MAKERERE UNIVERSITY

SCHOOL OF COMPUTING & IT

MCN 7105: STRUCTURE AND INTERPRETATION OF COMPUTER PROGRAMS END OF SEMESTER PROJECT ACADEMIC YEAR 2024/2025

Student Name: GODWIN NOMUGISHA

Degree Program: MDCSE

Reg. No: 2024/HD05/22077U

Question 1

Examining the implementation of the specified data science abstractions and describing their different levels of data and procedure abstraction layers.

1.Linear Model: A linear model is a tool used to find the relationship between two variables. For example, given a reader reviews a book, one may estimate how positive or negative the overall sentiment of the review will be depending on the frequency of certain words such as 'trust' or 'anger' in the book.

Data Abstraction: The model takes in a set of predictors i.e the number of times certain words occur and a target target value such as a sentiment score and undergoes training and preprocessing where the data needs to be cleaned.

Procedure Abstraction:

- Fit: This procedure examines the training data in order to determine the position of the line of best fit.
- Predict: After the model has been trained, it can be used to predict results on other datasets. For instance, one can begin to analyze a totally different book.
- Evaluate: After the predictions have been made, it is necessary to verify the efficacy of the model through other means such as testing with R^2 or measuring error rates.

2. list-sentiment: This piece of work is meant to analyze a given text in order to determine its overall tone by means of the words it contains. For example, in Reality of Our Situation, scores of individual words such as "trust" and "fear" are settled and the program estimates the overall mood with those scores.

Data Abstraction:

- Lexicon: A lexicon is a type of dictionary that assigns positive or negative ratings to words.
- Input List: This is a listing of words that were taken out of the particular text under analysis (e.g. trust, fear, positive).

Procedure Abstraction:

- Mapping: Each word in the input is matched to its score in the lexicon.

- Aggregation: Scores are summed-up or averaged in order to yield one outcome, the sentiment score of the entire analysis.

For example, in the first dataset, say, “positive” recorded 80 times and “negative” 70 times. Averaging the two gives an overview of the tone of the document.

3. read-csv: This abstraction is for extracting the data out of textual files with a CSV delimiter. If someone collects the sentiment scores for the stories, they could put it in a CSV and use this abstraction to read and analyze these scores.

Data Abstraction is where we examine both Raw CSV data and Structured Data

- Raw CSV data: It is a straight file that consists of list of rows and columns of figures only.
- Structured Data: Once processed, the information is structured into tables that can be utilized better.

Procedure Abstraction

- Syntax: This phase divides the source document into horizontal and vertical segments.
- Heading Normalization: Prepares the contents of each column in a coherent manner for example Sentiment and Frequency columns.
- Formating: Explains the process of converting raw text into required forms, which might be numbers, dates, etc.

4. qq-plot*: In statistics, a Q-Q plot is often a graphical tool for comparing a data set to a specific distribution, or checking if the data set follows a certain type of distribution. For example, one might be interested to know whether the sentiment scores of Reality of Our Situation are normally distributed.

Data Abstraction

- Data Source: This contains the word frequencies or the sentiment scores contained text.
- Statistical Distribution: Mathematical description of a variable, normal curve for this case, used to for comparison purposes.

Procedure Abstraction

- Quantile Calculation: Obtains data values indicating the proportion of the dataset that they represent.
- Plotting: The comparison of quantiles with the same probability in two datasets; one being the sample data and the other being the theoretical data. If the plot is a straight line the two data sets are proportional.

5. Hist: Histograms are a very effective way to tell the distribution of data. In this case, they can express the how many times affective labels, such as positive, negative, trust and others have been used in a story.

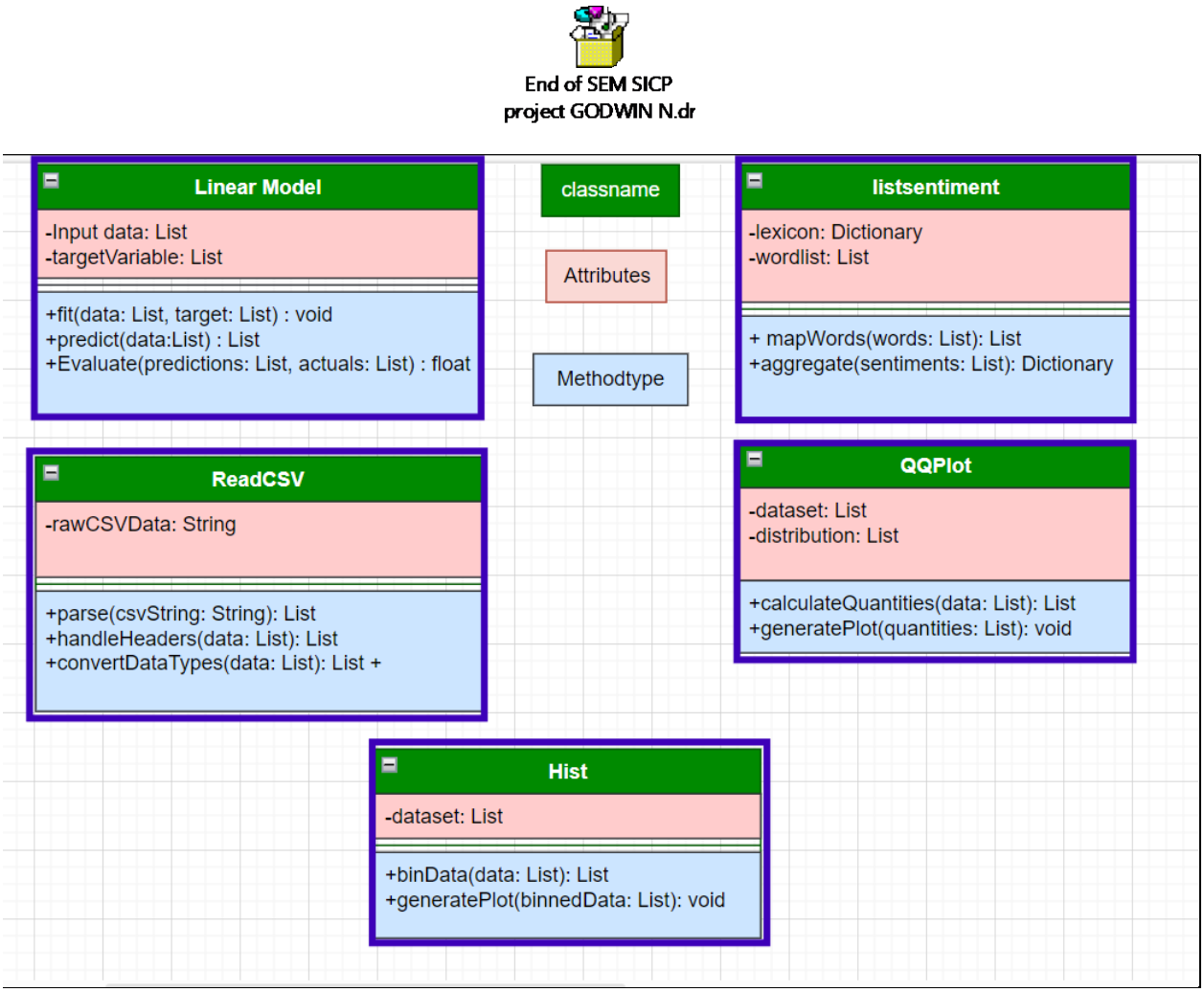
Data Abstraction:

- Dataset: the affective labels eg “trust”, “anger” and the number of times they are found in a text.
- Bins: Makes the data more easily understood by splitting it into groups and or ranges.

Procedure Abstraction:

- Binning: This involves dividing the data into intervals such as 10–20 and 20–30 and so on.
- Plotting: During this stage, a histogram is constructed using bars where the length of the bar used to represent a value is proportional to the frequency represented.
- For instance, in Reality of Our Situation, one will see “positive” being used more than joy, disgust, etc; its frequency is over 80.

Diagram of Abstraction Levels generated using draw.io



Question2

We are to build a set of abstractions for analyzing the moods of tweets from Uganda over a period of 12 months.

1. **tweet-sentiment.** This quote addresses how sentiment is formed on a page with a single tweet with a #hashtag, emojis and other forms of language that are quite 'informal' and are found on tweets. It is an enhancement of the earlier list-sentiment abstraction, however unused features of this abstraction include the features of tweets such as hashtags (#UgandaDecides) and emojis.

Motivation: It is not uncommon to see tweets use contractions, neologisms, unique hashtags or an emoji at the end of a tweet as those features capture sentiment. Otherwise, without those integrating features, the meaning of the tweets would be lost i.e;

Data Abstraction: The object of this abstraction is the main text of the tweet posted with its hashtags, mentions, and emojis and each of these forms of expression is part of the main text of the tweet. From the text, sentiment lexicon assigns scores to every word contained in it, hashtag or an emoji.

Procedure Abstraction: Preprocessing steps involve deletion of links and conversion of text to lowercase. Words, hashtags, and emojis are remapped to their corresponding new sentiment scores. These computed scores when combined yield the overall sentiment score for that particular tweet.

2. **location-tweets.** As with the parameter, this abstraction seeks to restrict the analysis of tweets only to those that have been geotagged to Uganda. This incorporates the mentioned geotags on the metadata as gives by the Twitter API.

Motivation: To use sentiments that accurately reflect the opinion of Ugandans, it is necessary to eliminate the analysis of sentiments from other countries regions;

Data Abstraction: Considering the metadata, the geotag is one of the most important, as it indicates where a tweet is originating from along with the times of that particular tweet. Furthermore, there is a filtering criterion that takes out tweets that do not originate from Uganda.

Procedure Abstraction: So here, tweets are filtered out on the basis of the country from which they are generated, and hence the country under study will also be Uganda. The tweets are then subdivided according to the regions, for instance Kampala, Gulu, Fort Portal, etc. to provide regional analysis.

3. **time-trend.** This abstraction analyzes sentiments of tweets on a monthly basis, to show variations in moods and how they change over time and are impacted by events such as the elections.

Motivation: Changes in mood tend to have meanings or other descriptions such mouth movements when a gif is incompletely captured during editing or the entire gif fails to capture the action; all of which describe depth, for example, people actually feel more positive during national holidays and during political unrest, many feel sad.

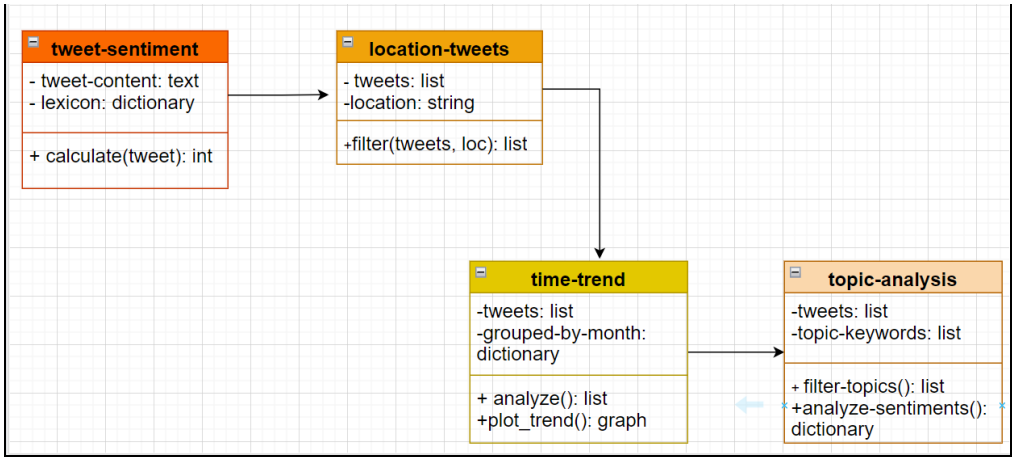
Data Abstraction: Employs the timestamps of each tweet and uses the generated sentiment scores (tweet-sentiment).

4. **topic-analysis.** This abstraction organized tweets into different topics, such as politics, entertainment, or the economy. Consequently, this would enable a more detailed sentiment analysis of particular issues.

Motivation: So many Tweets are sent related to different subjects on daily basis that they can be combined into groups to get a better understanding of how people feel about different aspects of life.

Data Abstraction: Has a list of pre-defined topics (ex. Politics, Economy, Sports, etc.) and an accompanying list of keywords and hashtags for each topic.

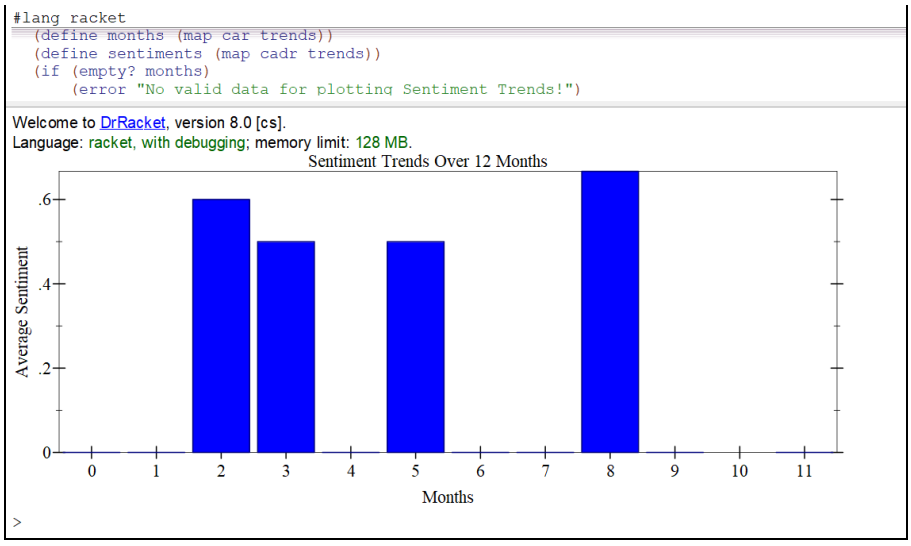
Procedure Abstraction: Automatically groups tweets into topics based on keywords or hashtags then applies tweet-sentiment for each of the topics to derive net sentiment scores.



Code output images sample data run using the built source code,

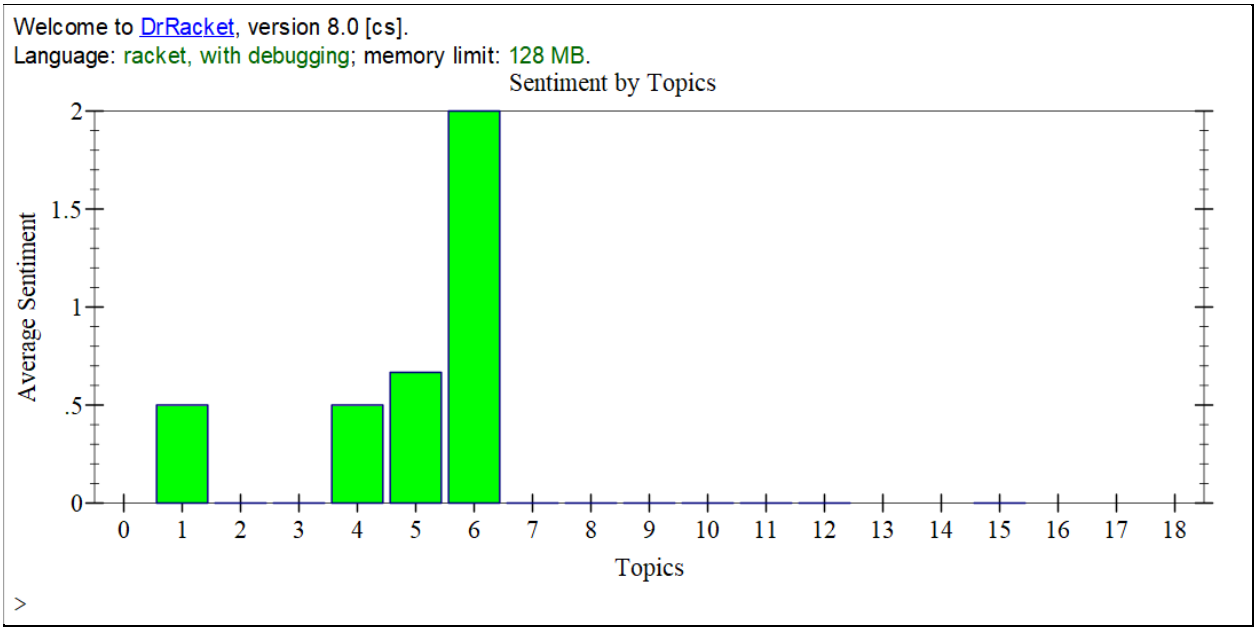
1. **Graph 1: Sentiment Trends Over Months**

✓ **X-Axis**: Months (e.g., January, February, etc.)
✓ **Y-Axis**: Average sentiment scores calculated from tweets for each month



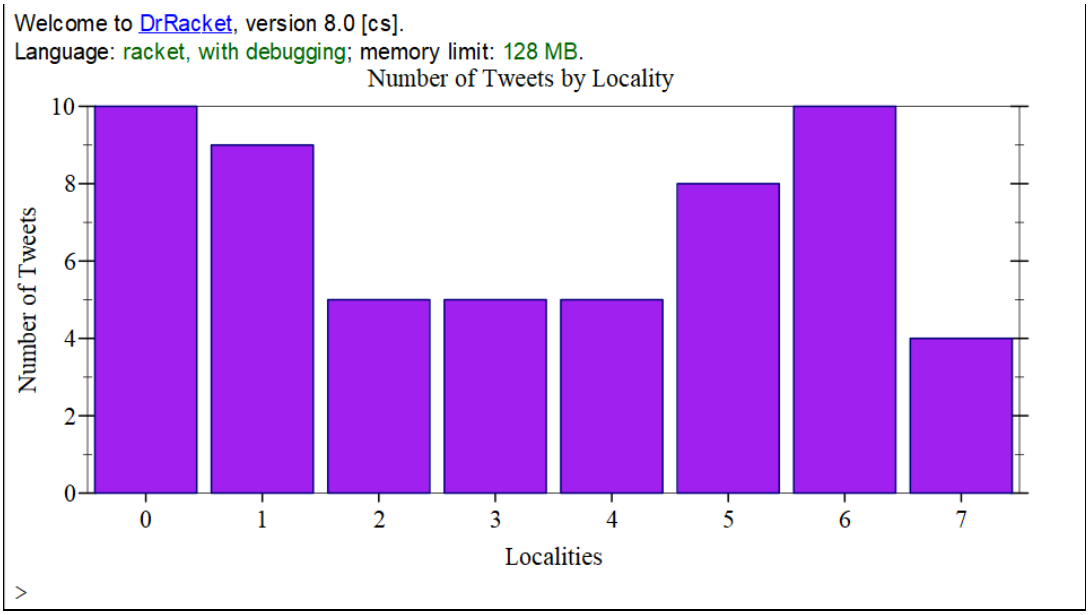
Graph 2: Sentiment by Topics

- ✓ **X-Axis**: Topics (e.g., Politics, Entertainment, Economy, etc.)
- ✓ **Y-Axis**: Average sentiment scores calculated from tweets for each topic



Graph 3: Number of Tweets by Locality

- ✓ **X-Axis**: Localities (e.g., Kampala, Gulu, Entebbe, etc.)
- ✓ **Y-Axis**: Count of tweets from each locality



GitHub repository: https://github.com/Gysh-Wyn/uganda-tweets-sentiment-analysis