CAB330: Assignment 2

Task Descriptions

The purpose of this assignment is to give you an understanding of how the methods you learnt from this unit can be applied to various types of datasets, such as structured record data, transactional data, unstructured text data, and web log data. This assignment includes four tasks for Clustering Analysis, Association Analysis, Text Analysis, and Web Log Analysis. You can use Python and all the libraries you've learned so far.

Task 1: Clustering Analysis

Dataset '*Music_shop_v1.csv*' is from a music shop, which stores technical details of all audio tracks that it has been producing. The dataset includes various audio track and their features. Each row represents an audio track defined by its eight attributes. Detail of the dataset is given below:

Attribute	**Data Type**	**Description**
ID	String	The ID for the track
Name	String	Name of the track
Energy	Float	Energy is a perceptual measure of intensity and activity and is recorded in the range of 0.0 to 1.0. Typically, energetic tracks (with high values of this variable) feel fast, loud, and noisy. For example, if the audio track has a high energy, then it could indicate a "death metal" while a low value on the scale could indicate "bach prelude". Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Loudness	Float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Typically tracks have this attribute values ranged between -60 and 0 dB.
Speechiness	Float	It detects the presence of spoken words in a track. The more exclusively speech-like a recording is (e.g., talk show, audio book, poetry), this attribute value is closer to 1.0. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Instrumentalness	Float	It indicates whether a track contains vocals or not. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Type	String	It denotes the object type of the track.
Time_signature	Int	It is a notational convention used in Western music to specify how many beats are contained. The range is from 0 - 5.

Assume that the music shop wants to segment the audio tracks (data points) based on the eight attributes. In order to allow efficient use of the information embedded in each segment, it is better to find a smaller number of audio track segments.

Your task is to conduct ***k*-means clustering** on this dataset and find and describe the **minimum number of effective clusters**. Answer the following seven questions in relation to this dataset and your clustering analysis.

1) Identify data quality issues in "Music_shop_v1.csv" file such as unusual data types, missing values and others. Describe your data cleaning approach and clean the dataset.
2) Which variables were included in your analysis and what were their roles and measurement level set? Justify your choice. You may need to review week 2 lecture.
3) Build a default clustering model with $k = 3$ and answer the following sub-questions:
    a) How many data points (audio tracks) are assigned into each cluster?
    b) Plot the cluster distribution using *pairplot*. Explain key characteristics of each cluster/segment.
4) What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better model? (Hint: you can use *StandardScaler* and *MinMaxScaler* from the *sklearn* package)
5) Interpret cluster analysis results using the best model by characterizing the nature of each cluster. (Hint: you may use functions *histplot, kdeplot or displot*)
6) Use Euclidean distances to compute distances between cluster centers based on the optimal *k*. Answer what are the maximum, minimum and average distances separately, and explain your method for finding the optimal *k*.
7) How the outcome of the clustering analysis can be used by decision makers? Given an application or an example where this clustering outcome can be used by the music shop.

Task 2: Association Analysis

A supermarket store is interested in determining associations between items purchased by its customers. The store has chosen to conduct an association analysis of items purchased. The dataset, "*POS_TRANS_v1.csv*". Detail of the dataset is given below:

Attribute	**Description**
LOCATION	Point of sale device identification number
TRANSACTION_ID	Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
TRANSACTION_DATE	Date of transaction
PRODUCT_NAME	Product Purchased
QUANTITY	Quantity of this product purchased (always set to 1 by a point-of-sale device)

Your task is to conduct association analysis on this data set and answer the following seven questions in relation to this dataset and your association analysis.

1) Identify data quality issues in "*POS_TRANS_v1.csv*" file such as unusual data types, missing values and others. Describe your data cleaning approach and clean the dataset.
2) What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3) Conduct association mining and answer the followings:
    a) What is the highest *lift* value for the resulting rules? Which *rule* has this value?
    b) What is the highest *confidence* value for the resulting rules? Which rule has this value?
4) Plot the *confidence*, *lift*, and *support* of the resulting rules. Interpret them to discuss the rule set obtained.
5) The store is particularly interested in products that individuals purchase when they buy "Tea".
    a) How many rules are in the subset?
    b) Based on the rules, what are the other products these individuals are most likely to purchase?
6) If the store wants to increase the sales of "Shampoo", what products should be placed near it based on the association rules?
7) How the outcome of this analysis can be used by decision makers? Given an application or an example where this outcome can be used by the supermarket store.

Task 3: Text Analysis

A cinema is planning to provide an online recommendation service to its customers. The cinema has collected a metadata dataset, "*M_metadata_v1.csv*", which contains some information on movies. This task requires you to perform text analysis on this dataset to determine clusters of movies based on similar topics that can be obtained from the movie descriptions. Detail of the data set is given below:

Attribute	**Description**
Cast1, Cast2, Cast3, Cast4, Cast5 and Cast6	The group of popular actors/actresses who acted in the movie
Description	This provides a short synopsis of the movie
Director 1, Director 2, Director 3	The list of directors for this movie. If it is directed by only one director then Director 2 and Director 3 will have "Director Not available".
Genre	A movie genre is a motion picture category based on similarities in either the narrative elements or the emotional response to the film. It has values like Documentary, Kids&Family, Romance and SciFi. Hint. It can be used to name the derived clusters.
Rating	Using the Motion Picture Association of America (MPAA) film rating system, each movie is rated for its suitability for certain audiences based on its content. It includes G, NC17, NR, PG, PG-13 and R
Release Date	The date of release for the movie.
Runtime	Runtime is the time between the starting of the movie up to the end of the credits scene.
Studio	The facility that was used to make that movie.
Title	Title of the movie
Writer1, Writer2, Writer 3, Writer 4	A list of screenplay writers or the scriptwriters or scenarists who has written the screenplay for this movie.

Year	The Year the movie was released

Answer the following seven questions in relation to this dataset and your text analysis.

1) What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
2) Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.
3) Did you disregard any frequent terms? Justify your answer.
4) Justify the term weighting option selected, and what is the number of input features available to execute *k*-means clustering? (Please note that how the original text data is converted into a feature set to support text analysis)
5) Find the optimal *k*. Plot and explain your answer.
6) How many clusters are generated? Provide a way to meaningfully name each cluster and then display the results.
7) How the outcome of your analysis can be used by the decision maker? Given an application or an example where this outcome can be used by the cinema.

Task 4: Web Log Analysis

The dataset you will use for this task is a web log file, "*Weblog_v1.csv*" which provided by an e-commerce company. This task requires to determine user browsing patterns of the website and analyze those patterns to provide recommendations to improve the website. The dataset is the original text file that needs to be processed with the steps required for web usage mining as explained in the practical. Detail of the data set is given below:

Attribute	**Description**
IP address	Client's IP address
Timestamp	The time, in coordinated universal time (UTC), at which the activity occurred.
Request	Represents the data of the HTTP requests that are recorded in the Web log file.
Status	200 (Successful request) 206,302,304,404(Unsuccessful requests)

Your task is to pre-process the given dataset and apply a suitable data mining method, such as classification, clustering, or association mining, to the web log dataset. Answer the following four questions in relation to this data and the analyses that you have chosen.

1) Pre-process the log data to identify useful attributes based on columns in the text file such as IP_address, Timestamp, Request, or Status.
2) What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3) Apply a data mining task (method) on the processed dataset. Explain the rationale behind selecting the data mining task/method.
4) Discuss the results obtained, and the applicability of your findings. You should include only a high-level managerial kind of discussion on the findings. It should not just be an interpretation of results as shown in results.

Marking Guidelines

Your marks are determined based on your group work (project report, 85%) and your demo (15%) in the week 13 Practical session.

In the Week 13 Practical session, you should be prepared to show your final diagrams and results to your marker. The marker will ask each individual student questions and will assign individual marks. The marker will also check the code, plots and results, along with the report. The entire group should be present to show the project result and answer the questions raised by marker. We expect you to complete a draft project report prior to the demonstration.

Please note that in data mining or data science, there is hardly ever a single solution. Also, many times, there is no correct or wrong solution. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked.

The marks are distributed as follows:

Task 1 (7 marks)

Task 2 (7 marks)

Task 3 (7 marks)

Task 4 (4 marks)

Instructions

1. You should submit the project report via **Canvas (Assignment 2 Submission Link).**

2. Name the project report as **asm2_groupNumber.doc.** This word file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group number (name). Combine this file with your **team agreement** and your **source code** and name the compressed file as **asm2_[groupNumber].zip.**

2. The project report should be divided into four parts according to each task, each part starting from a new page. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to all questions for each task.

3. This is a group assignment. The group size is three. You can continue the same group as in Case Study 1. If you have formed a new group after Assignment 1, please notify the lecturing staff. They will remove you from the existing group. In this case, you need to register your new team at Canvas.

4. The group is to be ARRANGED and MANAGED by you based on the operating norms in your team agreement document. As in real life, the performance of the individuals in the team

shall be judged by the performance of the team together. Please contact the teaching staff with any team conflict issues.

5. Of course, the work your group hand in must be your own; no collaboration or borrowing from other groups is permitted. Read the Assessment Policies on Canvas or QUT Website.

Marking Criteria

Criteria	**Comments and Scoring**
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components.	0-5 (Very Poor)
Has demonstrated a task with a working model having a data source, and diagram with substantial but incorrect implementation of at least one of the components. Questions were poorly answered.	6-10 (Poor)
Has implemented all tasks with at least two being substantially correct. Shows some understanding of concepts with some success in applying knowledge. Only basic questions were answered.	11-12 (Unsatisfactory)
Has implemented all four tasks: One task is fundamentally correct, with substantially correct workflow diagrams which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts.	13-15 (Acceptable level of Achievement)
Has fundamentally correct implementation of all tasks i.e., selection of correct variables in data, correct allocations, understanding, and explanation of clusters, findings association rules, finding clusters in text data with good term features, and application of an appropriate data mining operation to the log data. Shows competency in applying data mining. Many questions have been reasonably answered. Demonstrate a good understanding of the methods and terms used in clustering, association mining, text mining and web mining, during written and verbal analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	16-18 (High level of achievement)
Has implemented all of the requirements above with very few errors. A strong focus on application of tools, and evaluation and interpretation of results is evident.	19-21 (Very high level of achievement)
All of the criteria above are met, extensive model generation and analyses have been conducted to produce exceptional outcomes. Have applied principles learnt in lectures to enhance the results.	22-25 (Exceptional level of achievement)