1.

Personal journals, as repositories of human experiences, play a crucial role in unravelling the complexity of life (Craig and Dillon, 1983). Journals provide insight into individual's lives, exposing their aspirations, apprehensions, joys, sorrows, secrets, fears, and a multitude of events, people and activities, that mould their existence, both positively and negatively (Hamilton, 2018). The objective of this research is to investigate the themes, patterns and sentiments within a personal journal, striving for a thorough life-impact analysis and a refined understanding of how diverse elements can influence emotions and daily experiences.

Recognising that opinions and sentiments are pivotal influences on human behaviour (Taboada, 2016), the research employs sentiment analysis and K-means clustering to categorize journaling entries into distinct sentiment-based clusters. The rationale stems from the desire to understand the emotional essence of journal entries through sentiment analysis. Meanwhile, using K-means involves revealing the inherent subjectivity, polarity and strength of polarity within each entry and comparing them against one another. The application of K-means serves to partition the dataset into subsets of entries that share the highest similarity, enabling the identification of natural groupings in the unlabelled data (Ghosal *et al.*, 2020). To delve into more profound connections between associations and key terms within each cluster, the Apriori Algorithm is applied. This method uncovers relationships among attributes of a dataset and derives rules based on support and confidence metrics (Kettlewell *et al.*, 2020). This step reveals patterns and themes that might be overlooked by solely relying on K-means clustering. Employing these tools can retrospectively reveal not just the factors influencing the writer's emotions, but also explore the connections tied to these sentiments and identify the key aspects of their life that hold the utmost significance to them.

2. Data Exploration

My dataset called 'sdata' comprised of a single variable and 80 records featuring a woman's journal entries. I acquired this information from a Kaggle dataset featuring varied journal entries from different individuals. However, I tailored the data to meet my research objectives by centring the entries around the life of a single person. Upon preliminary analysis of the structure, the dataset compromises 'characters', and each entry record is organised in rows,

offering insights into different moments of the writer's life (figure 1). It became evident that quantifying the data was a necessary step before clustering.

One of the initial steps involved correcting spelling mistakes and addressing missing values using Excel. Subsequently, a word frequency analysis (Appendix A) was conducted in Rstudio to explore and identify relevant and irrelevant words for my analysis, and these results were visualised through the ggplot2 function. Some of the top 20 words appeared non-informative (Figure 2), prompting me to designate them as custom words and remove them before the data pre-processing stages. This process was repeated until the top 20 words were insightful (Figure 3). Data exploration revealed the combined potential of clustering and association rules to elucidate the strength and contextual links among these words in the writer's life and assist in explaining positive and negative associations between them.

Data exploration highlighted valuable insights into the pre-processing requirements, with the importance of standardizing to lowercase, removing punctuation, eliminating numbers, clearing whitespaces and discarding non-informative stop words to enhance data quality. While stemming and lemmatization could simplify words, I opted against these procedures to preserve the value and richness embedded in each journal entry.

```
> head(sdata)
# A tibble: 6 × 1
  Entries
  <chr>
1 I had friends over which felt great but I also had some social anxiety about it.
2 Yesterday I did some research about getting my calendars printed out-- what they will cost, different m…
3 spending time with my husband and kids makes me happy. I was glad to spend quality time with family
4 Work has been slow the last week or so. I haven't been able to hit my daily goal near as often as I wou…
5 I learned that my husband will mis another one of our dinner this Sunday because of her work. I've been…
6 I managed to finish my work relatively early for the day. I'm glad because this opened up free time for…
```

Figure 1: Dataset Overview

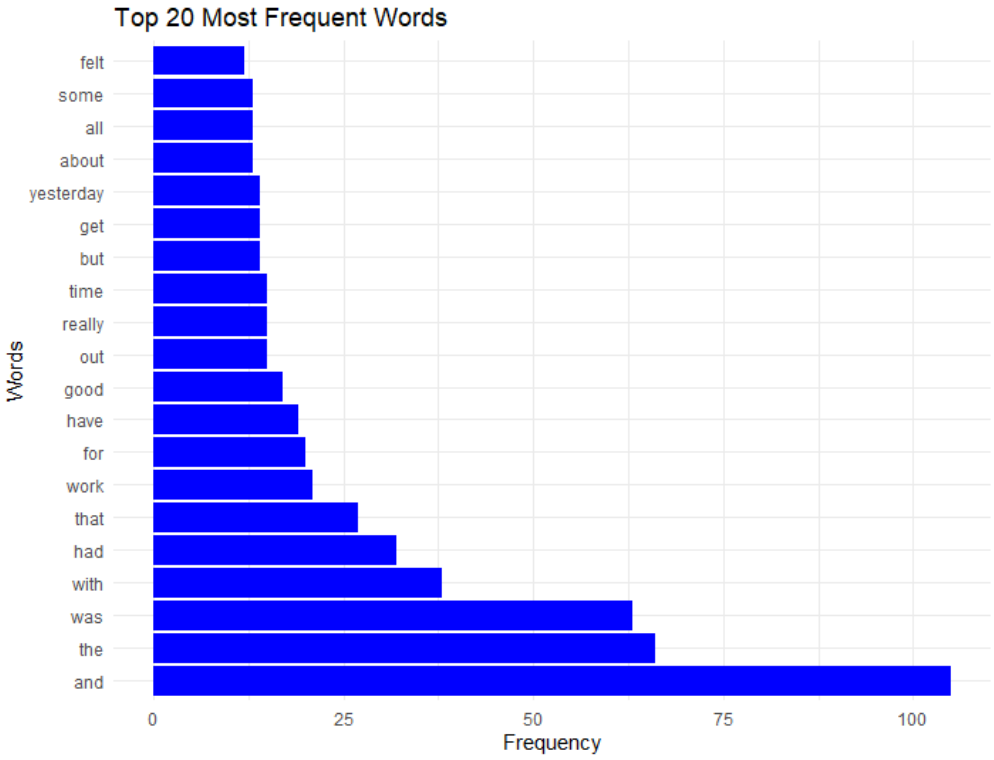


Figure 2: Most frequent words (before data-processing)

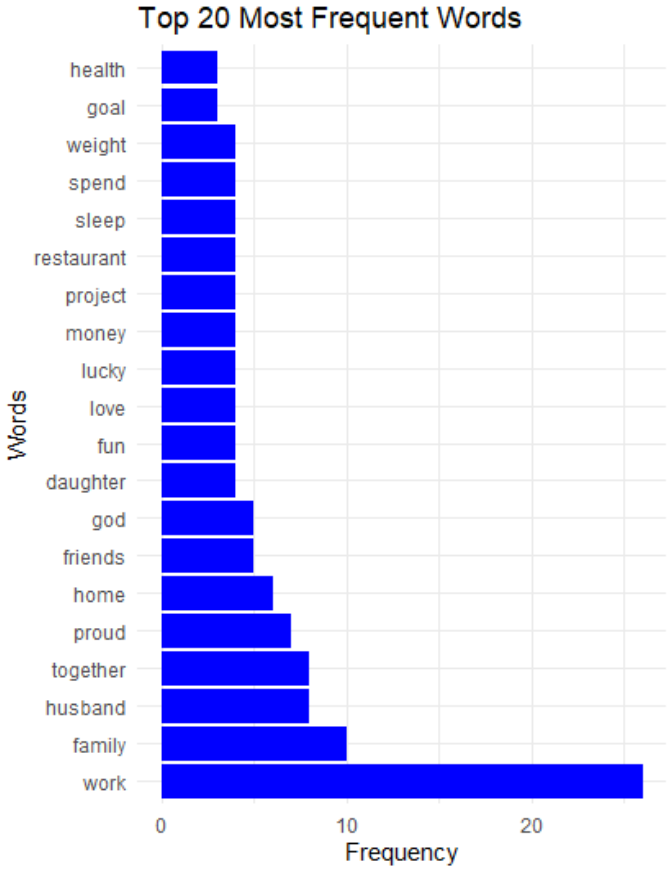


Figure 3: Most frequent words (after data-processing)

2/3. K-means algorithm

Employing the 'syuzhet' package in RStudio, I quantified my qualitative data by computing sentiment scores for each text entry using the 'get_sentiment' function. After reviewing the scores, I removed entries with a sentiment score of zero, considering them neutral and irrelevant. I also normalised the sentiment scores to a range of -5 to +5. Before conducting K-means clustering (Appendix B), I used the 'Elbow Method' to identify the optimal number of clusters (Figure 4). The plot revealed a sharp decline from 1 to 2 clusters, followed by a more gradual decline from 2 to 4. Beyond 4 clusters, the decrease within-cluster sum of squares (WCSS) became subtle. The 'Elbow' in the plot indicates that the optimal number of clusters is 3, as increasing the number of clusters beyond this point would not significantly enhance within-cluster variance. Given the one-dimensional nature of the data, I introduced a pseudo-second dimension for visualization and confirmed the establishment of three separate clusters through a scatter plot (Figure 5).

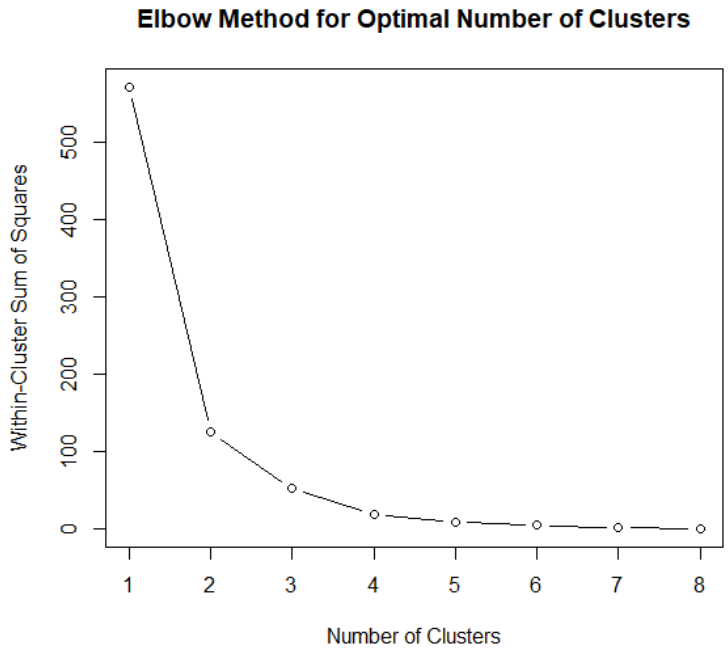


Figure 4: Elbow Method

Figure 5: Clusters Scatter Plot

Results

Figure 6 presents a tibble summarising the clustering analysis results. Cluster 1 comprises of 25 records and an average sentiment of 2.24, suggesting a slightly positive sentiment in the entries within this cluster. Cluster 2, consisting of 15 entries with an average score of -2.87, reflects a predominantly negative sentiment. In contrast, Cluster 3, encompassing 22 entries and an average score of 4.73, suggests a highly positive sentiment. These average scores serve as centroids, demonstrating that K-means clustering has organized text entries into three distinct sentiment categories: slightly (moderately) positive, negative and very positive. Given this outcome, I have chosen to explore the distribution of sentiment scores across the three clusters. Cluster 3 reinforces the notion of slightly positive sentiments, ranging between 3 and zero. In contrast, Cluster 2 also contains negative scores, but unlike Cluster 1, a few scores extend into the positive range, which might indicate texts with mixed sentiments. Cluster 3 stands out, featuring consistent sentiment scores exclusively on the positive end. Furthermore, the graph illustrates that Cluster 1 has the highest number of entries, followed by Cluster 3 and then Cluster 2. After examining Figures 6 and 7, I recognised the need for additional refinement in my analysis to clarify the presence of positive scores in Cluster 2 in

comparison to Cluster 1. Moreover, I am keen to conduct a detailed analysis of each cluster to comprehend the contextual patterns of these sentiments.

```
# A tibble: 3 × 4
  cluster count avg_sentiment
    <int> <int>         <dbl>
1       1    25          2.24
2       2    15         -2.87
3       3    22          4.73
# i 1 more variable: sd_sentiment <dbl>
```

Figure 6: Clusters distribution

Figure 7: Clusters Sentiment Distribution

2/3. Apriori Algorithm

```
> head(cluster_1_data)
# A tibble: 6 × 4
  Entries                                                  sentiment cluster random_dimension
  <chr>                                                        <dbl>   <int>            <dbl>
1 " friends walking us we enjoy spending them"                     2       1           0.0562
2 "everyday sleep is peaceful attitude"                            2       1           0.785
3 " appreciate productive independent"                             2       1           0.103
4 " work issues deadline approaching almost missed luckily c…      3       1           0.355
5 "everyone loves sleep past months created schedule myself …      3       1           0.145
6 " chili cheese quesadillas we chocolate mousse cake lieu a…      3       1          -0.794
> head(cluster_2_data)
# A tibble: 6 × 4
  Entries                                                  sentiment cluster random_dimension
  <chr>                                                        <dbl>   <int>            <dbl>
1 " friends social anxiety "                                      -2       2          -0.425
2 " learned husband mis another sunday work ive patient am s…     -5       2          -0.182
3 " health compare use be ill am "                                -2       2           0.881
4 " tough project work afraid wouldnt int on "                    -2       2          -0.909
5 "god life me go forward im bad no matter how bad know god …     -2       2          -0.0868
6 " so many projects on tight deadline work hectic frustrati…     -1       2           0.914
> head(cluster_3_data)
# A tibble: 6 × 4
  Entries                                                  sentiment cluster random_dimension
  <chr>                                                        <dbl>   <int>            <dbl>
1 "spending husband kids me glad spend quality family"             5       3           0.577
2 " work relatively early im glad opened up free me do work …      5       3           0.766
3 " watched pro wrestling aew nxt fun me love wrestling espe…      5       3          -0.0933
4 " daughter cooked us simple delicious enjoyed sitting toge…      5       3          -0.508
5 " husband plus years worked together cook meal kitchen cle…      5       3          -0.916
6 " morning routines start reading prayer minutes pray god c…      5       3          -0.344
>
```

Figure 8: Cluster's first rows

Examining each cluster (Figure 8), Cluster 1 seems predominantly about references on enjoyable activities and positive attitudes. In contrast, Cluster 2 exhibits references to anxiety, fear and stress, while Cluster 3 includes themes related to family and enjoyment. To establish connections between words within each cluster, association rules were employed. The data for each cluster was downloaded into distinct Excel files and then uploaded into R for the individual application of the Apriori algorithm (Appendix C). Data pre-processing involved identifying words deemed irrelevant for the analysis by inspecting the rows in each cluster. White spaces were also removed. Subsequently, the pre-processed list was converted into a transaction object, the required format for the Apriori algorithm. The 'apriori' function was utilized to mine these transactions for rules. This process involved multiple attempts, adjusting confidence and support levels to define meaningful rules. Plots were generated to identify the top ten frequent items in each cluster, potentially revealing hidden themes overlooked by K-means. The rules were sorted in descending confidence order, and a throughout inspection was carried out.

Results

```
> inspect(head(sort(rules, by = "confidence", decreasing = TRUE)))
    lhs           rhs         support confidence coverage lift     count
[1] {family}   => {proud}     0.08    0.6666667  0.12     4.166667 2
[2] {daughter} => {proud}     0.08    0.6666667  0.12     4.166667 2
[3] {proud}    => {family}    0.08    0.5000000  0.16     4.166667 2
[4] {proud}    => {daughter}  0.08    0.5000000  0.16     4.166667 2
>
```

Figure 9: Cluster 1 Rules

Figure 9 reveals connections among the item's 'family', 'daughter' and 'proud'. A similarly strong association between 'daughter' and 'proud' with identical coverage, support, confidence and lift is observed. The last two rules, in reverse order, show the strong bond between 'family' and 'proud' and 'daughter' and 'proud'. Considering insights from k-means clustering, these rules may explain the positive sentiments in Cluster 1, indicating that the writer's family and her daughter are associated with positive feelings. The top ten item frequency list (Appendix D) suggests other themes associated with and important to her moderate happiness, such as staying healthy, work life, and sharing delicious food/meals.

```
> rules_sorted <- sort(rules, by = "confidence", decreasing = TRUE)
> inspect(rules)
    lhs             rhs    support   confidence coverage  lift count
[1] {}           => {work} 0.4000000 0.4        1.0000000 1.0  6
[2] {frustrating} => {work} 0.1333333 1.0        0.1333333 2.5  2
[3] {boring}     => {work} 0.1333333 1.0        0.1333333 2.5  2
[4] {life}       => {work} 0.1333333 1.0        0.1333333 2.5  2
[5] {frustrated} => {work} 0.1333333 1.0        0.1333333 2.5  2
```

Figure 10: Cluster 2 Rules

Figure 10 highlights an anomaly, indicating that 'work' is present in all transactions based on the coverage score. Subsequent rules consistently associate 'frustrating' and 'boring' with 'work', with a confidence score of 1. 'Work' is also mentioned when 'life' is referenced. These associations can explain the negative sentiments found in cluster 2, suggesting that 'work' is perceived as frustrating and boring to the writer, yet, her life predominantly revolves around it. This is further reinforced in Figure 3 at the data exploration stage. The top ten items diagram underscores that 'bills' contribute to these sentiments and are a negative influence in her life (Appendix D).

```
> itemFrequencyPlot(trans, topN=5)
> rules_sorted <- sort(rules, by = "confidence", decreasing = TRUE)
> inspect(rules)
    lhs          rhs       support    confidence coverage   lift      count
[1] {proud}   => {work}    0.09090909 1          0.09090909  3.142857 2
[2] {enjoyed} => {family}  0.09090909 1          0.09090909  5.500000 2
[3] {team}    => {success} 0.09090909 1          0.09090909 11.000000 2
[4] {success} => {team}    0.09090909 1          0.09090909 11.000000 2
[5] {calm}    => {work}    0.09090909 1          0.09090909  3.142857 2
> |
```

Figure 11: Cluster 3 Rules

Figure 11 suggests that 'proud' consistently appears whenever 'work' is mentioned. 'Family' is five times more likely to appear whenever enjoyed is mentioned. 'Team' and 'success' exhibit the same support as these rules but with a higher lift, indicating a very strong connection between these words. 'Work' is also three times more likely to appear with 'Calm'. These findings can elucidate the mixed sentiments represented in cluster 2, as work seems to evoke both negative and positive feelings. Despite work being occasionally boring and frustrating, it is rewarding, making the writer proud and serving as a source of calmness in her life. Looking at Figure 3, it can also be learned that 'team', 'project' and 'work' are related, as is 'team' with 'success'. Therefore, the writer's source of happiness comes from her work team. Other themes suggesting a profound source of joy include the writer's marriage, her family and sharing meals with her loved-ones (Appendix D).

Discussion

Previous research has exclusively employed personal diaries for the documentation of medical and educational matters. Diaries serve as reflective tools, particularly in ethnographic methods (Browne, 2013). In medicine, diaries are utilized in thematic analysis, such as nurses maintaining personalised diaries to enhance their comprehension of patients. For instance, during COVID-19, intensive care units documented patient clinical status and recovery (Galazzi *et al.*, 2023). Similarly, resource diaries have been proven effective in stabilizing mental health during psychiatric treatment (Mikkelsen, 2018).

Exploring data mining in personal journals, which document our private lives, remains relatively uncharted. This is likely due to concerns about the sensitive nature of the data and ethical considerations regarding privacy. However, this novel application has the potential to

yield fresh insights for the fields of mental health and wellness, as well as for medical domains dealing with memory loss illnesses and cognitive impairment.

Moreover, while research investigated the impact of major life events on well-being, the influence of daily, casual events, both positive and negative, remains largely unexplored. Existing studies explore how major life events affect cognitive and affective well-being (Kettlewell *et al.*, 2020), as well as the quality of life in relation to social support and personality factors (Kettlewell *et al.*, 2020). However, the underexplored aspect lies in understanding how smaller, day-to-day aspects of life influence overall life and emotional development, and this application serves as a starting point to address this gap.

Appendix

Appendix A: Word frequency code

```
> tdm <- TermDocumentMatrix(sdata$Entries)
>
> m <- as.matrix(tdm)
>
>
> word_freq <- rowSums(m)
>
>
> word_freq_df <- data.frame(word = names(word_freq), freq = word_freq)
>
> word_freq_df <- word_freq_df[order(-word_freq_df$freq), ]
> ggplot(data = word_freq_df[1:20, ], aes(x = reorder(word, -freq), y = freq)) +
+     geom_bar(stat = "identity", fill = "blue") +
+     coord_flip() +
+     labs(x = "Words", y = "Frequency") +
+     ggtitle("Top 20 Most Frequent Words") +
+     theme_minimal()
```

Appendix B: Sentiment Analysis and K-means Code

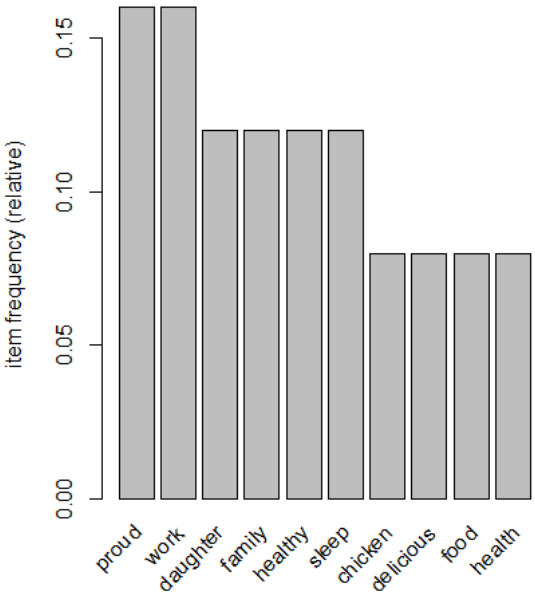
```
library(tm)
library(syuzhet)
library(ggplot2)
library(slam)
library(dplyr)

# Function to clean and preprocess text
clean_and_preprocess_text <- function(text) {
  custom_stopwords <- c(
    "yesterday", "today", "tomorrow", "time", "day", "feel", "felt",
    "lot", "something", "well", "take", "first", "one", "some",
    "important", "now", "best", "thing", "finally", "also", "always",
    "back", "bit", "got", "get", "a", "the", "of", "with", "was",
    "it", "my", "to", "i", "in", "didn't", "nice", "great", "really", "things",
    "just", "makes", "come", "dont", "done", "week", "like", "can",
    "good", "like", "really", "happy", "made", "just", "days", "able",
    "great", "night", "going", "much", "always", "trying", "even", "way",
    "long", "hours", "think", "last", "new", "didnt", "that", "had", "for", "about", "all", "out",
    "feeling", "helped", "make", "will", "getting", "most", "they", "from", "not", "its", "were", "very", "been", "this", "but", "have",
    "her", "our", "after", "more", "could", "each", "then", "through", "what", "went", "did", "which",
    "end", "having", "due", "has", "those", "and", "when",
    "because", "evening", "being", "before", "there",
    "dinner", "doing", "ready", "than", "whole",
    "since", "fairly", "pretty",
    "eat", "over",
    "different", "would", "talk", "finish", "job", "managed", "need", "often", "talked",
    "daily",
    "ate", "at", "any", "are"
  )
  corpus <- Corpus(VectorSource(text))
  corpus <- tm_map(corpus, content_transformer(tolower))
  corpus <- tm_map(corpus, removePunctuation)
  corpus <- tm_map(corpus, removeNumbers)
  corpus <- tm_map(corpus, removeWords, custom_stopwords)
  corpus <- tm_map(corpus, stripWhitespace)
  return(unlist(sapply(corpus, as.character)))
}


sdata$Entries <- clean_and_preprocess_text(sdata$Entries)

# Sentiment Analysis

sdata$sentiment <- get_sentiment(sdata$Entries, method = "afinn")
sdata <- sdata[sdata$sentiment != 0, ]
sdata$sentiment <- pmin(pmax(sdata$sentiment, -5), 5)
```

```r
#K-means clustering

set.seed(123)
max_clusters <- 8  #maximum number of clusters is set to the number of unique sentiment scores
wcss <- numeric(max_clusters)

for (i in 1:max_clusters) {
  kmeans_result <- kmeans(as.matrix(sdata$sentiment), centers = i, nstart = 20)
  wcss[i] <- kmeans_result$tot.withinss
}

plot(1:max_clusters, wcss, type = "b", xlab = "Number of Clusters", ylab = "Within-Cluster Sum of Squares",
     main = "Elbow Method for Optimal Number of Clusters")

num_clusters <- 3
kmeans_result <- kmeans(as.matrix(sdata$sentiment), centers = num_clusters, nstart = 20)
sdata$cluster <- kmeans_result$cluster

 # Visualizations and Analysis

set.seed(123)
sdata$random_dimension <- runif(nrow(sdata), min = -1, max = 1)

ggplot(sdata, aes(x = sentiment, y = random_dimension, color = factor(cluster))) +
  geom_point() +
  labs(title = "Cluster Visualization on Sentiment Scores",
       x = "Sentiment Score", y = "Random Dimension",
       color = "Cluster") +
  theme_minimal()

sdata %>%
  group_by(cluster) %>%
  summarise(
    count = n(),
    avg_sentiment = mean(sentiment),
    sd_sentiment = sd(sentiment)
  )

ggplot(sdata, aes(x = sentiment, fill = factor(cluster))) +
  geom_histogram(binwidth = 0.5) +
  facet_wrap(~cluster) +
  labs(title = "Sentiment Score Distribution by Cluster",
       x = "Sentiment Score", y = "Count",
       fill = "Cluster") +
  theme_minimal()
```

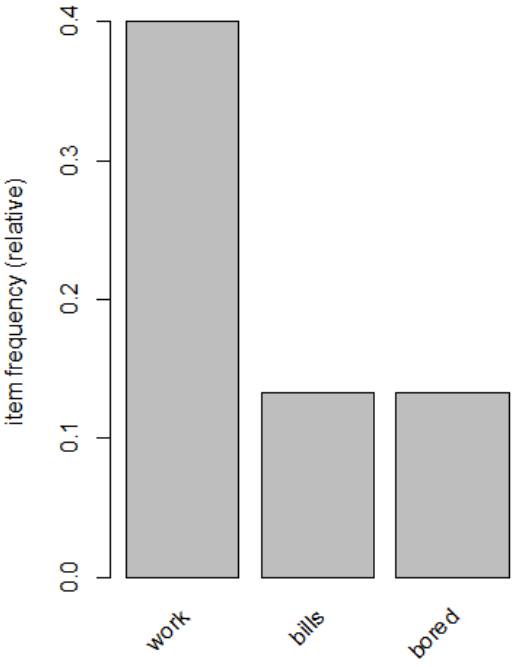
Appendix C: Apriori Algorithm code

```r
library(tm)
library(arules)
library(stringr)

#pre-processing stopwords removal performed separately for each cluster

# Cluster 1
stop_words <- c("we", "everyday", "is", "she", "do", "me", "you", "so",
                "am", "im", "up", "how", "try", "be", "an", "working", "again")

# Cluster 2
stop_words <- c("we", "everyday", "is", "she", "do", "me", "you", "so",
                "am", "im", "up", "how", "try", "be", "an", "working", "again", "int",
                "go", "no",
               "on", "use", "everything", "vs", "wasnt", "by", "far", "or")

# Cluster 3
stop_words <- c("we", "everyday", "is", "she", "do", "me", "you", "so",
                "am", "im", "up", "how", "try", "be", "an", "working", "again", "these", "us", "thought", "lol", "two",
                "on", "other", "everyone", "better", "decided", "anything", "came", "by", "together")


remove_words <- function(text, words) {

  removed <- str_replace_all(text, paste0("\\b(", paste(words, collapse = "|"), ")\\b"), "")
  removed <- str_trim(removed)
  words <- unlist(str_split(removed, " "))
  words <- words[words != ""]
  return(words)
}

preprocessed_docs <- lapply(cluster3$Entries, function(x) remove_words(x, stop_words))

trans <- as(preprocessed_docs, "transactions")

#support and confidence adjusted as needed
rules <- apriori(trans,
                 parameter = list(supp = 0.08, conf = 1))
itemFrequencyPlot(trans, topN=6)

rules_sorted <- sort(rules, by = "confidence", decreasing = TRUE)

inspect(rules)
```

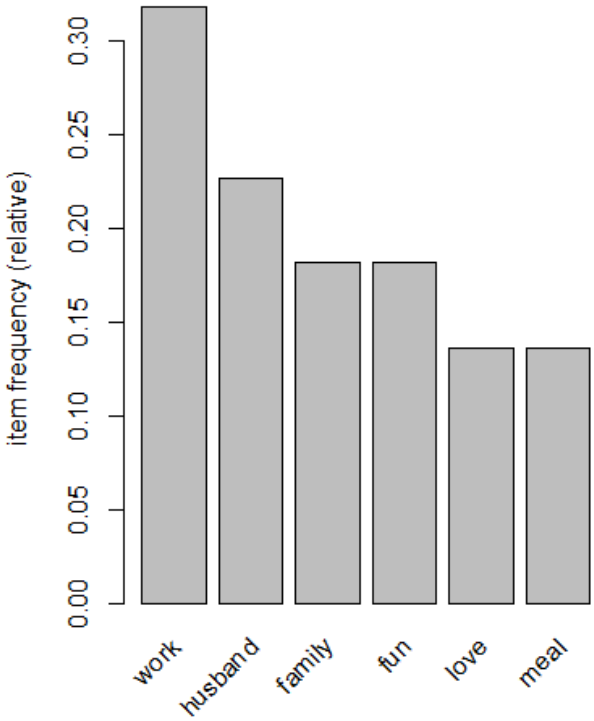
Appendix D: Top frequent items in each cluster



Cluster 1



Cluster 2



Cluster 3