**Nanyang Technological University  
Nanyang Business School**

**BC2402 – Designing and Developing Databases**Semester 1, 2024/25

**Individual Assignment**

**Hello… I am an AI. How can I assist you?**

**1. INTRODUCTION**

*A. The Preamble*

SINGAPORE: Singapore has proposed a new framework for generative artificial intelligence (AI) and is now seeking views from the international community on it.

The Model AI Governance Framework for Generative AI expands on the existing framework covering traditional AI and which was last updated in 2020.

"With generative AI, there is a need to update the earlier model governance framework to holistically address new issues that have emerged," said the AI Verify Foundation (AIVF) and Infocomm Media Development Authority (IMDA) in a joint press release on Tuesday (Jan 16).

AIVF is a network that was launched last year, tapping the likes of IBM, Microsoft and Google to develop AI testing tools for responsible use and shape international standards.

The proposal seeks to set forth a "systematic and balanced approach" for generative AI concerns while facilitating innovation, and involves all key stakeholders like policymakers and the research community to "collectively do their part".

Feedback from the international community will be collated and used to support the finalisation of the Model AI Governance Framework for Generative AI in mid-2024, said the agencies.

Tham, A. (2024) “Singapore proposes generative AI framework”, Channel News Asia, retrieved from <https://www.channelnewsasia.com/singapore/generative-ai-artificial-intelligence-proposal-framework-4051526>

*B. AI is too chatty, creating bogus refund policies.*

Air Canada is being held responsible for a discount its chatbot mistakenly promised a customer, the Washington Post reported.

The airline must refund a passenger, Jake Moffat, who two years ago purchased tickets to attend his grandmother's funeral, under the belief that if he paid full price, he could later file a claim under the airline's bereavement policy to receive a discount, according to a ruling by Canada's Civil Resolution Tribunal (CRT).

He didn't invent the idea, rather a support chatbot with which he communicated on Air Canada's website provided him the false information, ultimately costing the airline several hundred dollars. The tribunal's judgment could set a precedent for holding businesses accountable when relying on interactive technology tools, including generative artificial intelligence, to take on customer service roles.

Cerullo, M. (2024) “Air Canada chatbot costs airline discount it wrongly offered customer”, CBS News, retrieved from https://www.cbsnews.com/news/aircanada-chatbot-discount-customer/

**2. DATA DESCRIPTIONS**

You are provided with a dataset extracted from a local car rental company. The dataset consists of multiple tables. Specifically,

1. classTBL

*This table contains data about flight classes.*

[ClassID] uniquely identifies each record

[Class] is the class name (i.e., econ, econ premium, business)

1. customerTBL

*This table contains data specific to each customer.*

[KFlyerID] uniquely identifies each record

[CustName] corresponds to the key in surnameTBL

[CustGen] gender

[CustAge] age

[MemeberSince\_m] the month in which a customer’s membership commences

[MemeberSince\_y] the year in which a customer’s membership commences

[PostalSect] corresponds to the key in postalsectTBL

[MembershipType] is the type (level) of a customer’s membership

1. destTBL

*This table contains data specific to various destinations.*

[DestID] uniquely identifies each record

[AirCode] is the airport code

[LocName] is the name of the destination

[Dist] is the distance (in miles) between Singapore and the destination

1. fulllogTBL  
   *This table contains data specific to a customer-chatbot interaction.*

[ChatID] uniquely identifies each record

[UserID] corresponds to the key in customerTBL

[ChatSource] whether the chat content is sent by a human or chatbot

[Content] is the chat message

[Date\_d] the day on which the chat message is sent

[Date\_m] the month in which the chat message is sent

[Date\_y] the year in which the chat message is sent

[Time\_hh] the hour at which the chat message is sent

[Joy] the probability that the chat message illustrates joy emotions

[Anger] the likelihood that the chat message illustrates anger emotions

[Disgust] the likelihood that the chat message illustrates disgust emotions

[Surprise] the likelihood that the chat message illustrates surprise emotions

[Fear] the likelihood that the chat message illustrates fear emotions

[Sadness] the likelihood that the chat message illustrates sadness emotions

[Contempt] the likelihood that the chat message illustrates contempt emotions

[Sentimentality] the likelihood that the chat message illustrates sentimentality

[Confusion] the likelihood that the chat message illustrates confusion

See the Appendix for details on how these emotions can be computed.

1. postalsectTBL

*This table contains data about postal code locations*

[PostalSect] uniquely identifies each record

[GeneralLoc] the general location represented by the postal section

1. redeemTBL

*This table contains data about mileage redemptions.*

[RedeemID] uniquely identifies each record

[Redeem] whether redemption has been performed

1. seatTBL

*This table contains data about cabin seats.*

[SeatID] uniquely identifies each record

[Seat] the specific seat

1. surnameTBL

*This table contains data about customers’ surnames.*

[SurnameID] uniquely identifies each record

[Surname] the specific surname

1. tripsTBL

*This table contains each trip detail.*

[KFlyerID] corresponds to the key in customerTBL

[TripID] uniquely identifies each record

[RouteID] corresponds to the key in destTBL

[Outbound] 1 = departing from Singapore; 0 = returning to Singapore

[Trip\_m] the month in which the trip occured

[Trip\_y] the year in which the trip occured

[SeatRow] the specific row of cabin seat

[SeatSpc] corresponds to the key in seatTBL

[SMealRq] whether special meal request is made

[Class] corresponds to the key in classTBL

[Redeem] corresponds to the key in redeemTBL

[EliteMilesMod] is the elite miles modifier (e.g., the actual distance = 1000 miles, EliteMilesMod = 2.00, Elite Miles = 1000 x 2 = 2000)

[Kshop] is the type of Kshop purchased

You are encouraged to use the SQL database implementations provided with this document to manage your workload.

Note that you must submit your database implementation if you use the datasets to implement your databases. Otherwise (if you shall be using the provided databases), you are not required to submit the database implementations.

**3. PROJECT DELIVERABLES**

***The due date for the assignment is 4 October 2024 (23:59 hrs NTULearn server time)***

You are not expected to modify the provided mySQL database implementation. You are only expected to submit one file, as follows:

1. 1 x SQL script file

**A. SQL script file**

You are tasked to develop some SQL scripts to query the data, as follows:

You are to create SQL scripts (e.g., SELECT statements) that generate answers for the following queries.

1. Table considered: <customerTBL*>*

How many customers are there?

1. Table considered: < customerTBL >

What are the membership types?

*Important: Note the data issues. You decide what to do about them. Using comments, explain your decision-making.*

1. Table considered: < customerTBL >

For each membership type, how many customers are there?

1. Tables considered: <customerTBL> + < postalsectTBL>

For each GeneralLoc, on each membership type, display the total number of customers, and the respective breakdowns between females and males

1. Tables considered: < customerTBL> + < postalsectTBL>

For each general location, on each membership type, display the number of customers who have been members since 2000.

1. Tables considered: < customerTBL> + < TripTBL> + <destTBL>

For each membership type, display the total trip distance (among members with the same membership type) for trips completed between the year 2020 and 2022.

1. Tables considered: <TripTBL> + <destTBL>

Who are the top travelers during holiday seasons (i.e., July, August, November, and December) and non-holiday seasons (i.e., remaining months)?  
Using a single result grid, display the top two travelers (total longest trip distance) for the holiday and non-holiday seasons, respectively.

You should only consider outbound flights.

You should not consider trips going to NRT, MAN, and LGW.

The output should contain the following columns.

**

The first two rows of results will be the top 2 travelers during the holiday season. The subsequent two rows will be the top 2 travelers during the non-holiday season.

*Note: Pay attention to the data. There are hidden issues.*

1. Tables considered: <fullogTBL> + < customerTBL >

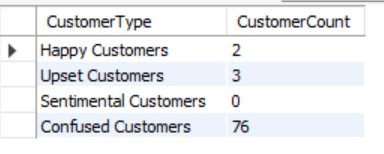
Which membership type is the most frequent chatbot user?

For each membership type, display the number of sustained conversations (i.e., a sustained conversation involves more than 1 customer-chatbot exchange in a conversation instance).

*Important: Note the data issues in fullogtbl. You decide what to do about them. Using comments, explain your decision-making.*

1. Tables considered: <tripsTBL> + <destTBL> + <fulllogTBL> + <customerTBL>

Customer Analytics. You may present the final output in the following format.

Your results may differ. Do provide your assumptions (if any) in the inline comments.  
 

Suggested Logic  
i. Generate a list of userid, and the corresponding averages of joy, anger, disgust, surprise, fear, sadness, contempt, sentimentality, confusion.

ii. Generate a list of KFlyerID and the corresponding modified miles (Dist \* EliteMilesMod).

iii. Combine (1) and (2) based on a logical condition.

iv. Generate a list of KFlyerID and the corresponding Positive Emotions (averages of Joy, generated above), Negative Emotions (averages of Anger + Disgust + Fear + Sadness), Sentimentality (generated above), and Confusion (generated above).

v. Generate a list of KFlyerID and the corresponding ratios. Specifically,

Positive Emotions Ratio = log (Positive Emotions) / log (modified miles)

Negative Emotions Ratio = log (Negative Emotions) / log (modified miles)

Sentimentality Ratio = log (Sentimentality) / log (modified miles)

Confusion Ratio = log (confusion) / log (modified miles)

vi. Retrieve records satisfying the following conditions:

Positive Emotions Ratio > Negative Emotions Ratio AND  
Positive Emotions Ratio > Sentimentality Ratio AND

Positive Emotions Ratio > Confusion Ratio

The number of records is the number of “Happy Customers”.

Repeat (vi) for each of the following conditions.  
**[Upset Customers]**Negative Emotions Ratio > Positive Emotions Ratio AND  
Negative Emotions Ratio > Sentimentality Ratio AND

Negative Emotions Ratio > Confusion Ratio

**[Sentimental Customers]**Sentimentality Ratio > Positive Emotions Ratio AND  
Sentimentality Ratio > Negative Emotions Ratio AND

Sentimentality Ratio > Confusion Ratio

**[Confused Customers]**Confusion Ratio > Positive Emotions Ratio AND  
Confusion Ratio > Negative Emotions Ratio AND

Confusion Ratio > Sentimentality Ratio

vii. Perform unions to merge on the above results vertically.

*Important: Note the data issues in fullogTBL. You decide what to do about them. Using comments, explain your decision-making.*

1. Tables considered: <FulllogTBL>

MySQL SOUNDEX() function returns soundex string of a string.

SOUNDEX is a phonetic algorithm for indexing names after the English pronunciation of sound. For details, see <https://dev.mysql.com/doc/refman/8.4/en/string-functions.html#function_soundex>

Audio Analytics. Using SOUNDEX(), identify the most frequent soundex string (4 rightmost characters) from the content column (i.e., 1423).

Display records with the content column containing the soundex string (i.e., %1423%).

*Important: Note the data issues in fullogTBL. You decide what to do about them. Using comments, explain your decision-making.*

**4. SUBMISSION**

A submission folder will be made available on NTULearn. You can make as many submissions as necessary, but only the latest submission will be evaluated.

The submission must be made by **4 October 2024, 23:59**.

**5. APPENDIX**

Sentiment analysis is contextual mining of text, which identifies and extracts subjective information in source material and helps a business to understand the social sentiment of their brand, product, or service while monitoring online conversations. However, analysis of social media streams is usually restricted to fundamental sentiment analysis and count-based metrics. This is akin to scratching the surface and missing out on those high-value insights waiting to be discovered.

With recent advances in deep learning, algorithms' ability to analyze text has improved considerably. Using advanced artificial intelligence techniques can be an effective tool for conducting in-depth research.

In the dataset, we focus on seven core emotions: Joy, Anger, Fear, Surprise, Sadness, Contempt, and Disgust, as well as complex emotions: Sentimentality and Confusion.

Sentimentality is defined as "happy sadness” and is often seen when watching emotionally resonant material. Contrary to other Emotion scores, Sentimentality has several possible combinations of core emotions that can contribute to a positive score, commonly seen together.

Confusion was created in an effort to parse different contexts for the use of features suggestive of Anger and negative valence in general but can also suggest mental effort or confusion.

Sentiment analysis uses natural language processing (NLP) and machine learning (ML) technologies to train computer software to analyze and interpret text in a way similar to humans. The software uses one of two approaches, rule-based or ML—or a combination of the two known as hybrid. Each approach has its strengths and weaknesses; while a rule-based approach can deliver results in near real-time, ML based approaches are more adaptable and can typically handle more complex scenarios.

1. Rule-based sentiment analysis

In the rule-based approach, software is trained to classify certain keywords in a block of text based on groups of words, or lexicons, that describe the author’s intent. For example, words in a positive lexicon might include “affordable,” “fast” and “well-made,” while words in a negative lexicon might feature “expensive,” “slow” and “poorly made”. The software then scans the classifier for the words in either the positive or negative lexicon and tallies up a total sentiment score based on the volume of words used and the sentiment score of each category.

1. Machine learning sentiment analysis

With a machine learning (ML) approach, an algorithm is used to train software to gauge sentiment in a block of text using words that appear in the text as well as the order in which they appear. Developers use sentiment analysis algorithms to teach software how to identify emotion in text similarly to the way humans do. ML models continue to “learn” from the data they are fed, hence the name “machine learning”. Here are a few of the most commonly used classification algorithms:

Linear regression: A statistics algorithm that describes a value (Y) based on a set of features (X).

Naive Bayes: An algorithm that uses Bayes’ theorem to categorize words in a block of text.

Support vector machines: A fast and efficient classification algorithm used to solve two-group classification problems.

Deep learning (DL): Also known as an artificial neural network, deep learning is an advanced machine learning technique that links together multiple algorithms to mimic human brain function.

1. The hybrid approach

A hybrid approach to text analysis combines both ML and rule-based capabilities to optimize accuracy and speed. While highly accurate, this approach requires more resources, such as time and technical capacity, than the other two.

References

https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

https://www.ibm.com/topics/sentiment-analysis

Friesen, W. V., & Ekman, P. (1984). EMFACS­7.

D'Mello, S.K., Craig, S.D., Graesser, A.C. (2009). Multimethod assessment of affective experience and expression during deep learning. International Journal of Learning Technology