UNIVERSITI
KEBANGSAAN
MALAYSIA
The National University of Malaysia

SEMESTER 2, 2022/2023

STQD6114 – ANALITIK DATA TAK BERSTRUKTUR

Analitik data tak berstruktur

LECTURER:

DR. NOR HAMIZAH BINTI MISWAN

Name	Matric Number
Hongyu An	P120996

CHAPTER I

INTRODUCTION

The dataset provided is a subset of a larger dataset consisting of book reviews from the Amazon Kindle Store category. It covers a period from May 1996 to July 2014 and contains a total of 982,619 entries. The dataset follows a 5-core structure, meaning that each reviewer and each product in the dataset have at least five reviews associated with them.

The dataset is organized into several columns, including:
asin: ID of the product, such as "B000FA64PK".
helpful: helpfulness rating of the review, represented as a fraction (e.g., 2/3).
overall: rating of the product.
reviewText: text of the review.
reviewTime: time of the review in a raw format.
reviewerID: ID of the reviewer, like "A3SPTOKDG7WBLN".
reviewerName: name of the reviewer.
summary: summary of the review.
unixReviewTime: unix timestamp representing the time of the review.

This article only retains the comment column as a data table analysis.
This project serves as a demonstration of various techniques used to analyze a dataset of book reviews from the Amazon Kindle Store category. The main objective of this project is to uncover prominent themes occurring in the reviews and gain insights into the sentiments expressed by the reviewers.

The project begins by creating a word frequency table, which provides a quantitative representation of the occurrence of different words in the reviews. By visualizing this information through a word cloud, the most prominent themes or topics can be identified.

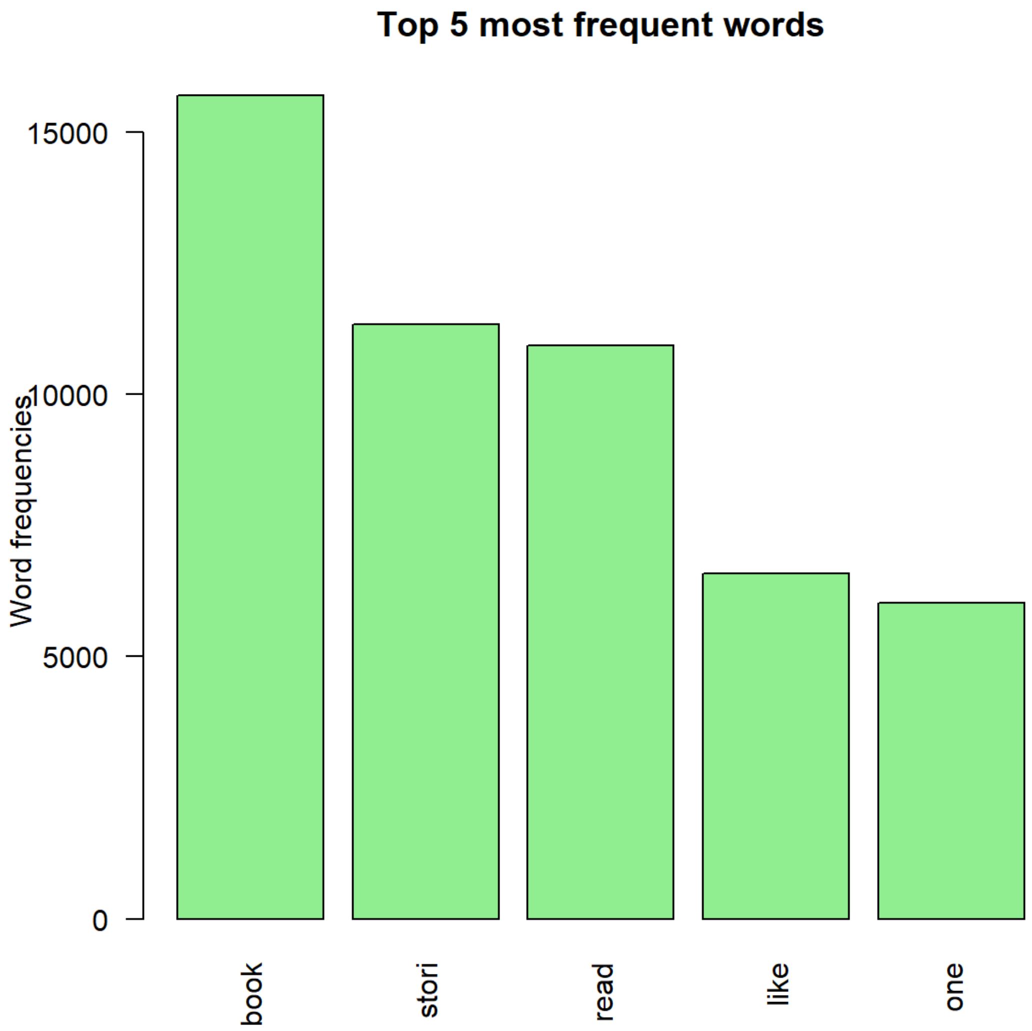
To further understand the context surrounding these themes, word association analysis using correlation is employed. This technique helps establish relationships between words and provides additional insights into the connections between different themes.

The project also explores four methods to generate sentiment scores. These scores allow for assigning a numeric value to the strength of positivity or negativity in the review texts. By analyzing the average sentiment across the dataset, it becomes possible to determine the overall sentiment trend, whether it is positive or negative.

In addition to sentiment analysis, the project implements an emotion classification approach using the NRC sentiment lexicon. This allows for the identification and analysis of different emotions expressed in the text. The project presents two plots to visualize and interpret the emotions found in the reviews.

CHAPTER II

CONTENTS



As can be seen from the above figure, the words with the highest frequency in the comments are book, stori, read, like, one, and all of them exceed 5000 times.

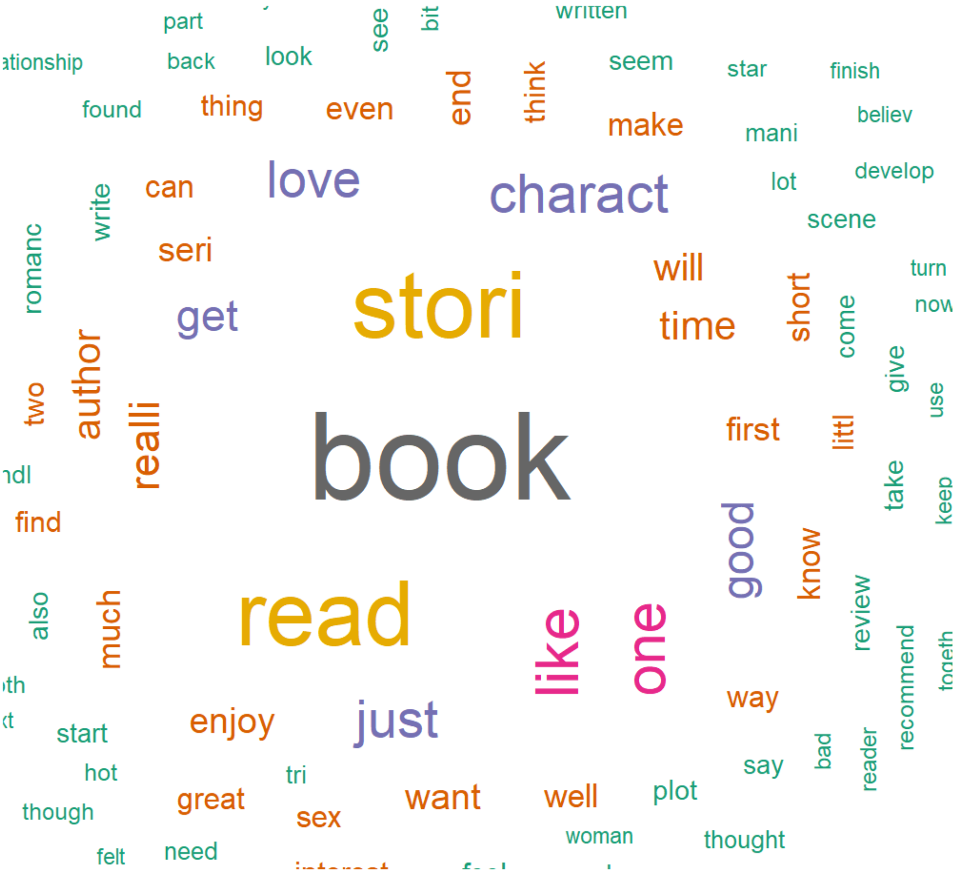
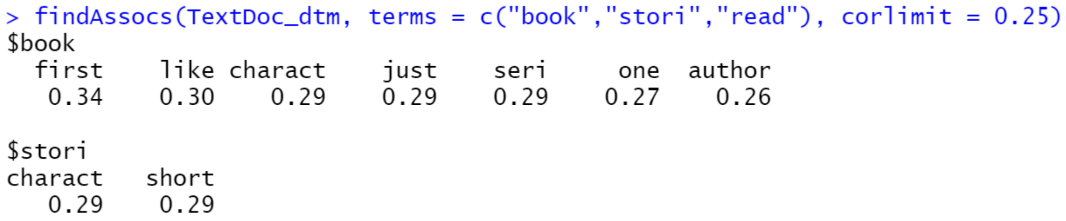


Figure cloud



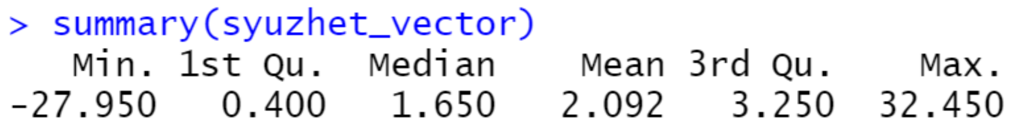
```
> findAssocs(TextDoc_dtm, terms = c("book","stori","read"), corlimit = 0.25)
$book
  first    like charact    just    seri     one  author 
   0.34    0.30    0.29    0.29    0.29    0.27    0.26 

$stori
charact   short 
   0.29    0.29 
```

This script shows which words are most frequently associated with the top three terms (corlimit = 0.25 is the lower limit/threshold I have set.

We can set it lower to see more words, or higher to see less).

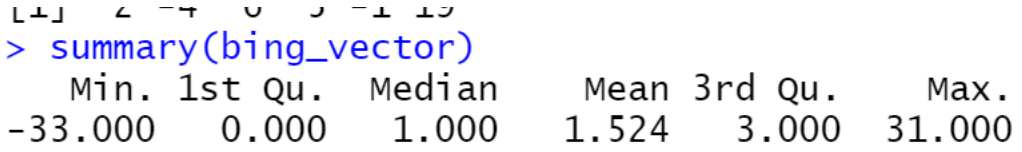
The output indicates that “first”, “like”, “charact”(which is root word of character),"just", "seri"(which is root word of series),"one" and "author" have occur ~30% of the time with the word “book”.

Similarly we can say that "charact"(which is root word of character),"short" have occur ~ 30% of the time with the word "stori"(which is root word of stories).



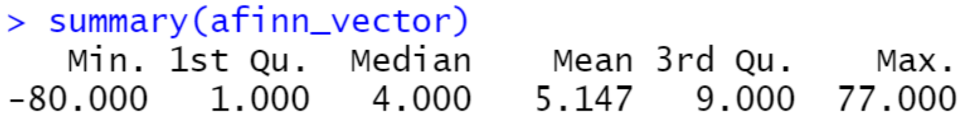
```
> summary(syuzhet_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-27.950   0.400   1.650   2.092   3.250  32.450 
```

The summary statistics of the suyzhet vector show a median value of 1.6, which is above zero and can be interpreted as the overall average sentiment across all the responses is positive.



```
> summary(bing_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-33.000   0.000   1.000   1.524   3.000  31.000 
```

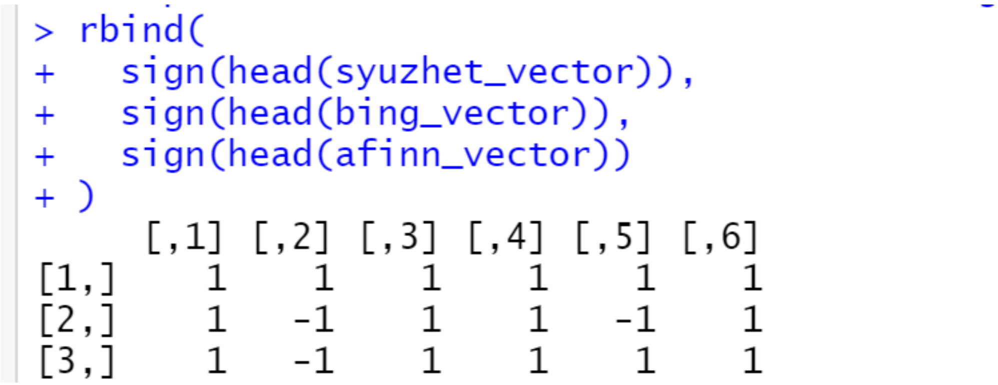
Finally, I compared the duration of the most popular movies.



```
> summary(afinn_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-80.000   1.000   4.000   5.147   9.000  77.000 
```

The summary statistics of bing and afinn vectors also show that the Median value of Sentiment scores is above 0 and can be interpreted as the overall average sentiment across the all the responses is positive.

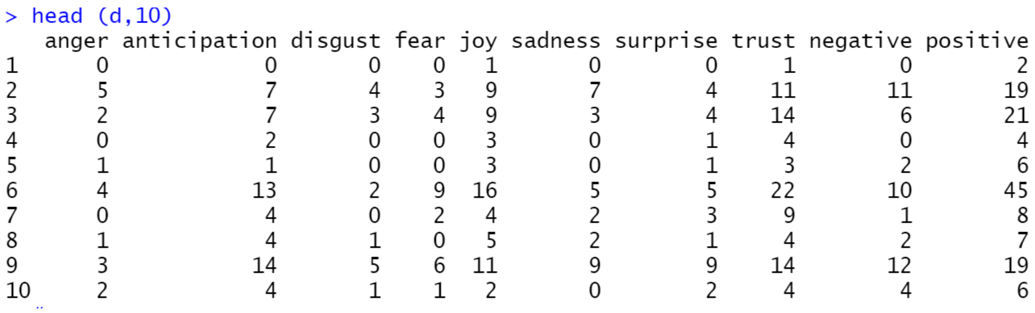
Because these different methods use different scales, it’s better to convert their output to a common scale before comparing them. This basic scale conversion can be done easily using R’s built-in sign function, which converts all positive number to 1, all negative numbers to -1 and all zeros remain 0.



```
> rbind(
+   sign(head(syuzhet_vector)),
+   sign(head(bing_vector)),
+   sign(head(afinn_vector))
+ )
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    1    1    1    1
[2,]    1   -1    1    1   -1    1
[3,]    1   -1    1    1    1    1
```

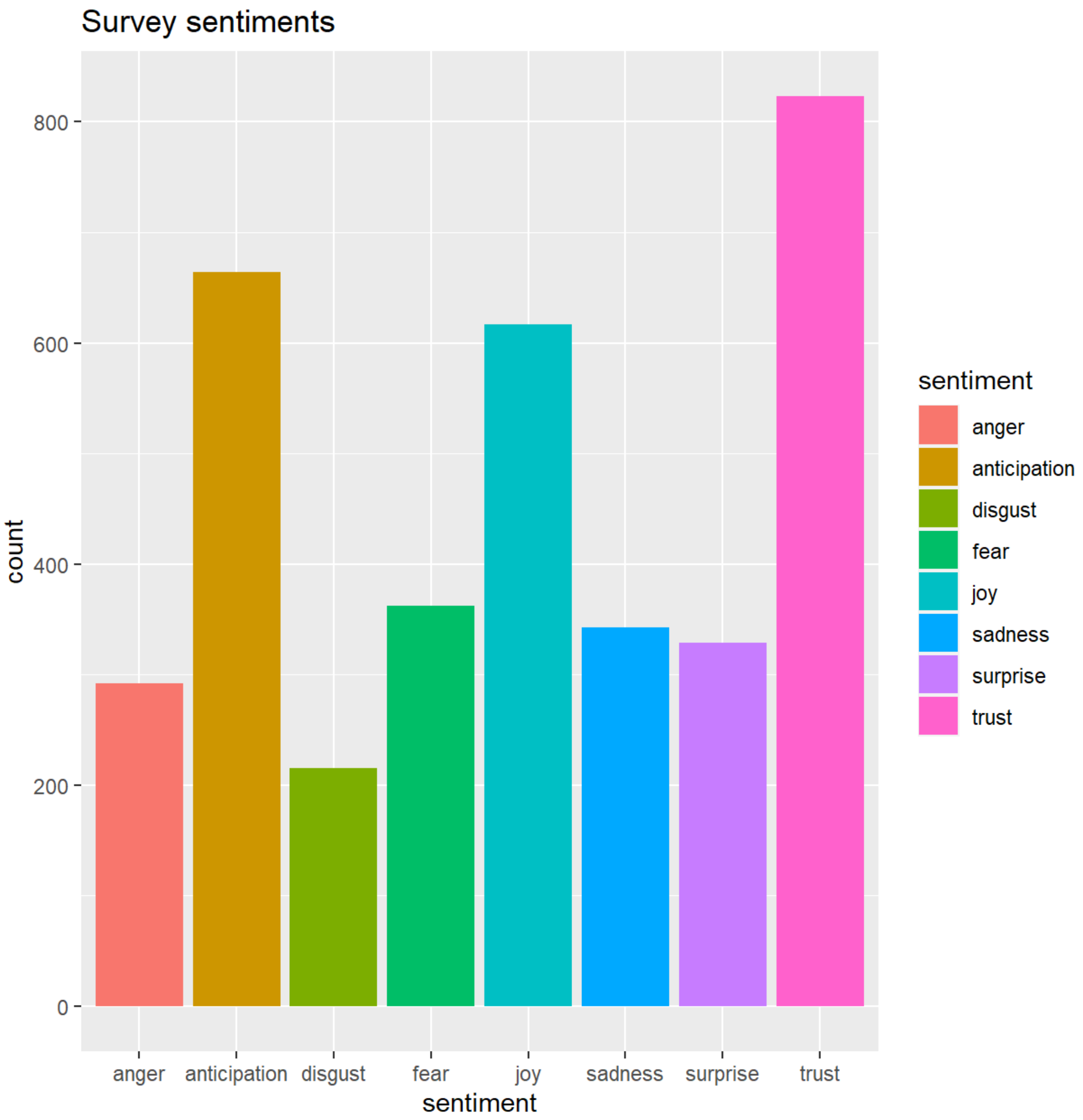
Figure 9. Normalize scale and compare three vectors

Note the first element of each row (vector) is 1, indicating that all three methods have calculated a positive sentiment score, for the first response (line) in the text.

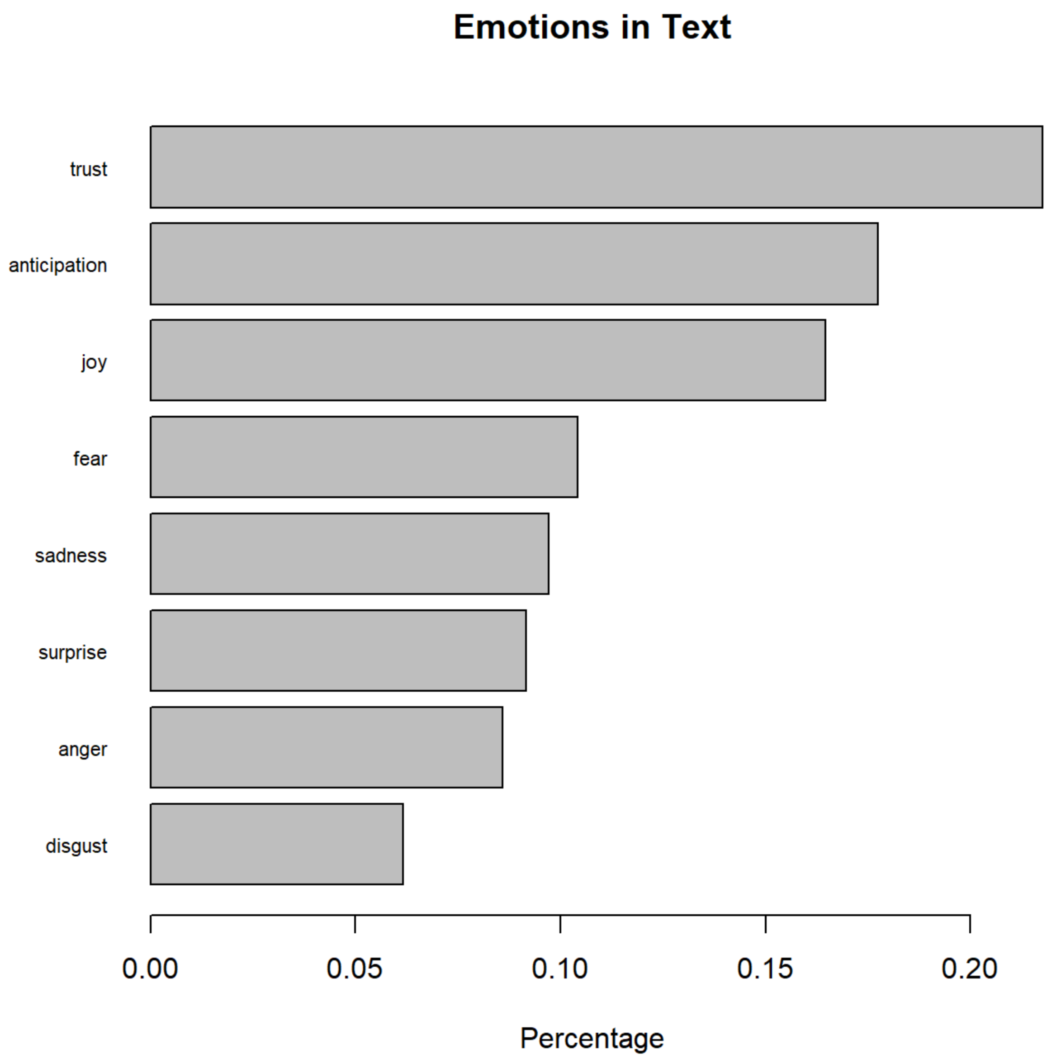


```
> head (d,10)
   anger anticipation disgust fear joy sadness surprise trust negative positive
1      0            0       0    0   1       0        0     1        0        2
2      5            7       4    3   9       7        4    11       11       19
3      2            7       3    4   9       3        4    14        6       21
4      0            2       0    0   3       0        1     4        0        4
5      1            1       0    0   3       0        1     3        2        6
6      4           13       2    9  16       5        5    22       10       45
7      0            4       0    2   4       2        3     9        1        8
8      1            4       1    0   5       2        1     4        2        7
9      3           14       5    6  11       9        9    14       12       19
10     2            4       1    1   2       0        2     4        4        6
```

It can be seen from this table that the number of positive words in each comment is greater than negative.



This bar chart demonstrates that words associated with the positive emotion of “trust” occurred more than eighth hundred times in the review, whereas words associated with the negative emotion of “disgust” occurred 200 times. The word “surprise” occurred in the review The word "anger" appeared more than 300 times, while the word "anger" appeared about 300 times. The word "surprise" appeared more than 300 times in the comments, while the word "sadness, fear" appeared about 350 times.



From this figure, we can clearly draw a conclusion that the words that are positively related to all comments and trust account for the highest proportion, indicating that the evaluation of this book is very good.

CHAPTER III

CONCLUSION

This project was demonstration of how to create a word frequency table and plot a word cloud, to identify prominent themes occurring in the review.

Word association analysis using correlation, helped gain context around the prominent themes.

It explored four methods to generate sentiment scores, which proved useful in assigning a numeric value to strength (of positivity or negativity) of sentiments in the text and allowed interpreting that the average sentiment through the text is trending positive.

Lastly, it demonstrated how to implement an emotion classification with NRC sentiment and created two plots to analyze and interpret emotions found in the text.

We can conclude that customers are optimistic and positive about this book. This type of book may be more popular with readers. We can increase the purchase and writing of this type of book to increase profits.