UNIVERSITY OF WOLVERHAMPTON
KNOWLEDGE ▪ INNOVATION ▪ ENTERPRISE

6CS030 Big Data
2023/24
Portfolio

Learning Outcomes for the Module:

LO1 – Apply appropriate theory, tools and techniques to problems associated with big data.

LO2 – Synthesise solutions to problems from the big data domain.

LO3 – Analyse and evaluate solutions to big data problems.

LO4 – Present results of solutions using appropriate methods.

The portfolio is made up of the following elements:

Element	Type	Learning Outcome	Portfolio Percentage
1	Coursework. Hand-in date (Week 12): **29th April, Friday 2:00 pm**	LO1, LO2, LO3, LO4	70%
2	Time Constrained Assessment. Provisional date (Week 12): **Monday during workshop time**	LO1, LO2, LO3, LO4	30%

Passing the Module and Resits

To pass the module you must achieve 40% overall in the portfolio. Should you not achieve a pass mark, you will need to resit the part of the portfolio that you did not pass. For the coursework you will have to resubmit the report and code detailed in this document, feedback will be given on what needs to be done to complete the work. For the TCA a new test will run during the University's resit week (Week 12: **Monday during workshop time**).

Note, the overall mark will be capped to 40% if you are resitting the module.

Students with mitigation will have to complete the part(s) not attempted.

Ensure you check your feedback to find out if you need to resit any part of the module.

Teamwork

You can work in groups of 2 for the Coursework (Element 1). The TCA (Element 2) is individual work.

UM2 Apprenticeship Students

You should follow the same requirements as detailed here, but the datasets should be related to your work environment.

Coursework Requirements (Element 1 – 70% of Module Total)

Part a: Report on Practical Work (80%)

You are required to either

- produce a report on a topic related with Big Data analytics (**data analysis topic**) or
- produce a report on creating a search engine to provide access to a data collection (**search engine topic**).

A set of sample topics and datasets will be provided in the Appendix A.

To aid this, you are required to carry out the following:

- The topic should cover some material taught in the lecture (e.g., big data analysis process, search engines, case studies)
- Select a number (one or more) of datasets to analyse or to make searchable:
    - The datasets should have a connection. For example, you may want to look at crime patterns and relate it to the socio economic data found in Census data.
    - Or you might create your own dataset, e.g. a university search engine could scrape information from the university web site
    - Example data sources can be found in Appendix B.
- Clean the data as required (this can be before or after importing it).
- Perform some analysis of the data to derive some useful information, as can be obtained by the data set selected.
- Provide visualisation of the analysis through any technologies you think suitable or, in the search engine case, index the data and make it searchable.
- Produce a report on the process undertaken and findings.

Further information

For the data analysis topic, the investigations of the above datasets need to use the techniques used on the module to process and analyse the data. Lecture 2 outlines a Pipeline to follow, which you may wish to follow:

- Acquire data
- Prepare or Process data
- Analyse data
- Report, index or visualise data
- Act

You should write up the results of your investigation in a report.

The Report

Data Analysis Topic

For a data analysis topic, your report should cover the following areas:

Item	Description	Marks
Title	The title must reflect your topic. Please choose an interesting topic which will answer three questions: For example: Big Data Analytics for Mitigating the Adverse Economic Impact of Covid-19 Lockdown Practices: - What is the Problem that you are going to solve? - How are you going to solve? - Where is the targeted domain?:	3
Abstract	Abstract should have the following points: - Generic Information (1 Sentence) - Problem Statement (2 Sentences) - Aim/objective of the Report: (1 Sentence) - Contributions of the work connected with Methodology (2-3 Sentence) - Social/Research community Impact of the work (1 Sentence)	5
Background of the study	- Generic Information (1 Paragraph) - Problem Statement: Need to explain as a layman aspects, with statistical information and pictorial view. (1 Paragraph) - Aim/objective of the work: (1 Paragraph) - Contributions of the work connected with Methodology (should be itemised) - Organization of the Report (1 Paragraph)	10
Related Work	This section discusses existing relevant works. This can be based on the category of the existing work; in the final paragraph you should mention how your work differs from existing works.	7
Methodology	Here you should discuss the methodology of your entire work process with a Block Diagram/Phases. Discuss each block/phase in details. Should also cover aspects (if appropriate) such as • Data Cleaning • Chosen model	20
Result and Discussion	Should discuss: - Experimental Setup - Discussion and analysis of the findings	20

	Please use suitable data visualisation techniques as appropriate	
Conclusion	Highlight the overview of your work and the findings	5
Reference	Around 15 references from, e.g, IEEE Xplorer, ACM, Elsevier and Springer (not more than last 5 years)	3
Template	IEEE conference format (Max 8 pages): https://www.ieee.org/conferences/publishing/templates.html	2
Plagiarism Report	Maximum 20% excluding the reference	5
Total		80

Part b: Code Appendix (20%)

You must submit sufficient technical supporting material to verify what you have done. You may use any analysis tools or techniques covered in the module.

Your submission must include all supporting technical material required to replicate your findings. This could be in the form of an appendix in your document for Part a, which keeps a record of what steps were taken to process and clean the data if necessary. If you are familiar with Python, you might want to consider using a Notebook to do this instead. You must provide sufficient information to allow others to follow and verify your findings.

Search Engine Topic

For the search engine topic, your report should cover the following areas:

Item	Description	Marks
Title	The title must reflect your topic. Please choose interesting topic which will answer three questions: For example: Search Engine for Covid-19 publication - What are the information needs your search engine supports? - What is the targeted domain? - How do you want to provide a search service?	3
Abstract	Abstract should have the following points: - Generic Information (1 Sentence) - Problem Statement(2 Sentences) - Aim/objective of the Report: (1 Sentence) - Contributions of the work connected with Methodology (2-3 Sentence) - Social/Research community Impact of the work (1 Sentence)	5
Background of the study	- Generic Information (1 Paragraph) - Problem Statement: Need to explain as a layman aspects, with statistical information and pictorial view as appropriate. (1 Paragraph) - Aim/objective of the work: (1 Paragraph) - Contributions of the work connected with Methodology (should be itemised) - Target domain - Information needs (what would people search for, and why?) - Organisation of the Report (1 Paragraph)	10

Related Work	This section discusses the existing related works. You can discuss this section based on the category of the existing work; in the final paragraph you should mention how your work differ with the existing works.	7
Methodology	Here you should discuss the details methodology of your entire work process, e.g. with a Block Diagram/Phases. Discuss each block/phase in details.	20
Artefact and Discussion	Describe your search engine in terms of - Software architecture - Retrieval function or approach - Ranking details (e.g., calculation of scores) Also should cover the following things: 1. Data Set 2. Search interface 3. Indexing details 4. Example searches and results	20
Conclusion	Highlight the overview of your work and the findings	5
Reference	Around 15 references from, e.g., IEEE Xplorer, ACM, Elsevier and Springer (preferably not more than last 5 years)	3
templet	IEEE conference format (Max 8 pages): https://www.ieee.org/conferences/publishing/templates.html	2
Plagiarism Report	Maximum 20% excluding the reference	5
Total		80

Part b: Code Appendix (20%)

You must submit sufficient technical supporting material to verify what you have done. You can implement your solution in a programming language of your choice.

Your submission must include all supporting technical material required to replicate your findings. This could be in the form of an appendix in your document for Part a. You must provide sufficient information to allow others to follow and verify your implementation.

Submission Deadline: Week 12 (Friday-2:00 PM)

Each student should upload an electronic copy on Canvas. The version must be the same for each team.

Ensure you keep a copy of your work.

Teamwork

PLEASE NOTE:

You can work in groups of two for the Coursework, but no more than two. However, on no account should you work on the assignment with others groups to produce a larger group answer.

Ensure your coursework clearly states who the two team members are. Include a peer review that states what each member's contribution was to the assessment (should be equal).

The same mark will be allocated to both team members for this part, unless there are issues highlighted by the peer review. In such cases the module leader will discuss the issues with both members of the team.

Grade Attainment Criteria

See the above 6CS030 marksheets for further details on the marking criteria for all of the above.

Time Constrained Assessment (Element 2 – 30% of Module Total)

Time Constrained Assessment (TCA). This will be a multiple choice question (MCQ) style test with 30 questions.

Provisional date: Monday April 22 during workshop time

Appendix A

Sample Topics:

- Big Data on Healthcare
- Big Data on Education
- Big Data on Agriculture
- Big Data on Security
- Big Data on e-Governance
- Big Data on Weather Forecasting
- Big Data on Online Social Networks
- Big Data on Computational Social Science
- Big Data on System Modelling
- Big Data on Parallel Computing
- Big Data on Time Series Analysis
- Big Data on Business analytics
- Big Data on Internet of Things
- Big Data on Supply Chain Management
- Big Data on Cyberattacks
- Big Data on Image/video Processing
- Big Data on Energy management
- Big Data on Medical and Health Informatics
- Big Data on Urban Informatics
- Big Data on Environment and climate
- Big Data on Data Security and Privacy
- Search engine on COVID-19 publications
- Academic search engine (e.g. to find a suitable supervisor)
- Search engine for selected files on your laptop (desktop search)
- Dataset search (find a suitable dataset)
- Any dataset from Kaggle (http://www.kaggle.com/)
- etc.

Appendix B

Sample datasets

The following are given as example sources for data:

Name	URL
Nomis: Provided by the Office of National Statics. Includes Census data and a number of other datasets.	https://www.nomisweb.co.uk/ https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp
Police data	https://data.police.uk/data/archive/ (CSV) https://data.police.uk/docs/ (JSON)
CPS case outcomes	https://www.cps.gov.uk/cps-case-outcomes-principal-offence
Central government data, local authorities and pubic bodies (UK) For example: • destinations of key stage 4 and 16 to 18 (KS5) students • database of registered common land	https://data.gov.uk/ https://www.gov.uk/government/collections/statistics-destinations https://data.gov.uk/dataset/database-of-registered-common-land-in-england
Northern Ireland Open Data Sets	https://www.opendatani.gov.uk/dataset?res_format=JSON
US Government data	https://catalog.data.gov/dataset https://catalog.data.gov/dataset?res_format=JSON
Historical weather data	https://www.ceda.ac.uk/blog/uk-weather-station-records-now-freely-available-to-all-midas-open/
Getting Twitter data for Academic Research See these for how to get twitter data using Python	https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data http://www.tweepy.org/
NBN atlas databases (Natural World)	https://registry.nbnatlas.org/public/datasets
List of datasets for machine-learning research	https://stringfixer.com/tr/List_of_datasets_for_machine_learning_research

Public Datasets for Machine Learning and Data Science	https://pub.towardsai.net/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f
Machine learning datasets	https://www.datasetlist.com/
CORD-19 dataset containing COVID-19 and coronavirus-related research	https://allenai.org/data/cord-19

You are not restricted to these websites only. Do note, any data acquired should not be offensive in any way!

You may need to setup an account to access some of these datasets, you are not required to subscribe to any websites that require payment.