220CT – Data and Information Retrieval
Courswork

This is your 220CT coursework. There are 4 tasks in this CW. All Tasks are mandatory. All submissions through moodle.

Task 1, Part 1: DB Design (15 marks)

The International Space Station (ISS) is a habitable artificial satellite in low Earth orbit. It is the ninth space station to be inhabited by crews following previous orbital stations that were launched by the US the former Soviet Union and later Russia. The ISS is intended to be a laboratory, observatory and factory in space as well as to provide transportation, maintenance, and act as a staging base for possible future missions to the Moon, Mars and beyond. In order to support the crew and overall operation of ISS the space agencies in charge of running the station conduct regular missions to launch spacecraft carrying payloads of essential or replacement equipment up to ISS. A payload inventory, see table below, is recorded of each mission, consisting of the space agency leading the mission and the equipment payload to be sent up to ISS. The overall weight of the payload is also determined in order to calculate the fuel needed for orbital insertion of the spacecraft to successfully rendezvous with ISS.

Mission No.	Agcy No.	Lead Agency	Country	Mission Date	Equipment	Qty	Item Weight	Total Weight
ISS-2237	178	JAXA	Japan	14/12/2013	Potable water dispenser	2	*100kg*	211kg
					Flexible air duct	6	0.5kg	
					Small storage Rack	4	2kg	
ISS-3664	526	ESA	EU	16/01/2014	Bio filter	6	0.20kg	1.20kg
ISS-2356	167	NASA	USA	12/02/2014	Small storage Rack	3	2kg	69kg
					Battery pack	2	5kg	
					Urine transfer tubing	2	1.5kg	
					O2 scrubber	1	50kg	
ISS-1234	032	Roskosmos	Russia	16/04/2014	Small storage Rack	1	2kg	2.5kg
					Flexible air duct	2	0.5kg	

Deliverable:
Using SQL, implement the database above. To do so, you'll need to normalise the table (should that be required), identify the attributes, create and Entity-Relationship Diagram and create the tables using SQL commands.

Task 1, Part 2: (15 marks)

The NASA exoplanet dataset archive can be found here:

http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets

You are asked to design a Database solution for the data above. What would you use? How? Why did you make that choice? What are the advantages and disadvantages?

Deliverable:

Your solution must include the following:

1. The DB solution of your choice.
2. A detailed explanation of how these data will be stored and accessed in the DB you suggest.
3. The benefits of this solution in relation to the data above and its size.
4. The QoS (such as scalability) provided/should be provided to the user should this solution be adopted.

Task 2: Presentation (20 marks)

You will be required to conduct research one of the subjects below and explain how big data/data science is being currently used as a solution to help prevent them. Then present your arguments on the ethical and privacy issues you may encounter.

Domains:

- Fraud detection
- Phishing detection
- Identity theft

Deliverable:

A scientific poster containing the following sections:

- A description of the domain/problem and why you chose it.

- A description of how Big Data/Data Science is being used within the domain, including an overall description of the specific techniques and technologies that are used (showing evidence of research).

- Reflection on whether Big Data/Data Science solutions are successfully meeting business objectives.

- A description on how you think that Big Data/Data Science can be further used within that domain.

- Analysis of how the Big Data/Data Science ideas and solutions in your domain of study could be expanded to other Security domains and how knowledge and experience can be transferred.

- Conclusion and closing remarks.

The report should be accompanied by a list of resources.
Poster templates on moodle.

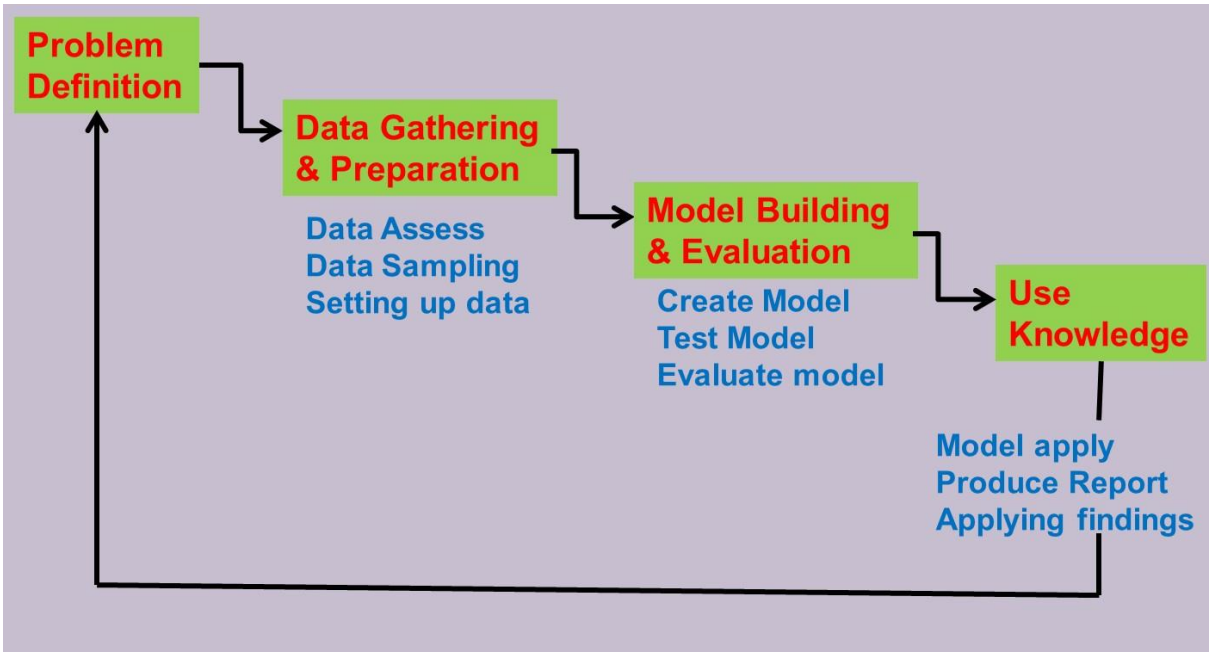
.

Task 3 – A data mining system for a Hosptial (25 marks)

A hospital has been collecting a great deal of data on their patients and have heard that use of data mining could improve their service. They would like you to create a brief report that includes the following.

(i) What data mining is and an appropriate application for the hospital.

(ii) How you would go about creating the system using the data mining lifecycle below.



(iii) If the small amount of data (diabetes.arff) collected so far by the hospital is appropriate for assessing if a person has diabetes.

(iv) The use of a data mining model such as a multilayer perceptron or decision tree to determine whether a person has diabetes. Note, you will need to use a data mining tool like WEKA to create your model and use the diabetes.arff data to train and test this model.

Deliverable:

Include a report section that addresses the four sections above and fulfils the marking criteria.

Task 4: Your Big Data Big Idea (25 marks)

This is your chance to shine!

Identify and implement an idea that you have about how you'd use Big Data for something cool. *Check the lecture 'Data in Real Life' for inspiration.*

1. Purpose an idea and clearly outlined (What is the purpose of your data collection and analysis). In the lecture notes above, you can see that each idea is specific and has a specific purpose.
2. Acquire the Data. You can do that in many ways including using available public large data sets.
3. Analyze the data in order to achieve the objective you set out for yourself in step 1.
4. Produce a report the includes your results, data visualization and thoughts.

(No, this shouldn't be a book, just a short report about what you did and how it worked out). No more than **1000 words.**

A helping hand: Sample samples are available on moodle for you to check out. Here: https://cumoodle.coventry.ac.uk/pluginfile.php/1047293/mod_resource/content/1/Sample%201.pdf