Southampton Solent University

Assessment Brief

Assessment Details

Unit Title:	Data Science
Unit Code:	SWD603
Unit Leader:	Cedric Mesnage
Level:	6
Assessment Title:	Evaluation Report
Assessment Number:	1
Assessment Type:	Software product and corresponding report
Restrictions on Time/Length :	2500 words (not including development artefacts and their explanations)
Consequence of not meeting time/word count limit:	There is no penalty for submitting below the word/count limit, but students should be aware that there is a risk they may not maximise their potential mark. Assignments should be presented appropriately in line with the restrictions stated above; if an assignment exceeds the time/word count this will be taken in account in the marks given using the assessment criteria shown.
Individual/Group:	Individual
Assessment Weighting:	100%
Issue Date:	26th September 2017
Hand In Date:	19th January 2018
Planned Feedback Date:	19th February 2018
Mode of Submission:	Online
Mode of Feedback:	Online
Number of copies to be submitted:	Not applicable: Online submission via 'MyCourse/Turnitin'
Anonymous Marking	This assessment will be exempt from anonymous marking as it falls within an exempt category under the University's Anonymous Marking Policy

Assessment Task

A detailed description of the milestones starts from page 4.

Assessment Criteria

Please refer to the detailed assessment criteria and grading scheme on page 8.

Learning Outcomes

This assessment will enable students to demonstrate in full the learning outcomes:

Knowledge and Understanding

K1 Discuss tools, technologies, methods and techniques associated with business database application development.

Cognitive Skills

C1 Analyse, design and evaluate elements of centralised, distributed and decision support database applications.

Practical and Professional Skills

P1 Undertake research of advanced database technologies and corresponding tools.

P2 Apply tools for the development of elements of centralised, distributed and decision support database applications.

Late Submissions

Students are reminded that:

i. If this assessment is submitted late i.e. within 5 working days of the submission deadline, the mark will be capped at 40% if a pass mark is achieved;

ii. If this assessment is submitted later than 5 working days after the submission deadline, the work will be regarded as a non-submission and will be awarded a zero;

iii. If this assessment is being submitted as a referred piece of work (second or third attempt) then it must be submitted by the deadline date; any Refer assessment submitted late will be regarded as a non-submission and will be awarded a zero.

http://portal.solent.ac.uk/documents/academic-services/academic-handbook/section-2/2o-assessment-policy-annex-1-assessment-regulations.pdf?t=1411116004479

Extenuating Circumstances

The University's Extenuating Circumstances procedure is in place if there are genuine circumstances that may prevent a student submitting an assessment. If students are not 'fit to study', they can either request an extension to the submission deadline of 5 working days or they can request to submit the assessment at the next opportunity (Defer). In both instances students must submit an EC application with relevant evidence. If accepted by the EC Panel there will be no academic penalty for late submission or non-submission dependent on what is requested. Students are reminded that EC covers only short term issues (20 working days) and that if they experience longer term matters that impact on learning then they must contact a Student Achievement Officer for advice.

A summary of guidance notes for students is given below:

http://portal.solent.ac.uk/support/official-documents/extenuating-circumstances/extenuating-circumstances.aspx

Academic Misconduct

Any submission must be students' own work and, where facts or ideas have been used from other sources, these sources must be appropriately referenced. The University's Academic Handbook includes the definitions of all practices that will be deemed to constitute academic misconduct. Students should check this link before submitting their work.

Procedures relating to student academic misconduct are given below:

http://portal.solent.ac.uk/support/official-documents/complaints-conduct/student-academic-misconduct.aspx

Ethics Policy
The work being carried out by students must be in compliance with the Ethics Policy. Where there is an ethical issue, as specified within the Ethics Policy, then students will need an ethics release or an ethical approval prior to the start of the project.

The Ethics Policy is contained within Section 2S of the Academic Handbook:
http://portal.solent.ac.uk/documents/academic-services/academic-handbook/section-2/2s-university-ethics-policy.pdf

Anonymous Marking
A copy of the University's Policy on Anonymous Marking, process details and student guidance on submission sheet completion can be found on the following links, which are also uploaded on the Student Portal.

Fact Sheet: http://portal.solent.ac.uk/documents/academic-services/policies-procedures-guidelines/anonymous-marking-fact-sheet.pdf

Process: http://portal.solent.ac.uk/documents/academic-services/policies-procedures-guidelines/anonymous-marking-process.pdf

Grade marking
The University uses a letter grade scale for the marking of assessments. Unless students have been specifically informed otherwise their marked assignment will be awarded a letter grade. More detailed information on grade marking and the grade scale can be found on myCourse.

Policy: http://portal.solent.ac.uk/documents/academic-services/academic-handbook/section-2/2o-assessment-policy.pdf

Fact sheet: http://portal.solent.ac.uk/documents/academic-services/academic-handbook/section-4/4o-grade-marking-briefing-for-students.pdf

1st Milestone

It is expected that at this level you have read extensively around the subject area and choices and decisions should be appropriately underpinned with suitable and current referenced material.

Case Study

You are working in an innovative music company which is planning on using the PostgreSQL Musicbrainz database as an initial point for the development of new music applications.

Question 1 (100 words written report) (10/100)

You have been assigned the task of designing a new music application(examples would be a recommendation system or a playlist sharing platform). Give a name to your application and describe briefly its purpose and how people will use it. List the tables from Musicbrainz your application will make use of.

Question 2 (200 words written report) (15/100)

Musicbrainz is a large database with a complex schema which does not enable for much development flexibility. Based on your knowledge of next generation databases discuss the advantages of the different database categories to migrate the database to.

Question 3 (SQL Queries and screenshots) (10/100)

Write the SQL queries required to get the data needed for your application from Musicbrainz. Execute your queries in the database on Alexandria and give screenshots of their executions. Remember to limit the number of results to only a few (LIMIT 100 for instance). Write a minimum of 3 queries. Submit your queries and screenshot outputs.

Question 4 (Python code and XML schema) (20/100)

The company wants to expose a Web service as a starting point for your application and asks you to produce XML documents from the database. Write a Python script which connects to the database and for each of your SQL queries from Question 3 a function which outputs an XML document representing the data fetched from the database. Include the script in your report and extracts from the XML outputs. Write an XML schema for the XML documents you are producing.

Question 5 (Python code and JSON) (15/100)

Write a similar Python script with functions which for each of your queries connects to the database and produces JSON documents. Give the script in your report and extracts from the JSON documents.

Question 6 (Python and MongoDB) (15/100)

The company has decided to use MongoDB as a database for your application and asks you to migrate the necessary data from Musicbrainz to MongoDB. Write a Python function which connects to a MongoDB database and inserts the JSON documents from the function created in Question 5 into a collection in the database. The collection should have the same name as your application. Give the code for this function.

Question 7 (Javascript Queries) (10/100)

Connect to MongoDB on Alexandria and query your MongoDB collection by writing Javascript queries which serve the purpose of your application. Include at least 3 queries in your report and show screenshots of their execution.

2nd Milestone

It is expected that at this level you have read extensively around the subject area and choices and decisions should be appropriately underpinned with suitable and current referenced material. For each question using orange produce a screenshot of your workflow in your report.

Case Study

The innovative music company wants to get some business intelligence on the data from the Musicbrainz database in order to make a decision on whether to use it for their new project.

Question 1 (200 words written report) (5/100)

Based on your knowledge of Data Science and your readings, discuss the different data mining methods that can be applied to data and what insights they can give you on the object of study.

Question 2 (SQL query) (10/100)

On Alexandria write the SQL query in Musicbrainz to get the number of artists and the number of labels by country. Produce screenshots of the query and its execution.

Here is an excerpt of the results :

Country | Artists | Labels
Kenya | 100 | 10
India | 2111 | 150

Question 3 (Python Script and CSV file) (10/100)

Write a Python script with a function which connects to the database and produces CSV data from the results of the query from question 2. Execute the function and print the output in a CSV file (“countries.csv”). Give the script in your report and extracts from the CSV file. Remember that the first line must list the column names.

Question 4 (First Scatterplot) (5/100)

Download your file “countries.csv” from Alexandria using the program “pscp.exe” which can be found on the Putty page. Start Orange, open your file and visualise the data in a data table and a scatterplot. Sort the table by number of artists. Add labels for the

countries names on the plot. Produce screenshots of the table and the plot in your report. Discuss which are the top countries.

Question 5 (Clustering) (10/100)

Use the kmeans algorithm to cluster the countries. How many clusters should we select? Discuss the optimisation of the number of clusters using various metrics (Silhouette, Distance inter-cluster and Distance to centroids). Visualise the clusters by colours in a scatterplot and in a data table for 2, 3 and 4 clusters. Discuss the clusters found. Produce screenshots of the tables and the plots in your report and to articulate your argumentation.

Question 6 (Classification Tree) (5/100)

Apply a classification tree and a classification tree viewer to the output of kmeans. Produce the screenshot of the tree and write down your interpretation of it.

Question 7 (Prediction) (10/100)

Connect a select column to choose the number of labels as target variable from your file. Sample the data and apply a linear regression to the sample connect the remaining data and the linear regression coefficients to a prediction and visualise the predicted values in a table and a scatterplot. Discuss the quality of the predictions. Additionally use a Test&Score widget for evaluation and compare with the SVM regression.

Question 8 (Frequent Itemsets) (5/100)

Explain the algorithm to compute frequent itemsets. What is the complexity of the problem of finding frequent itemsets and what is the technique used to improve its efficiency.

Question 9 (Digging Deeper) (30/100)

Using the same methodology as seen previously, query the database, produce a CSV file, download it and analyse it using Orange and its various widgets for other aspects of the Musicbrainz database. For instance discuss the distribution of the genres, classify cities, predict a label using the track length or any statistical distribution which seems relevant for you to analyse.

Question 10 (Conclusion) (5/100)

Discuss for what kind of applications you would recommend to use the Musicbrainz database based on the knowledge you gained through your analysis of the data.

Assessment Criteria and Grading

The summary grid below is the basis for grading achievement. Higher levels of achievement are described towards the right-hand side of the grid. Each level subsumes the previous level. Your grade will be the average of the two milestones with 5 marks for presentation and then converted to a grade from F to A. For each question of the milestones you will get a mark based on the marking scheme, depending on how well you answer the question you will get a higher mark. The following table describes what you achieved with your grades.

S, F1-3	D1-3	C1-3	B1-3	A1-4
Knowledge and Understanding				
Accuracy limited in breadth and depth; Range of information/evidence limited, mostly from familiar/given secondary sources, occasionally poorly organised; Largely but not wholly accurate understanding of the main concepts, theories and/or practice; Engagement with abstract/unfamiliar ideas or implications and applications is slight.	Largely accurate across most areas, with limited depth; Locates and organises an acceptable range of information/evidence often from given/familiar secondary sources; Adequate understanding of the main concepts, theories and/or practice; Limited ability to deal with abstract or unfamiliar ideas and their implications and applications	Accurate, with depth in several aspects; Locates and organises a satisfactory range of information/evidence, with limited use of primary sources; Satisfactory understanding of the relevant concepts, theories and/or practice and their main implications and applications; Understanding of more abstract aspects sometimes less developed.	Accurate and coherent in breadth, with depth in most areas; Explores and deploys information, including some aspects of new knowledge, from a wide range of secondary and several primary sources; Thorough understanding of abstract concepts, theories and/or cutting-edge practice and several of their implications and applications.	Exceptional depth in breadth; Contributes new knowledge to the subject/field of practice; Exemplary exploration and critique of information/ideas from a comprehensive range of sources (primary & secondary), many at the forefront of knowledge/practice; Advanced critical understanding of abstract concepts, theories and/or cutting-edge practice, their implications and applications, exceeds expectations for undergraduate work.
Cognitive Skills				
Superficial analysis of complex issues/problems, lacking in evaluation or synthesis; Little attempt to transfer and apply prior learning to new contexts; More descriptive than analytical and tends to rely on familiar/given material or approaches; Limited research; Sparse conclusions/practical solutions insufficiently argued/evidenced and mostly derivative, with marginally insufficient critical insight or creativity or originality.	Uses appropriate methods to analyse Complex issues/problems, with little evidence of evaluation or synthesis; Limited transfer/application of prior learning to new contexts; Tendency to description and reliance on familiar/given material or approaches; Limited range of research; Few conclusions/practical solutions sparsely argued/evidenced, mainly derivative and with little critical insight.	Uses appropriate (often given) methods to analyse complex/ unfamiliar and/or unpredictable issues/problems, with some evaluation and synthesis of information; Applies some aspects of prior learning to new contexts; Satisfactory research; Mostly relevant argument/evidence supports logical conclusions/practical solutions showing some critical insight and limited creativity or originality.	Selects and applies appropriate methods to address/solve complex, unfamiliar/unpredictable issues/problems; Largely consistent and critical judgement in analysis, evaluation and synthesis of information and application/transfer of prior learning in different contexts; Effective and wide-ranging research; Conclusions/practical solutions logically argued/evidenced, with some aspect of insight, creativity or originality.	Designs methods that convincingly address/solve complex, unfamiliar and unpredictable issues/problems; Exceptional critical judgement in analysis, evaluation, synthesis and application/transformation of prior knowledge to differing contexts; Systematic and extensive research which exceeds expectations for undergraduate work; Creative/original/compelling conclusions or practical solutions; convincingly justified/argued/evidenced.
Practical and Professional Skills				
Marginally fails to achieve basic competence in (some of) the required specialised practical, technical, creative, scholarly or work-related skills, and little awareness of professional contexts and expectations.	Basic competence in all the required specialised practical, technical, creative, scholarly or work-related skills, and partial awareness of professional contexts and expectations.	Achieves a basic level of competence in all the required specialised practical, technical, creative, scholarly or work-related skills, with more developed capability in at least one area and an awareness of professional contexts and expectations.	Competence in all the required specialised practical, technical, creative, scholarly or work-related skills, with indications of more developed ability in some areas and awareness of professional contexts and expectations.	Consistent high-level competence in all the required specialised practical, technical, creative, scholarly or work-related skills, with mastery in many areas and developed understanding of professional contexts and expectations.