COMP3008
Big Data Analytics

20 CREDIT MODULE

ASSESSMENT: 100% Coursework **W1: 30% Set Exercises**
W2: 70% Report

MODULE LEADER: Dr Lauren Ansell

MODULE AIMS

- To introduce students to the fundamentals of non-relational (NoSQL) databases.
- To critically evaluate the differences between relational and NoSQL databases.
- To gain experience pre-processing data files adequately for the use of NoSQL databases
- To gain experience with NoSQL databases through hands-on projects.

ASSESSED LEARNING OUTCOMES (ALO):

1. Critically compare and contrast the differences between relational and non-relational databases.
2. Critically appraise NoSQL database strengths and weaknesses.
3. Demonstrate the ability to perform all the CRUD operations (namely, create, retrieve, update and delete) on NoSQL databases.

Overview

This document contains all the necessary information regarding the assessment of the module COMP3008 (Big Data Analytics). The module is assessed 100% via coursework, and the final mark is composed of two elements: Set Exercises (30%) and Report (70%). You must achieve 40% overall across both elements to pass the module.

The sections that follow detail the assessment tasks that are to be undertaken. The submission and expected feedback dates are presented in Table 1 below. All assessments are to be submitted electronically via the DLE by the stated deadline.

	Submission Deadline	Feedback
Set Exercises (30%)	**Monday 17th of March 2025 at 15:00**	within 20 working days
Report (70%)	**Friday 2nd of May 2024 at 15:00**	within 20 working days

Table 1: Assessment Deadlines

All assessments will be introduced in class to provide further clarity over what is expected and how you can access support and formative feedback prior to submission. Whilst the assessment information is provided at the start of the module, it is not necessarily expected you will start this immediately – as you will often not have sufficient understanding of the topic. The module leader will provide guidance in this respect.

Available Support

Support is available during the weekly lab sessions or during my office hours every Friday 10am – 12pm either in person or online.

SET EXERCISES

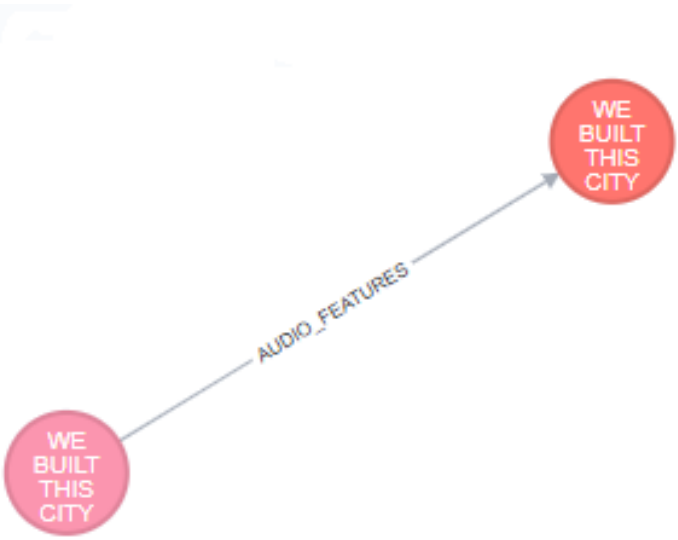
The UK chart is calculated from the sales and music and video streams. You have been provided with a dataset which contains the number-one singles from 2015 to 2023. The streaming service Spotify has an API which allows a user to download information about the audio features of tracks available on the platform.

You have been provided with 2 CSV files, one containing the information about the date and duration a track was at number one in the charts and a second file containing information on the audio features collected from Spotify, on the DLE (look for the Coursework section).

Using the data provided, you should carry out the following exercises.

Exercise 1 (15 marks)

Create a Neo4j database to store the data comprised in the CSV files. The database should respect the data model displayed in the example in the figure below. You have to provide all the commands needed to create the database and populate it with the data in the CSV files, and you must provide them in the exact order you propose to execute them. If you create indexes, you must also include the commands for index creation. Your database will be recreated, and the only way to do so is by following the commands that you will provide, in the order in which you provide them.



Exercise 2 (5 marks)

Produce a Neo4j query to list all the artists that spent more than 4 weeks at number one. The query should include the artist and be listed in descending order, that is the artist who spent the greatest amount of time at number one should be listed first, followed by the artist who is the second greatest, and so on. Your answer must show the query followed by the result with the columns appropriately named.

Exercise 3 (5 marks)

Produce a Neo4j query to identify all the tracks performed by Ed Sheeran as a solo artist and find the average valence of these tracks. You should submit the query and the result with the columns appropriately named.

Exercise 4 (5 marks)

Produce a Neo4j query to identify the most common artist in each year. The query should list each artist followed by the number of weeks in total spent at number one during that year. The artist should be listed in chronological order. You should submit the query and the result with the columns appropriately named.

The answers to the four exercises listed above should be submitted via the DLE on 17th March 2025 by 15:00. These exercises represent 30% of your final mark. Note that these exercises must be submitted in a plain text file (TXT file extension).

REPORT

You have been provided with 8 separate text files which each contain the text from several chapters from different books. The books belong to at most three different genres and your task is to group the books which belong to the same genre. The groupings will not be even and one book may belong to more than one genre, it is for you to decide which is the best fit and explain why.

To do this you need to use a range of big data analysis techniques. The report should contain commentary on the results of each of the analysis techniques you apply to the data and your final grouping of the books with justification.

Your report should include **at least** three different methods to classify documents and comment on the suitability and accuracy of each of the methods.

REPORT REQUIREMENTS

The report should fulfil the following requirements:

- You should write your report so that it is understandable and accessible, you can assume the reader is familiar with the COMP3008 material and do not wish to see it reiterated in your report.
- Your report should provide details of the methods applied for the document classification.
- Your work on this assignment should involve considerable reading and research from multiple appropriate sources—the Internet is a good place to start, but do not stop there; consider also books and preferably academic articles. Include a reference list, and make sure that your report contains citations to the sources in your reference list and using Harvard style referencing.
- When citing sources, ensure that you provide clear and explicit information about the contents of the cited source, so that the relationship between the contents of the source and the points being made are clear.

REPORT SPECIFICATION

The length of the report should be *4500 words (+/-10%)*. Include a word count immediately after your report title—a penalty of 5 marks will be applied for omitting the word count. Reports exceeding the maximum word length will be penalised. Your reference list is not included in the word count. If you produce appendices, these are not included in the word count.

Please note that a significant amount of substantive content is expected: Do not waste words on "conversational", verbose, or rambling English. Be concise!

Your report should be entirely your own work.

REPORT STRUCTURE

The following is a guide for the structure:
1. Introduction.
2. Overview of the classification methods.
3. Application of the methods.
4. Results.
5. Conclusions.
6. Reference list.

Consider the following points before starting your write-up:

You will be writing a report, *not* a journal or conference paper—hence, there is *no* need for an Abstract. You are not writing a dissertation or a book either. Thus, there is *no* need for Terms of Reference; Aims and Objectives; Methodology; Foreword.

Please avoid overly long sentences: 30 words is a long sentence and needs strong use of proper punctuation to be intelligible. Reconsider sentences once they go over 20 words.

Ensure that your writing is clear, and concise. Keep it simple, readable, and clear. Conciseness helps the reader to understand the intended message and will also help you to meet the word limit.

Ensure that your report has a clear structure in terms of numbered sections, sub-sections (if necessary) and paragraphs.

Ensure that any plots included are legible at 100% zoom.

ASSESSMENT CRITERIA

The set exercises listed in Section 1 represent 30% of your final mark. Note these exercises must be submitted via the DLE as a *plain text* file (TXT file extension).

Exercise	Weighting
Create a Neo4j database	15 marks
Exercise 2	5 marks
Exercise 3	5 marks
Exercise 4	5 marks

The report must be submitted as a *PDF* file (pdf file extension), and it should fulfil the following criteria. The report represents 70% of your final mark.

	Criteria	Weighting
1	Is there a properly constructed reference list and are the elements of such a list cited appropriately in the report?	5 marks
2	Is the research carried out appropriate in terms of breadth and depth? Is the relation between the contents of the cited sources and the contents of the report clearly explained?	10 marks
3	Does the report provide an overview of the specific methods (at least 3) used to classify the documents, and is this clearly and explicitly based upon evidence and citations from the literature?	10 marks
4	Does the report provide a detailed analysis that adequately describes the results of applying the different methods? Is the solution supported by appropriate evidence?	20 marks
5	Does the report propose a sensible grouping? Does the report identify reasons for its groupings? Is this clearly and explicitly supported by evidence? Are such examples presented and discussed in sufficient depth?	20 marks
6	Is there a set of conclusions, and do they provide a reasonable summary – at an appropriate level of abstraction – based upon the contents of the report?	5 marks

GRADE CRITERIA

When awarding marks, we shall employ the following guidelines.

Mark	Grade Criteria
Unprofessional (0-39%)	The quality of the work has not met the learning outcomes. Understanding and application of fundamental concepts and techniques is questionable. Work of this quality would not be acceptable in professional employment.
Poor (40-49%)	The quality of work has only met the threshold level but still requires further work to get it to a better standard. Your submission contains logical and analytical errors related to analysis and design techniques. Also, it only demonstrates a basic understanding of the subject competence. Further improvement is required to demonstrate personal thoroughness, effort and independent learning.
Fair (50-59%)	The quality of work submitted suggests that you have demonstrated a fair understanding of the analysis and design techniques. Still, the work you have submitted contains some errors and incomplete analysis and design. Also, it demonstrates you are able to apply your knowledge but need to improve understanding of the subject competence and personal thoroughness, effort and independent learning.
Good (60-69%)	The quality of the work submitted suggests that you are able to apply the analysis and design techniques well. The work you have submitted is substantially correct and complete. Also, it demonstrates a good understanding of subject competence and personal thoroughness, effort and independent learning.
Excellent (70-100%)	The quality of work is outstanding with no significant flaws. It demonstrates a high level of subject knowledge and competence; personal thoroughness, effort and independent learning; and possibly significant additional analytical/critical thought. Well done!

General Guidance

Extenuating Circumstances

There may be a time during this module when you experience a serious situation that has a significant impact on your ability to complete the assessments. The definition of these can be found in the University Policy on Extenuating Circumstances here.

Plagiarism

All of your work must be in your own words. You must use references for your sources, however you acquire them. Where you wish to use quotations, these must be a very minor part of your overall work.

To copy another person's work is viewed as plagiarism and is not allowed. Any issues of plagiarism and any form of academic dishonesty are treated very seriously. All your work must be your own and other sources must be identified as being theirs, not yours. The copying of another persons' work could result in a penalty being invoked.

Further information on plagiarism policy can be found here:

Plagiarism: https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations/plagiarism

Examination Offences: https://www.plymouth.ac.uk/student-life/your-studies/essential-information/exams/exam-rules-and-regulations/examination-offences

Turnitin (http://www.turnitinuk.com/) is an Internet-based 'originality checking tool' that allows documents to be compared with content on the Internet, in journals and in an archive of previously submitted works. It can help to detect unintentional or deliberate plagiarism.

It is a formative tool that makes it easy for students to review their citations and referencing as an aid to learning good academic practice. Turnitin produces an 'originality report' to help guide you. To learn more about Turnitin go to: https://guides.turnitin.com

AI Statement

As per the plagiarism section, any work submitted must be your own. The Faculty AI Policy is available for you on the DLE page to refer to. You may use AI to help structure your blog post and poster and to aid in generating ideas for the blog and poster. In the coding exercise, all code submitted must be your own.