informed of illness and other factors affecting their progress during the year and especially during the examination period. Students who believe that their performance on an examination or assessed coursework may have been impaired by illness, or other exceptional circumstances should follow the procedures set out in the "Mitigating Circumstances Policy", which can be found on e-Bridge in the 'Student Academic Services' section under the heading 'Policies and Regulations'.

- **Viewing Exam Scripts**

  Students are permitted to see their answer sheets for final and resit exams under the stated conditions. A student must apply to review their answer sheet within two weeks of the results being published on e-bridge. If a request is received before this deadline the relevant departmental administrative support staff will arrange a time for the student to view their script in the company of an appointed staff member.

  The following conditions apply for students to view their exam scripts.

  1. *Students will be afforded a set time period in which they are allowed to view their exam script (e.g. a limit of 5 minutes).*
  2. *Students have no right to discuss their mark, ask for their marks to be adjusted, or demand more marks.*
  3. *Students are only allowed to see their own script individually and cannot be accompanied by another party other than the appointed staff member.*

4. *Students are not allowed to write notes or take photos during the appointment.*
5. *There is no requirement to provide students with access to the exam paper or solutions.*
6. *The module leader may request an additional staff member to be on-site during review so as to avoid any potential conflict.*

Please note that the purpose of viewing exam scripts is to check for mistakes in your submission and gauge the quality of your work, not to discuss the grading or question your mark.

- **Learning Mall**

  Copies of lecture notes and other materials are available electronically through Learning Mall, the University's virtual learning environment.