# CAB330: Case Study 2

**Due date: 20th Oct, 2019**

**Weighting: 25%**

## Introduction

The purpose of this assignment is to give you an understanding that data mining methods can be applied to various types of data sets such as record data, transactional data, text data and web logs. This assignment is divided into four parts: Clustering, Association mining, Text Mining and Web Mining. You will create Python programs for Clustering, Association mining, Text Mining and Web Mining tasks.

## Task 1: Descriptive Data Mining - Clustering

A music company has been storing technical details of all audio tracks that it has been producing. The ***SPOTIFY*** data set includes various audio track and their features. Each row represents an audio track defined by its eight attributes. Detail of the data set is given below.

| Name | Data Type | Description |
|---|---|---|
| ID | String | The Spotify 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 | Speechiness 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 | 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 | Denotes the object type of the track. |
| Time_Signature | Int | The time signature is a notational convention used in Western music to specify how many beats are contained. The range is from 0 - 5. |

The audio company would like to segment the audio tracks based on the aforementioned features. They will prefer the lower number of audio track segments, to allow to use the information embedded within each segment effectively.

Your task is to conduct k-means clustering on the ***SPOTIFY*** data set, then find and describe the **minimum number of effective clusters**. Answer the followings in relation to this data and analysis.

## Task 1.1. Data Preparation for Clustering

1. Can you identify data quality issues in this dataset such as unusual data types, missing values and others?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

## Task 1.2. The first clustering model

1. Build a default clustering model with K= 3 and answer the followings:
    a. How many records are assigned into each cluster?
    b. Plot the cluster distribution using pairplot. Explain key characteristics of each cluster/ segment.
2. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
3. Interpret the (best out of the above two models) cluster analysis outcome. In other words, characterize the nature of each cluster by giving it a descriptive label by using distplot.

## Task 1.3. Refining the clustering model

1. Using elbow method and silhouette, find the optimal K. What is the best K? Explain your reasoning.
2. What is the best number of clusters that can describe the dataset effectively? Was this obtained with the default setting (i.e. K= 3) or refining the clustering model as above (Task 1.3.1)?
3. How the outcome of this study can be used by decision makers? Given an application where this clustrreing outcome can be used by the music company.

## Task 2: Descriptive Data Mining - Association

A supermarket store is interested in determining the associations between items purchased by its customers. The store has chosen to conduct a market basket analysis of items purchased.

The ***POS_TRANSACTIONS*** data set includes over 200,000 transactions made over the past three months. The following products are represented in the data set:

[Yoghurt, Jam, Shampoo, Bread, Egg, Milk, Tea, Cordial, Peanut butter, Dishwashing liquid, Cereal, Coffee, Conditioner, Butter, Sugar, Jelly, Cheese]

Detail of the data set is given below.

| Name | Description |
|---|---|
| LOCATION | Point of sale device identification number (e.g. for Register 3) |
| 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 the ***POS_TRANSACTIONS*** data set. Answer the followings in relation to this data and analysis.

1. Can you identify data quality issues in this dataset for performing association analysis?
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 following:
    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?
    c. Plot the confidence, lift, and support of the resulting rules. Interpret them to discuss the rule-set obtained.
4. The store is particularly interested in products that individuals purchase when they buy “Bread”.
    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?
5. How the outcome of this study can be used by decision makers?

## Task 3: Text Mining

A leading movie review and aggregation website is planning to start an online personalised movie recommendation service. They have a collection of individual movie descriptions. Perform text mining on this ***MOVIE*** 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.

| **Name** | **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 upto the end of the credits scene. |
| Studio | The facility that was used to make that particular 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 followings in relation to this data and 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?
4. Justify the term weighting option selected.
5. What is the number of input features available to execute clustering? (FYI: Note how the original text data is converted into a feature set that can be mined for knowledge discovery.)

6. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters which could be based on the movie genre?
7. Identify the first six high frequent terms (that are not stop words) in the start list?
8. Describe how these clusters can be useful in the planned online personalised movie recommendation service.

## Task 4: Web Mining

For an e-commerce business, the website structure and site plan were established with the efficiency and usability in mind, but its effectiveness was not verified. Only basic statistics have been produced through simple report and query techniques, but they provide no means for sophisticated web site analysis and predictions. Your task is to determine the user browsing patterns of the website and analyse those patterns to provide recommendations to improve the website.

You have been provided with a raw log file, ***WEBLOG***. This 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.

| **Name** | **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)<br>206,302,304,404(Unsuccessful requests) |

Your task is to **pre-process the given dataset and apply a suitable data mining operation**, such as classification or clustering or association mining, to the raw WEBLOG data set. Answer the followings in relation to this data and the analyses that you have chosen.

1. Pre-process the ***WEBLOG*** data to remove any unproductive items from the log file such as graphics, sound, etc and also identify UserID, SessionID and STEPS (within a given session, the order of the pages visited) based on IP address, date and time.
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Apply a datamining task on the processed dataset. Explain the rationale behind selecting the data mining task/method.
4. Discuss the results obtained. Discuss also the applicability of findings of the method. 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.

# Distribution of Marks (Total: 25 marks)

We would mark your data mining projects in the Week 13 practical class. You should be prepared to show your final work to your markerf. The marker will ask each individual student questions and will assign individual mark (~15%).

In data mining, 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. Submit the report discussing the final project components.

The marks are distributed as follows.

**Clustering analysis and pre-processing (6 marks)**
**Association Mining and pre-processing (5 marks)**
**Text Mining and pre-processing (6 marks)**
**Web mining (5 marks)**
**Report Writing (1 mark)**

# Instructions

1. The assignment is **due on 11:59 pm, 20th Oct**. It is a firm deadline.

2. You should submit the assignment report via Blackboard Assignment.

3. The assignment (Clustering, Association, text and web mining project DMProj2) will be **marked in the practical class in Week 13**. We will check the final code and results, along with the assignment report, to assign you marks. The entire team should be present to show the project result and answer the questions raised by the marker. We will ask questions to each student, and will assign about 15% of total marks as per individual performance.

4. The datasets required for this assignment can be found on BlackBoard with the file named as **casestudy2-data.zip**. It includes four datasets:

    a. ***SPOTIFY*** to perform clustering
    b. ***POS_TRANSACTIONS*** to perform association mining
    c. ***MOVIE*** to perform text mining
    d. ***WEBLOG*** to perform web mining

5. Name the case-study report as **casestudy2.docx.** The 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 name. Combine this file with your **team contract**, and name the compressed file as **casestudy2.zip.** Submit the zip file on **Blackboard (under assessment panel Assignment 2.**

6 . The **project report** should be divided into five 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 the questions set in this case-study specification. Some answers may require screen shots. Answer the questions in the case study for each model appropriately and succinctly. If a case-study step is about conducting a process, you do not have to provide an explanation or a screen shot. However if a question such as "Examine the results of clustering/association mining" is asked, you then need to explain what, why or why not? While you may like to go into extreme detail about, you may to do so but limit the project report to 20-30 pages. So, write down the important points and attach the important screen dumps, to show that you have thought the matter through.

7 . This is a group assignment. The team 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 Blackboard.

8. The group is to be ARRANGED and MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.

9. 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 Blackboard or QUT Website.

## Assignment Criteria Sheet

| Criteria | Scores |
|---|---|
| Non Submission of any of the data mining models/ evidence of | 0 |
| Has demonstrated a task with a working data mining model with /without submission and demonstrates the ability to run the program and add some components. | 1-5 |
| Has demonstrated a task with a working data mining model having a data source, and with substantial codes but incorrect output. Questions were poorly answered. | 6-11 |
| 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. | 12 |
| Has implemented all the tasks:<br>One mining task is fundamentally correct, with substantially correct work flow which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts. | 13-15 |
| Has fundamentally correct implementation of all tasks i.e.<br>Selection of correct variables in data, correct allocations, understanding, and explanation of clusters, finding association rules, finding clusters in text data with good term features, and application of an appropriate data mining operation to the web log data. Shows competency in applying text 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 |
| 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 |
| 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 |