# **ASSIGNMENT BY GROUP C**

# **PART A**

TWEET DATA ANALYSIS

# Introduction

Over the past decade, online data has become an important asset for any conglomerates in the world. Exploring this data into information benefitted these conglomerates in numerous ways like news channels can report local news, food industries can run local coupons and advertisers can run promotions based on the public interest. The major platform that confers this online information is social media especially Facebook and Twitter.

Out of many social media platforms recently twitter count alone escalated to near 330 million active users every month in 2019 (Luceri et al., 2019) from which we can draw conclusions on the amount of data being transferred online every day. It is estimated that more than 500 million users are posting their opinions each day (Rogovschi and Grozavu, 2014) which is limited to 140 characters pre tweet. This made twitter a popular public platform for sharing opinions. All this huge information flowing in twitter can be mined using knowledge mining techniques based on different categories and can be used in different fields. Thanks to twitter for introducing entities and tags where users can highlight the key words or the locations which made data mining more proficient. For instance, the tweets with the key words reflecting the symptoms of any kind of disease in a particular area can help public health organisations to get the command over the situation and perform medical operations (PaulM. and Dredze., 2012). In the similar way if tweets were found furnishing opinions on any key issues, those can be addressed by the representing bodies (Rogovschi and Grozavu, 2014). Yet, despite having geo tags on twitter few tweets can’t be located. These tweets can be identified using data mining techniques. Therefore, twitter serves as one of the social network platforms which can benefit the world in becoming more convenient.

In this paper, we aim to predict the missing locations for the data set collected and find the spatial distribution of the tweeters. Also, analyse the social issues or important days by associating the key words with them.

# Related work

As mentioned earlier, twitter can be mined for various purposes. Yet, all the attributes from the datasets are not suitable for the specific data mining. Each data mining technique requires a specific data pre-processing for the model to build. Particularly speaking about the prediction of location from the personal tweets, it is observed that using content of the tweet and social relationship of the tweeter together (Yuan et al., 2016) or content of the tweet and information on geographical label (Shen et al., 2018) is ideal for best results in most cases. Also, from the research done by (Luceri et al., 2019) it is found that using 10% data of the tweeters we can predict at least 50% of users’ location. Yet, even after the selection of attributes it is challenging to predict the location based on the complexity of the tweet content. In the work done by (Rogovschi and Grozavu, 2014) and (Steiger et al., 2016) it is clearly mentioned that to reduce the complexity, clustering of content using self-organising maps (SOM) is most entertained. Moreover, a study by (Steiger et al., 2016) shows that there is a high relevance in using temporal (Time), spatial (Location) and semantic (content) characteristics of the tweet where SOM can be used for clustering time, coordinates and the content at the same time. These clustered data in turn can be used for predicting the location by using neural networks which is found to be efficient by (Shen et al., 2018) and (Yuan et al., 2016). Another data knowledge creating we are going to present in this paper is association rule mining. This technique is more used for text mining from the social media platforms. This algorithm can be used for sentiment analysis and trend analysis for the twitter data and can relate the important words or phrases to the most trending topics around (KABIR et al., 2018) and also it can be applied for non-language features (J and G, 2018). This sentiment analysis is also used for stock market predicting by dividing the content into positive, negative and neutral reactions based on the words used in the data (Bing, Chan and Ou, 2014).

# Application 1 (SOM and Neural Network)

## Using SOM for clustering and NN for location prediction and spatial distribution

Before developing and training model, data pre-processing is required for the model to perform according to the requirement.

## Pre-processing

We used Rapid Miner for this application.

Step 1. After importing the metadata into Rapid miner, it includes 20298 rows and six set of attributes: “ID”, “User Id”, “Time”, “User Home Town”, “Location” and “Content”.

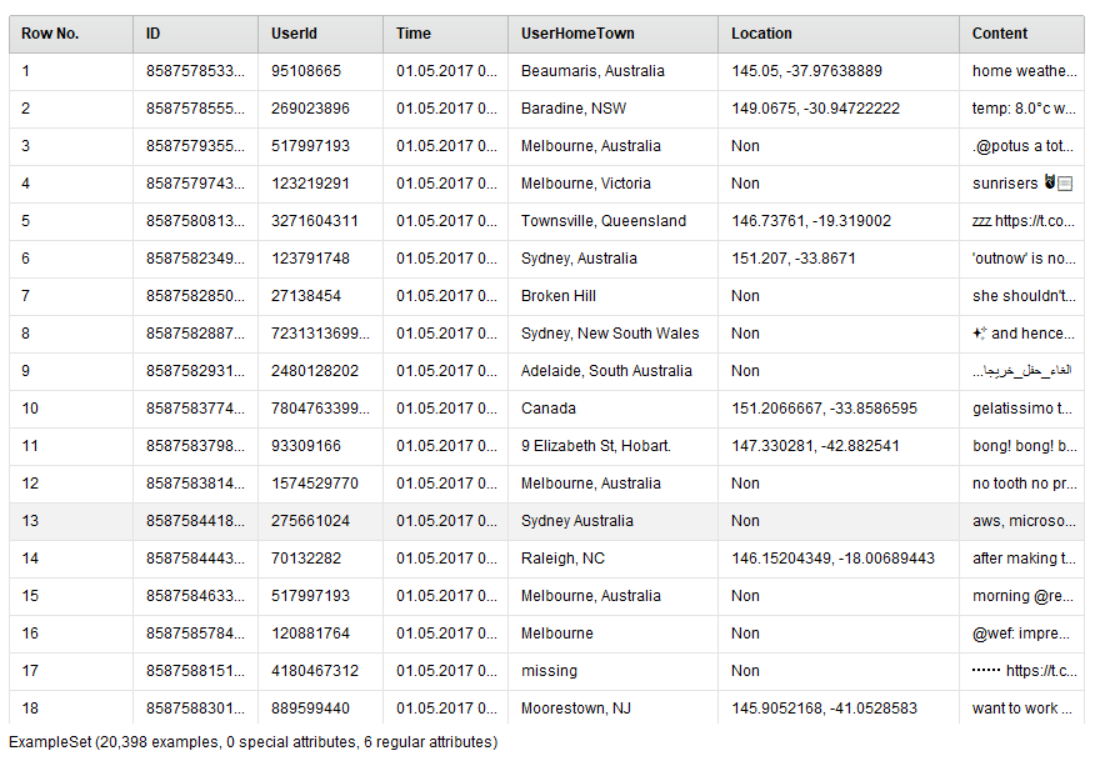


Figure Meta data

Step 2. Based on the related work we have collected Time, location and content are found to be more relevant to each other. Hence, we considered only those attributes. While addressing some data found difficult on the rapid miner, we used Microsoft Excel to address it and only 6000 rows were selected out of limitations.

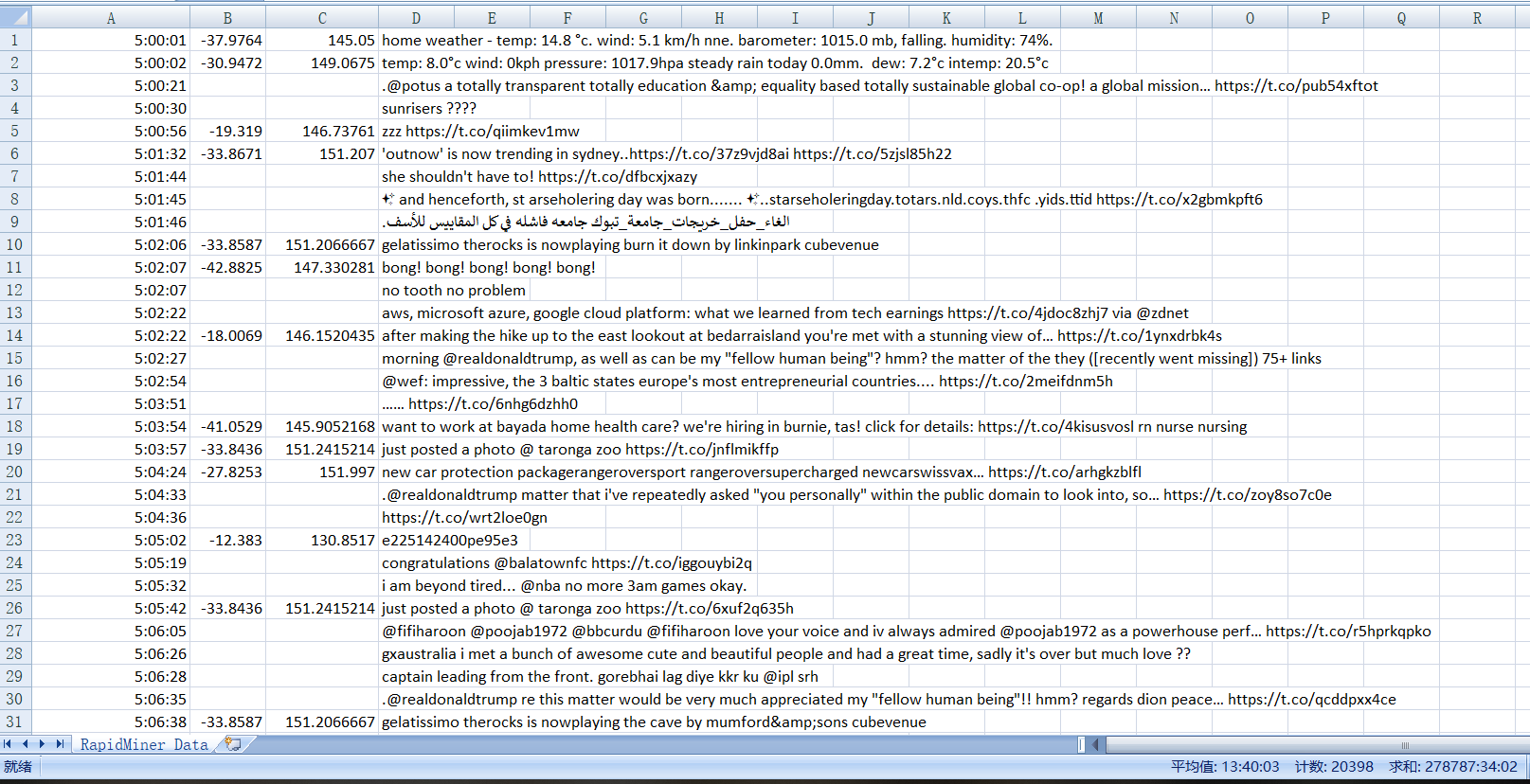


Figure pre-processed data

Step 3. Most of the data can’t be used straight away for SOM and neural networks. Pre-processing is required which include the steps like “Generate ID”, “Set Role”, transform type of attributes, processing for text data, “rename”, “Replace Missing Values”, “Select Attributes”, “Filter Examples”, “Join”, “Split”, “Split Data” in rapid miner.

Application of SOM: After the data is set for clustering, as we intend to find location from latitudes and longitudes, we apply SOM to cluster the values of locations. It is clustered into 30 neurons in this paper.

Table1. The coordinates of 30 areas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **geospatial characteristic** | **average Latitude** | **average Longitude** | **geospatial characteristic** | **average Latitude** | **average Longitude** |
| **0** | -37.828 | 144.958 | **15** | -15.206 | 148.219 |
| **1** | -38.097 | 144.470 | **16** | -23.926 | 151.545 |
| **2** | -36.800 | 144.314 | **17** | -27.179 | 153.111 |
| **3** | -36.930 | 143.477 | **18** | -28.244 | 153.443 |
| **4** | -44.834 | 136.730 | **19** | -31.059 | 152.676 |
| **5** | -39.187 | 133.034 | **20** | -33.295 | 151.466 |
| **6** | -34.885 | 138.712 | **21** | -33.872 | 151.197 |
| **7** | -33.059 | 137.514 | **22** | -33.981 | 150.888 |
| **8** | -27.292 | 134.415 | **23** | -33.603 | 150.335 |
| **9** | -32.664 | 118.331 | **24** | -32.107 | 148.923 |
| **10** | -32.131 | 115.799 | **25** | -35.260 | 148.913 |
| **11** | -21.109 | 116.590 | **26** | -36.781 | 148.854 |
| **12** | -13.217 | 117.205 | **27** | -43.663 | 148.066 |
| **13** | -6.440 | 121.054 | **28** | -39.633 | 146.331 |
| **14** | -10.597 | 133.973 | **29** | -37.786 | 145.154 |
|  |  |  |  |  |  |

Step 4. The content of the tweet has to be clustered using SOM considering its complexity. It has divided the content into 50 topics in this research.

Application of neural nets: Now, for predicting the missing values of spatial characteristic Neural Network is to be applied and build the final model. All the data which are not “?” (missing value) and “Non” should be filtered to train the model and test the performance. The rapid miner file process will be in the following way.

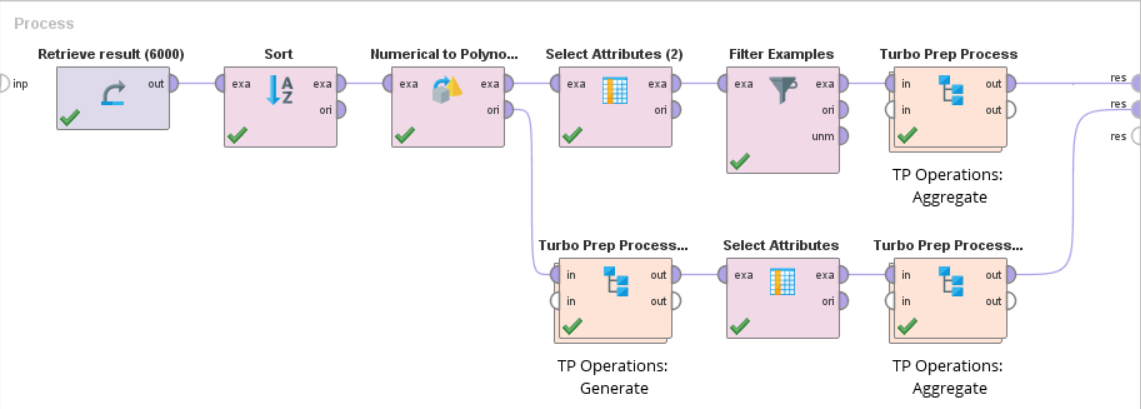
****

Figure final process in Rapid Miner

# Results

After the application of the neural nets, the missing values of the locations (spatial characteristics) were found. It has predicted the missing values by computing their distribution density. It is done by calculating the average values of the longitudes and the latitudes and spreading all over the area. As a result, 30 areas were found and tabulated as …

Table 2. tweet density

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **geospatial characteristic** | **Tweet Number** | **geospatial characteristic** | **Tweet Number** | **geospatial characteristic** | **Tweet Number** |
| 0 | 124 | 10 | 67 | 20 | 67 |
| 1 | 21 | 11 | 5 | 21 | 253 |
| 2 | 25 | 12 | 4 | 22 | 52 |
| 3 | 5 | 13 | 51 | 23 | 10 |
| 4 | 5 | 14 | 8 | 24 | 17 |
| 5 | 5 | 15 | 39 | 25 | 37 |
| 6 | 54 | 16 | 12 | 26 | 9 |
| 7 | 1 | 17 | 3031 | 27 | 24 |
| 8 | 16 | 18 | 1966 | 28 | 3 |
| 9 | 2 | 19 | 13 | 29 | 74 |



We mapped these coordinates on the google earth and found the following results.

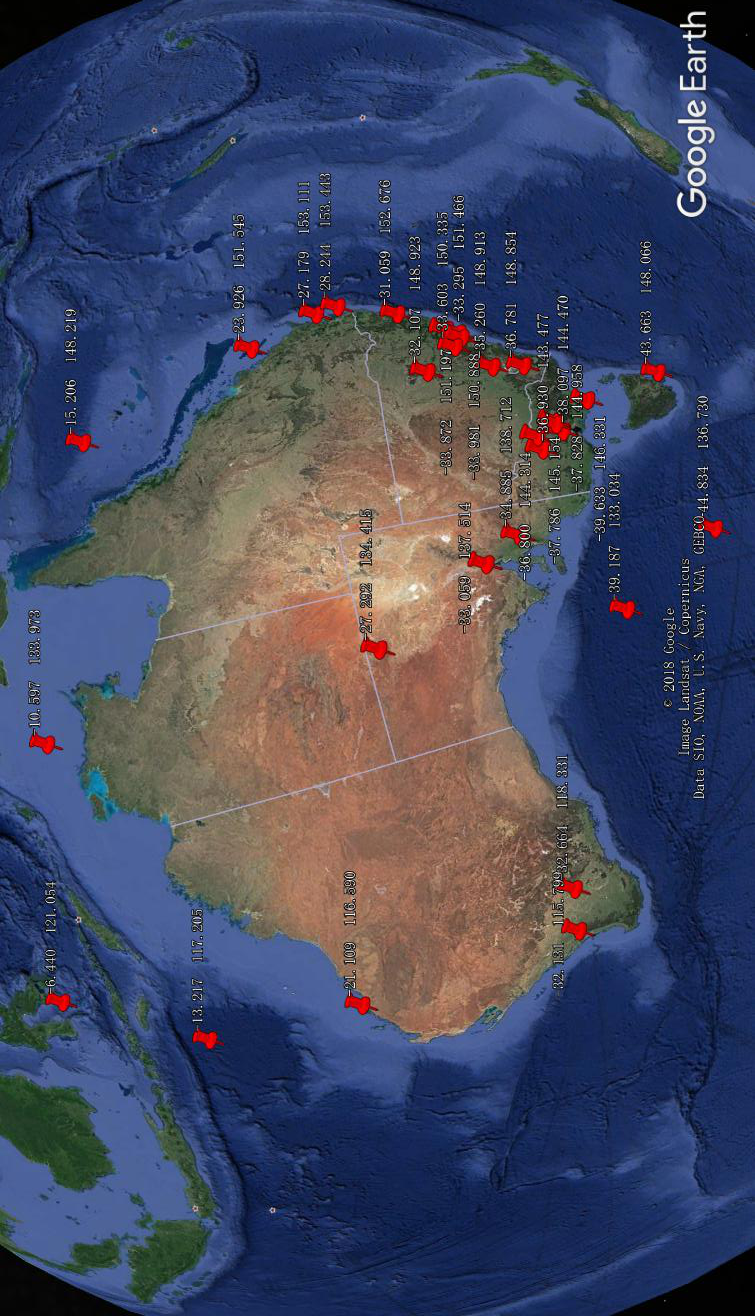


Figure Regional distribution of locations

We also mapped tweet density which shows the number of tweets from a particular area.

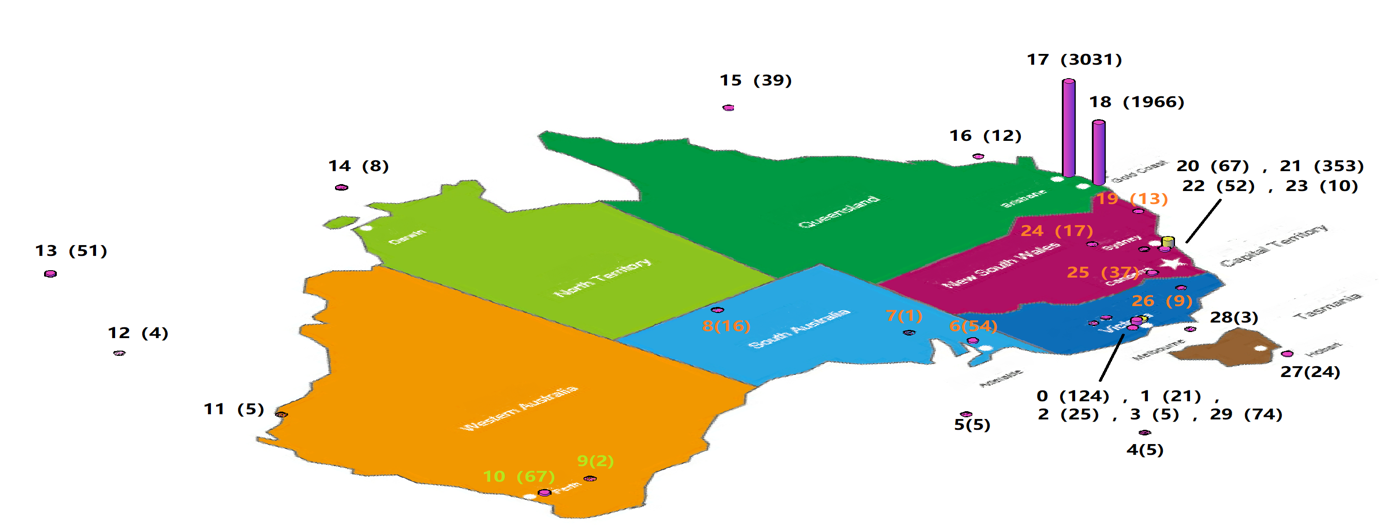


Figure tweet density

After training the Neural Net by 80% complete data, the generated model can be test by the last 20% data and shows the performance by “Root Mean Squared Error” and “Squared Error”.

Table3. Performance Vector

|  |  |
| --- | --- |
| Root Mean Squared Error | 8.683 +/- 0.000 |
| Squared Error | 75.391 +/- 98.258 |

# Application 2 (Association rule)

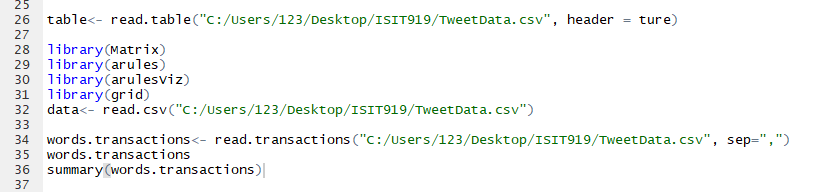
Association mining rule uses apriori algorithm to analyse the content of the tweet data. We can draw conclusions on the key objectives addressed in the tweets by mapping the date or social issues with the most used words.

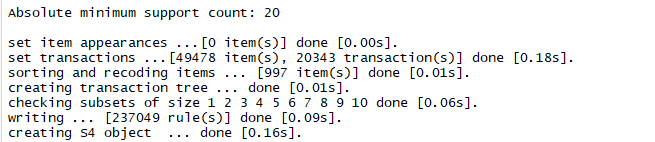
The basic pre-processing done with our data is eliminating the few unusual attributes and mapping the date with the content against the location. This helps us to find how people in a particular react to the key issues on a particular date.

## Pre-processing

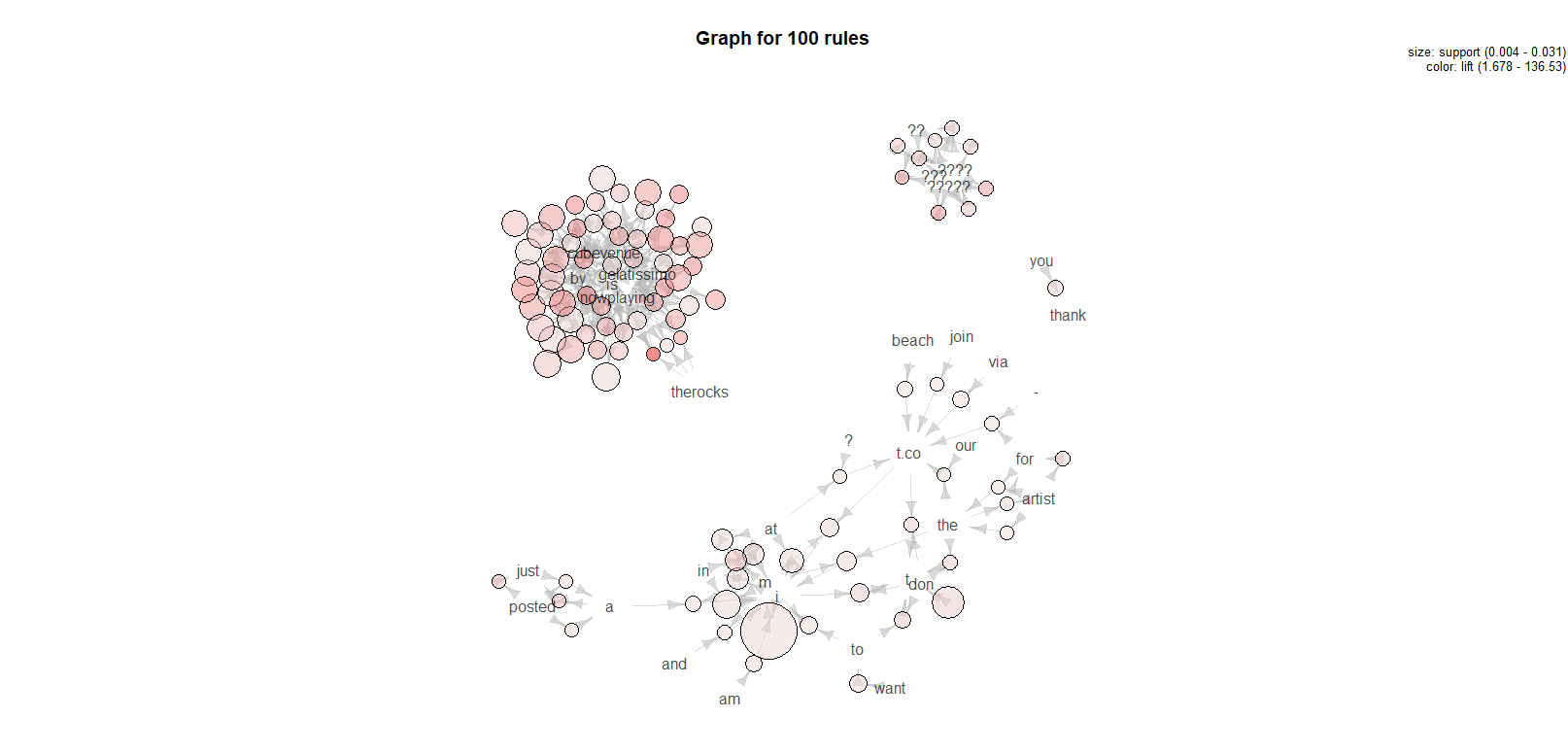
We used R-studios for our convenience for this application

Import data, apply association rule to mine the content of the tweet data.





Apriori algorithm is introduced to mine the association rules from dataset. The parameter support is set to 0.001 and confidence is set to 0.7 to obtain the association mining as below.



Association rule can draw knowledge based on the date in the dataset. We have mapped this date with the calendar and found it to be a Labor Day.

Date [01.05.2017] is Labor Day. In Queensland and northern territory, it is considered as a public holiday.

## Our observations for the rule mining are as follows:

Left-hand Rule and right-hand rule occur frequently at the same time: {labour}=>{day}



1. People discussed the essential issues in the society about transparent education which is a new concept for traditional educational field.

Rule (lhs- hrs): {transparent} => {education}



1. From the high lift value, it was noticed that people from Sydney in NSW are more mouthing on the words like advice, hazard and reduction. This rule indicates that people from this location are more concerned on social safety issues about Hazard reduction. This result can contribute to study social phenomena through social media on safety domain.



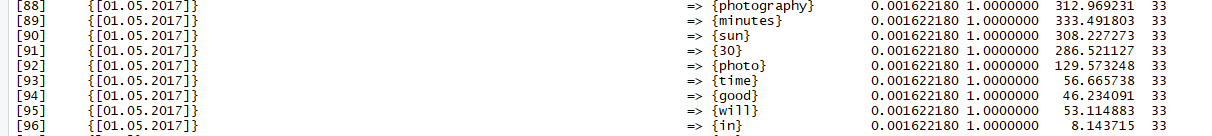
1. Twitter data can be applied to mine knowledge about news in terms of the art and its original location. This rule below shows that Country music artist jeffcrew is well known or came from Melbourne, AU. This association rule can indicate that social media can mine knowledge for social phenomena on art domain.



1. This rule below is relevant to word kawana that occur with gelatissimo, nowplaying, and cubevenue. This public knowledge can provide the other twitter about what is popular themes in gelatissimo kawana and the feeling of joining this kawana.



1. This rule below can be regarded as an itemset {01.05.2017}=> {photpgraphy}, {sun} { 30} {minutes} {good} { time} {photo} , which shows that people from different location like Brisbane, Melbourne, Darwin, Hobart prefer posting photos about sunrise when it’s about to rise in 30 minutes on tweeter. This indicate that most people enjoy sunrise early in the morning and people are happy to post photos to share on the tweeter social media to remember their special day on the May 1, 2017.



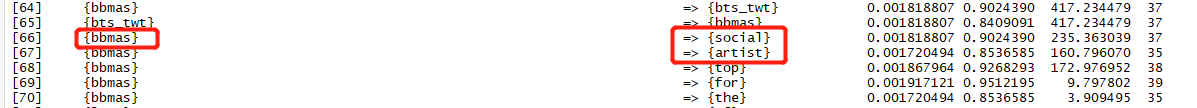
1. Rule {Posted} => { photo} show that people always prefer to post photos on tweeter to record their activities on this day. This social phenomenon can be beneficial to data scientists to analyze the public behaviors and people’s lifestyle in accordance to the development of new technologies. So, mining the public knowledge can help in studying social behavior.



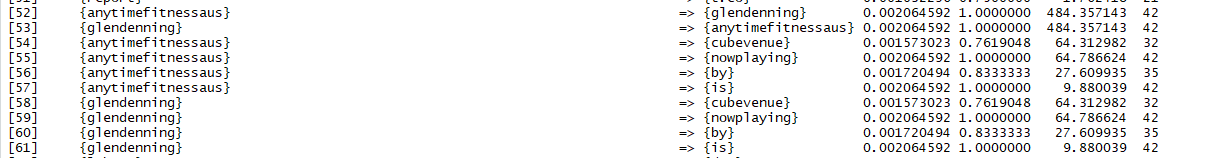
1. Rule {barometer} => {wind} from dataset shows that barometer always occur with wind which can describe the weather from a specific location and the rest can have a proper plan for daily activities according to weather and so on. This indicates the importance of weather forecasting, even this data from the twitter can help people in regulating their daily activities.



1. Rule {bbmas} => {social artist} always appear together which can provide relevant knowledge about social artist in this domain.



1. Rule { } => {nowplaying} {cubevenue} occur together which implies the opening hours of the gym anytimefitness and the importance of health.



# Discussion

As per the observation from figure 4, it is clear that the major part of the metadata we collected are the tweets from Australia and found to be from urban areas. This confirms that the clustering from SOM and prediction from neural networks gave the accurate results. This data has also been cross verified with objective laws. However, 5 areas are found in middle of the sea which is not possible. The logical explanation for this can be explained as there is a large dispersion degree in the data coordinates, the average values we created did spread from the middle point to the farthest distance of real area. In addition, some of the geospatial characteristics are very close yet SOM clustered them separately since the coordinates are too accurate. In further research, the highly dispersed data can be avoided for the better results and the accuracy of the coordinates should be to suitable degree such as keeping it to one to four decimal places. The other data mining technique is the association rule where we tried to map the words and the date present in the data. This gave us an insight on how people react on any special event also the concerns they have.

# Conclusion

According to the Figure 9, these tweets in the metadata are almost sent in Brisbane and Gold Coast (No.17 and No.18 area, about 3031 and 1966 respectively). Sydney (No 21 area, about 253) and Melbourne are the second largest areas (No 0, 1, 2, 3, 28 and 29 areas, sum up 272). It shows that social media are more active in travel areas than living city. Even this rule mining helped in putting the weather details in place where people can get benefitted besides from the weather forecast. The knowledge mined will lay a platform for the study on different aspects in the world which includes the social behaviour.

# Reference

LUCERI, L., ANDREOLETTI, D. & GIORDANO, S. 2019. Infringement of Tweets Geo-Location Privacy: an approach based on Graph Convolutional Neural Networks. *arXiv preprint arXiv:1903.11206*.

ROGOVSCHI, N. & GROZAVU, N. Opinion retrieval through unsupervised topological learning. 2014 International Joint Conference on Neural Networks (IJCNN), 2014. IEEE, 3130-3134.

SHEN, W., LIU, Y. & WANG, J. Predicting Named Entity Location Using Twitter. 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018. IEEE, 161-172.

STEIGER, E., RESCH, B. & ZIPF, A. 2016. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science,* 30**,** 1694-1716.

YUAN, G., MURUKANNAIAH, P. K. & SINGH, M. P. Percimo: A personalized community model for location estimation in social media. Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2016. IEEE Press, 271-278.

KABIR, A., KARIM, R., NEWAZ, S. and HOSSAIN, M. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. *Informatica Economica*, 22(1/2018), pp.25-38.

Bing, L., Chan, K. and Ou, C. (2014). Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. *2014 IEEE 11th International Conference on e-Business Engineering*.

Paul, M.J. and Dredze, M., 2012. A model for mining public health topics from Twitter. *Health*, *11*(16-16), p.1.

J, A. and G, J. (2018). Sentiment Classification of Tweets with Non-Language Features. *Procedia Computer Science*, 143, pp.426-433.

# **PART B**

In this section we are handling a huge dataset of around 30000 rows. This data represents the record of 70 diabetic patients. We have combined these 70 records into one dataset where the attributes represent the date, time, code and value.

Based on the code values given, we can understand the corresponding value it has.

33 = Regular insulin dose

34 = NPH insulin dose

35 = UltraLente insulin dose

48 = Unspecified blood glucose measurement

57 = Unspecified blood glucose measurement

58 = Pre-breakfast blood glucose measurement

59 = Post-breakfast blood glucose measurement

60 = Pre-lunch blood glucose measurement

61 = Post-lunch blood glucose measurement

62 = Pre-supper blood glucose measurement

63 = Post-supper blood glucose measurement

64 = Pre-snack blood glucose measurement

65 = Hypoglycemic symptoms

66 = Typical meal ingestion

67 = More-than-usual meal ingestion

68 = Less-than-usual meal ingestion

69 = Typical exercise activity

70 = More-than-usual exercise activity

71 = Less-than-usual exercise activity

72 = Unspecified special event

Our aim is to find the pattern among those codes based on the other attributes. As a first step we have gathered the data of three months comparing the glucose measurement and the insulin value. We divided the result into sections of the day like morning (5 am to 12pm), afternoon (12:00 to 16:00) and evening (16:00 to 23:59).

Observed data for January is as follows:

Glucose levels in the morning, afternoon and evening

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Insulin levels in the morning, afternoon and evening

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

It is observed that in the morning glucose levels are less with high insulin values when compared with the other sections in the day. It is observed the same for February and march as well.

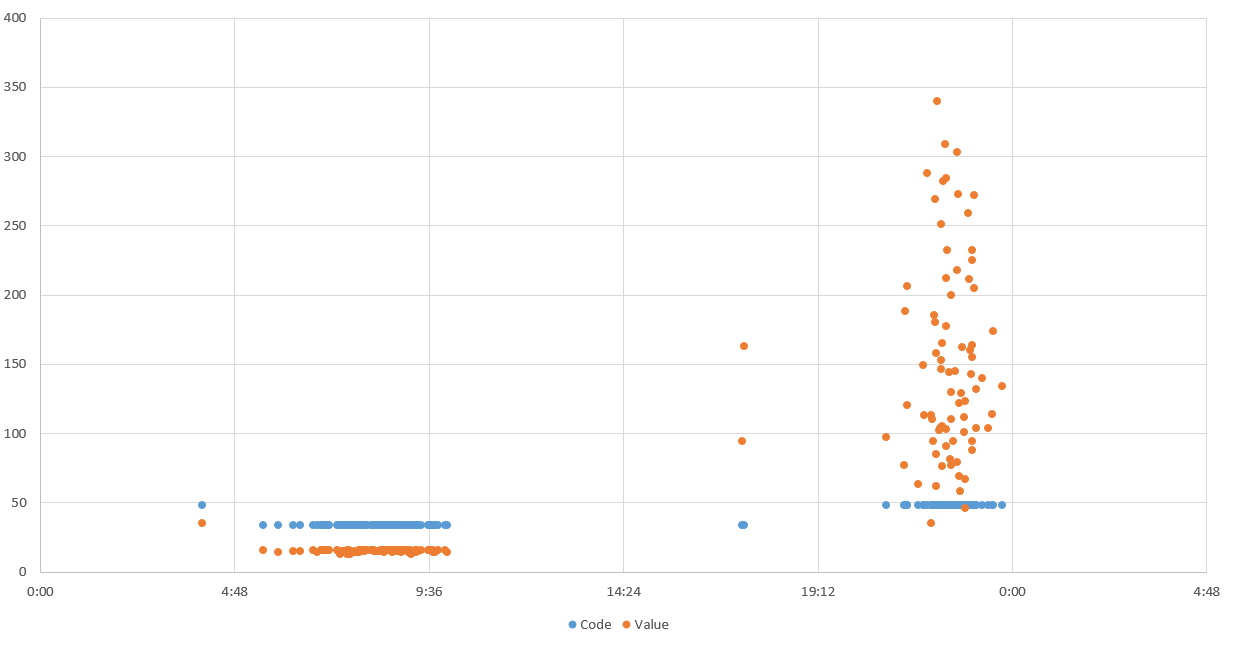
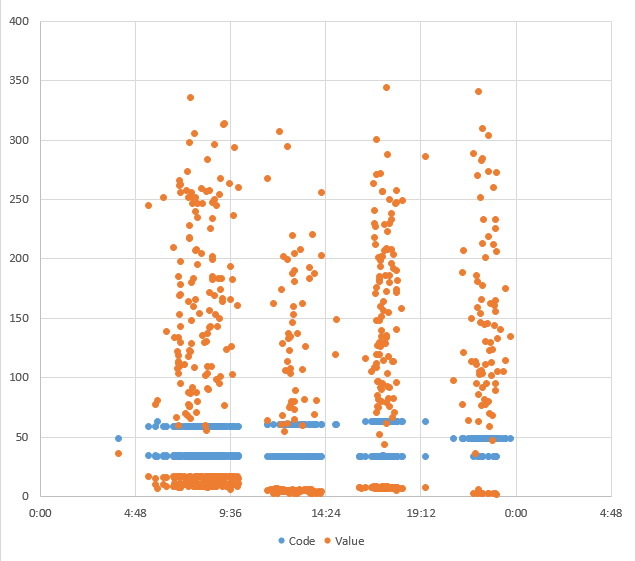
A more in detail analysis is done using R studio. We used clustering and association rule to find the patterns and treatment associated with the glucose value.

# Application

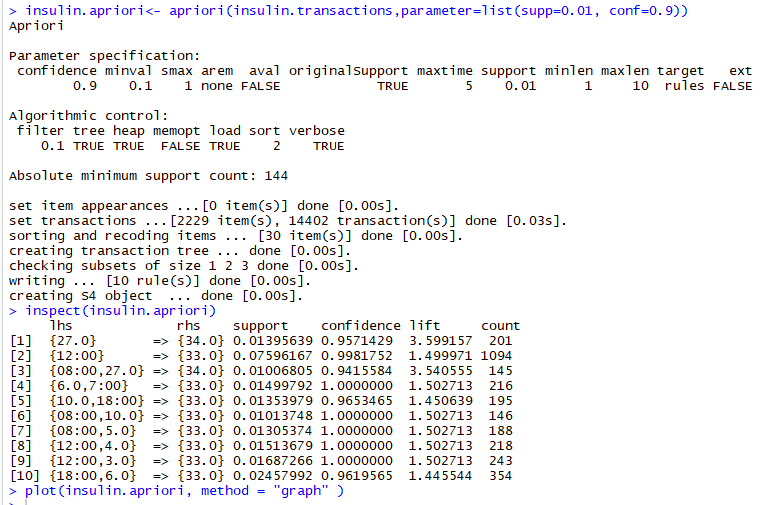
## Pre-processing

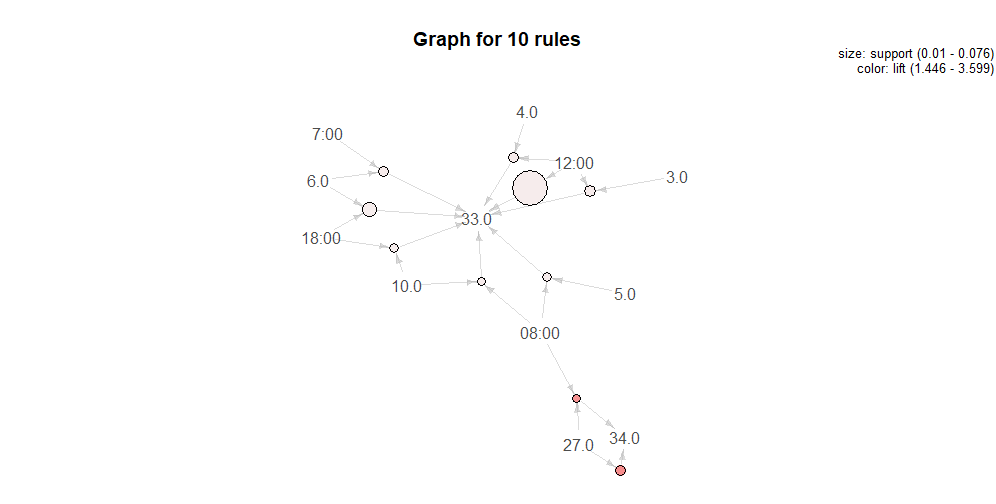
**Identify invalid data outliers and noises**

We have deleted rows with Code value =0, 33,56, since it has no explanation. This type of data is considered noisy and should be avoided for the better performance of the application. We mapped code and value using excel.



On observation through this graph we understood that code and value can be changed with respective to time. Hence, we used association rule mining to find the relations and patterns. Following are the results.

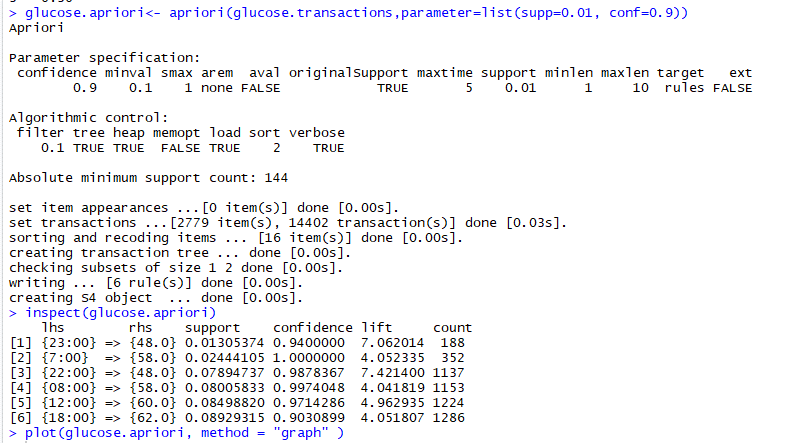


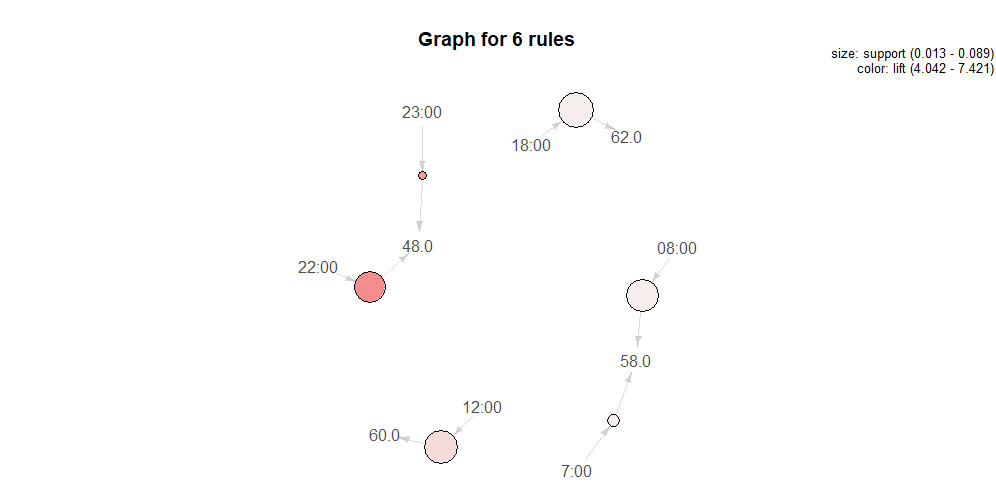


From the results it is observed that for the code 33 the most happening time is at 8:00 in the morning 12:00 in the noon and 6:00 in the evening.

Association rule is introduced to discover which pair of attributes that can occur at the same time frequently. The left-hand side rules and the right-hand rules can indicate blood glucose measurement

This plot shows that normally the time for patients to have regular insulin treatment and it is about 18.00 pm and the value for code regular insulin is under 10.





We can observe from the above that glucose can be clustered based on time which and where we can understand the glucose levels according to the sections of the day. This result matched with the excel graphs we mentioned above.