# STQD 6114 Project 2

## Muhd Hilman Bin Rozaini (P121535)

### Introduction

This project wishes to analyse and understand the behaviour of text data. Therefore, there are 2 parts of text analysis in this project, where part 1 analyses text documents and part 2 focuses more on sentiment analysis but using a different dataset.

In part 1 of the project, the objective is to identify the pattern on text data through understanding the term association between terms within documents and between documents using Latent Dirichlet Allocation (LDA) topic modelling analysis. To further analyse the similarity or correlation distance between text documents, clustering analysis is also performed to further understand any possible clustering relationship between the text documents.

In part 2, kindle review dataset is used as this part focuses on sentiment analysis. Lexicon is used to identify the pattern of the review on the dataset. The lexicon technique used includes Syuzhet, Bing, AFINN, and National Research Council (NRC) sentiment analysis.

### Dataset background for task 1 & task 2

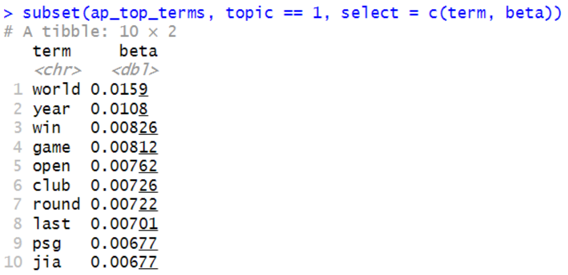
The dataset used for task 1 and task 2 are news articles related to sports. The total number of text documents involved in this analysis are 40 text documents, sourced mainly from local online news company such as New Strait Times, the Stars, Malaymail, and to name a few.

The language of the online news article are all in English, and the type of sports involved are varied, namely but not limited to badminton, golf, tennis, martial arts, swimming and hockey.

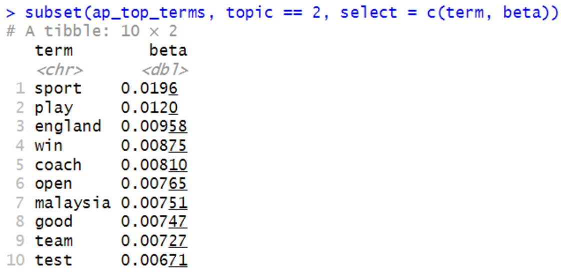
## Part 1

### Task 1:

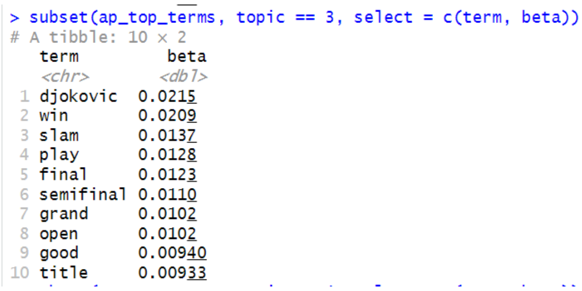
### i) Topic modelling analysis using LDA.



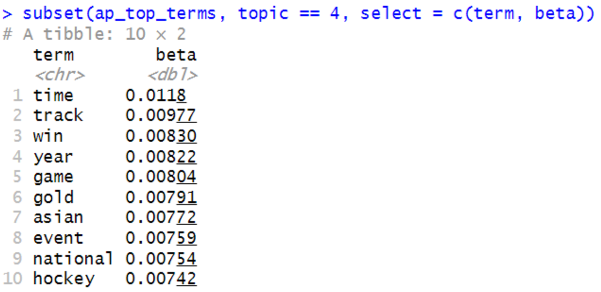
```
> subset(ap_top_terms, topic == 1, select = c(term, beta))
# A tibble: 10 × 2
   term     beta
   <chr>   <dbl>
 1 world 0.0159 
 2 year  0.0108 
 3 win   0.00826
 4 game  0.00812
 5 open  0.00762
 6 club  0.00726
 7 round 0.00722
 8 last  0.00701
 9 psg   0.00677
10 jia   0.00677
```



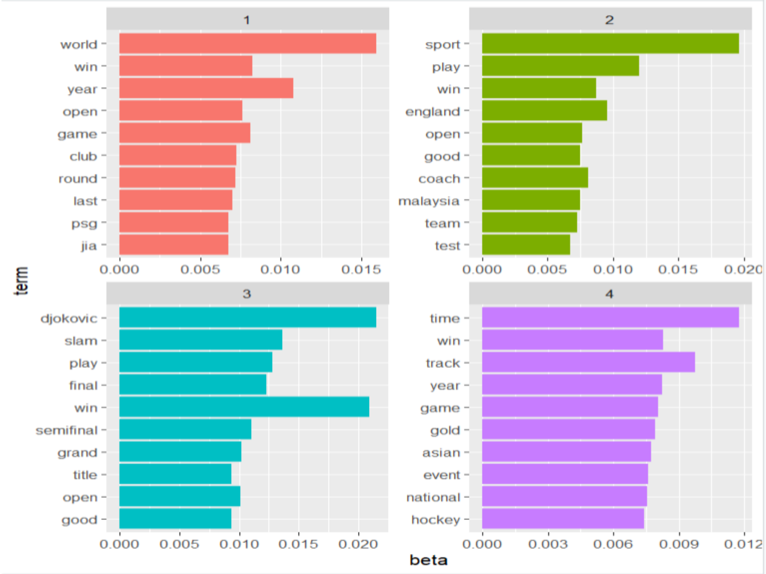
```
> subset(ap_top_terms, topic == 2, select = c(term, beta))
# A tibble: 10 × 2
   term        beta
   <chr>      <dbl>
 1 sport    0.0196 
 2 play     0.0120 
 3 england  0.00958
 4 win      0.00875
 5 coach    0.00810
 6 open     0.00765
 7 malaysia 0.00751
 8 good     0.00747
 9 team     0.00727
10 test     0.00671
```



```
> subset(ap_top_terms, topic == 3, select = c(term, beta))
# A tibble: 10 × 2
   term         beta
   <chr>       <dbl>
 1 djokovic  0.0215 
 2 win       0.0209 
 3 slam      0.0137 
 4 play      0.0128 
 5 final     0.0123 
 6 semifinal 0.0110 
 7 grand     0.0102 
 8 open      0.0102 
 9 good      0.00940
10 title     0.00933
```



```
> subset(ap_top_terms, topic == 4, select = c(term, beta))
# A tibble: 10 × 2
   term        beta
   <chr>      <dbl>
 1 time     0.0118 
 2 track    0.00977
 3 win      0.00830
 4 year     0.00822
 5 game     0.00804
 6 gold     0.00791
 7 asian    0.00772
 8 event    0.00759
 9 national 0.00754
10 hockey   0.00742
```



From the Latent Dirichlet Allocation (LDA) topic modelling analysis, we can observe that each topic with the highest beta value is distinctive between other topics. Namely, the highest beta word in topic 1 is "world", topic 2 is "sport", topic 3 is "Djokovic", and topic 4 is "time" with beta value of 0.0159, 0.0196, 0.0215, and 0.0118 respectively. These highest beta value words provide an inference into the main theme and concepts for each topic.

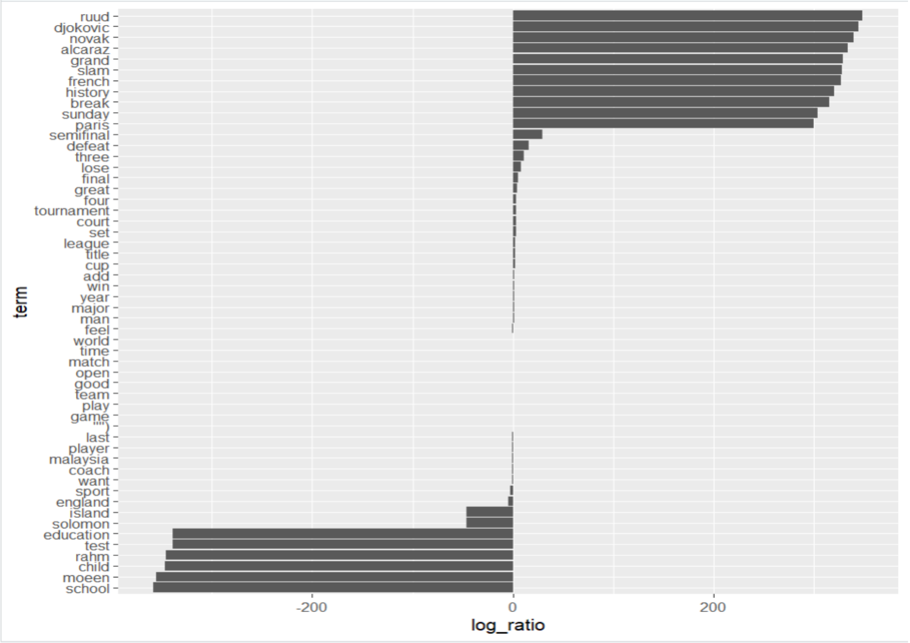
For example in topic 1, the highest beta value word is “world”, and the relevant terms from the plot includes “win”, “year”, “club”, “PSG” and to name a few. From these relevant terms, we can infer that the theme or concept revolving around topic 1 is related to soccer theme.

As for topic 2, the highest beta value is “sport”, while the relevant associated terms in topic 2 are “play”, “England”, “Malaysia”, “coach”, “open” and others. From here, we can infer that topic 2 revolves around the theme of “open international competition”. This is because the term “England” and “Malaysia” is involved, and the term “coach” signifies further that topic 2 is related to “open sports competition”, for example “badminton open” as usually there will be coaches present in all badminton open competition.

Further in topic 3, the highest beta value word is “Djokovic”. As “Djokovic” is the name of a tennis player, thus the associated words of “slam”, “final”, “grand”, “title”, and others are associated with tennis competition in topic 3.

Lastly in topic 4, the top 10 word’s beta value are quite high as compared to other topics. The word “time” has the highest beta value, and there are words such as “track” which might indicate cycling tracks or sprinting tracks, and there is another distinct word such as “hockey” which is another sport. Therefore, we can infer that the theme of topic 4 is a variety of sports that involves “time” as a crucial variable.

**ii) Terms that are most common within topic 2 and topic 3**

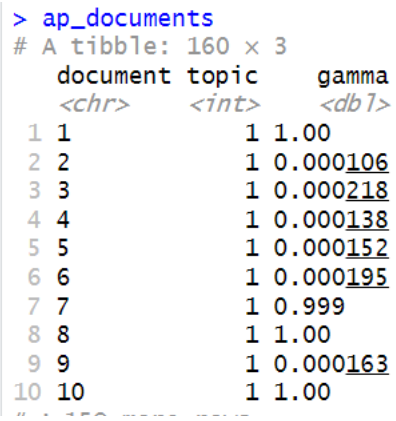


Between all possible combination of topics ranging from topic 1 to topic 4, an analysis of common words between topic 2 and topic 3 is chosen. A positive value indicates that the terms are more common and relevant in topic 3, while a negative value in log-ratio indicates that the terms are more common and relevant in topic 2. The opposite poles also signifies that the terms of the opposite poles are less common with the counterpart topics.

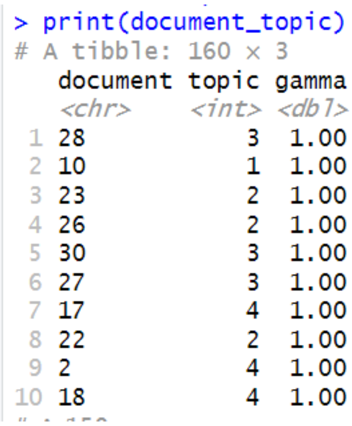
Based on the log-ratio value, the terms “ruud”, “Djokovic”, “novak”, “alcaraz”, “grand”, “slam”, French”, “history”, “break”, “Sunday” and “paris” are the most common terms in topic 3, but it is the least common terms in topic 2. Conversely, the terms “education”, “test”, “rahm”, “moeen”, and “school” are most common in topic 2 and least common in topic 3.

As the magnitude of the log-ratio value is large between the terms in topic 2 and topic 3, it provides an insight that the two topics have a distinctive language pattern where topic 2 revolves more on generic sports. That is why the terms such as “education”, “test”, “Rahm” are not relevant to topic 3 which revolves specifically on tennis sports.

### iii) Gamma values on per-document-per-topic-probabilities



```
> ap_documents
# A tibble: 160 × 3
   document topic    gamma
   <chr>    <int>    <dbl>
 1 1            1 1.00    
 2 2            1 0.000106
 3 3            1 0.000218
 4 4            1 0.000138
 5 5            1 0.000152
 6 6            1 0.000195
 7 7            1 0.999   
 8 8            1 1.00    
 9 9            1 0.000163
10 10           1 1.00    
```



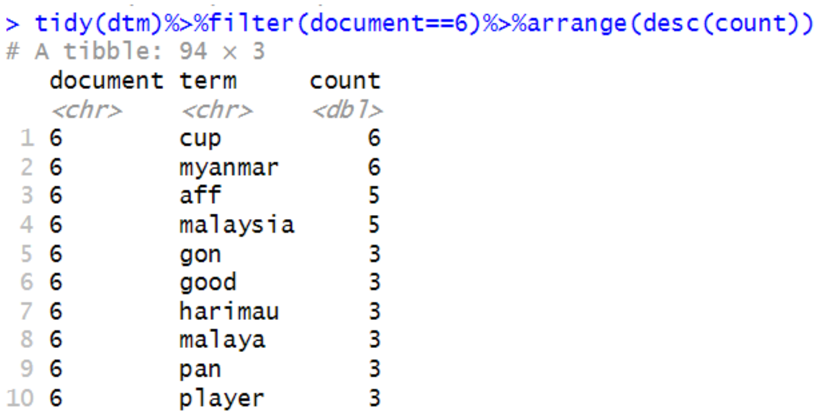
```
> print(document_topic)
# A tibble: 160 × 3
   document topic gamma
   <chr>    <int> <dbl>
 1 28           3  1.00
 2 10           1  1.00
 3 23           2  1.00
 4 26           2  1.00
 5 30           3  1.00
 6 27           3  1.00
 7 17           4  1.00
 8 22           2  1.00
 9 2            4  1.00
10 18           4  1.00
```

From the left tibble, we can observe the first 10 documents and their respective gamma values and assigned topic category.

Document 1, 7, 8 and 10 has gamma values of 1 or near 1 indicates that these documents are most relevant to topic 1, revolving around concept of soccer sports. The residual documents, namely document 2,3,4,5,6 and 9 have smaller gamma value and thus indicate that these documents are relatively not relevant to topic 1.

In regard to the right tibble, we can see each document is most relevant to a specific topic based on the highest gamma values.

### iv) The most common words in the document 6, chosen at random



```
> tidy(dtm)%>%filter(document==6)%>%arrange(desc(count))
# A tibble: 94 × 3
   document term     count
   <chr>    <chr>    <dbl>
 1 6        cup          6
 2 6        myanmar      6
 3 6        aff          5
 4 6        malaysia     5
 5 6        gon          3
 6 6        good         3
 7 6        harimau      3
 8 6        malaya       3
 9 6        pan          3
10 6        player       3
```

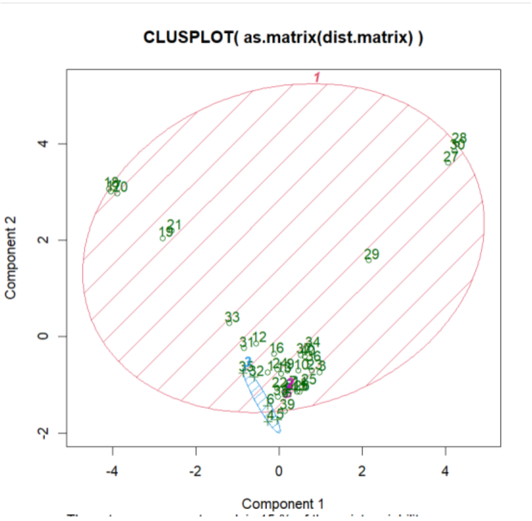
In regard to the tibble, we are observing the most common terms specifically in document 6. The most common terms in document 6 are “cup” and “Myanmar” with 6 counts, following with “aff”, “Malaysia” having 6 counts.

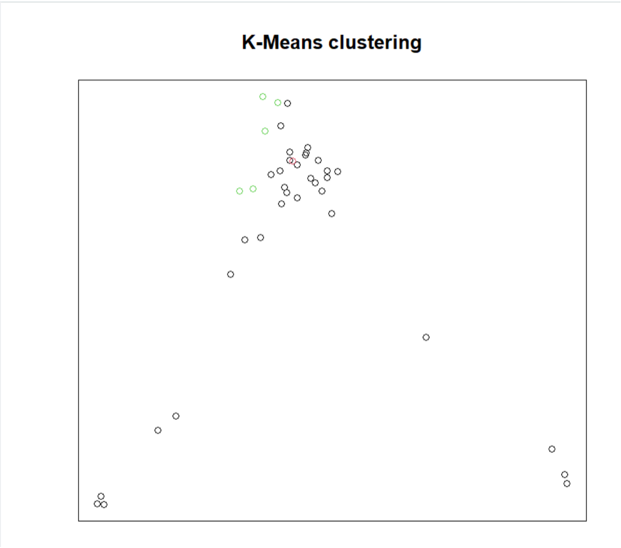
From the terms, we can infer that this document might be associated to topic 1, which is soccer theme topic.

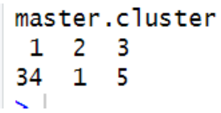
### Task 2: text clustering

For further analysis of the pattern in the text analysis, clustering is performed to identify the relationship distance between each document. There are three clustering performed, namely k-means clustering, hierarchical clustering, and density-based clustering (DBSCAN)

### k-means clustering







```
master.cluster
 1  2  3
34  1  5
```

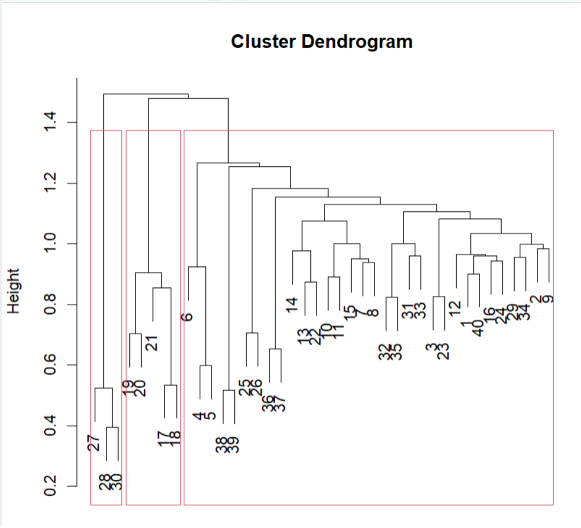
As for k-means clustering, 3 clusters are predetermined (k=3) prior to performing the analysis. This is to see the initial behaviour of the documents.

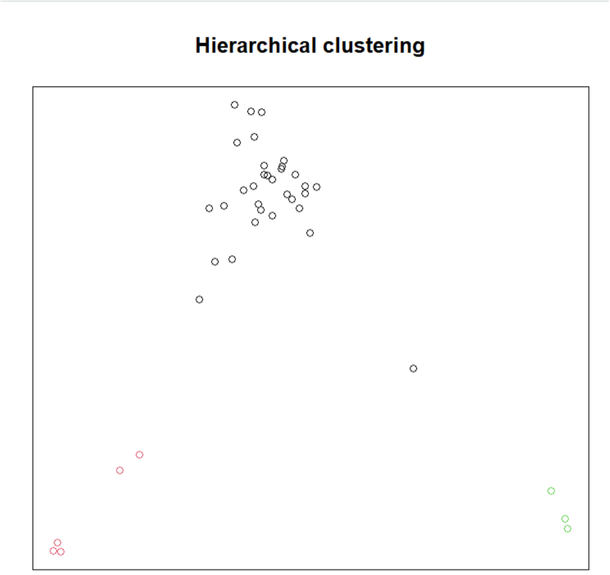
From the left plot, we observe that there has been a large cluster group occupying the plot, while cluster-2 and cluster-3 are very small in size.

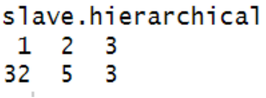
Furthermore, from the right plot, we can also observe that there are 34 documents being classified under cluster-1, while only 1 document is classified under cluster-2 and 5 documents are classified under cluster-3.

From k-means clustering, we can suggest that the most optimum clustering would be k=2 due to the insignificant difference when k=3 cluster is applied.

## Hierarchical clustering





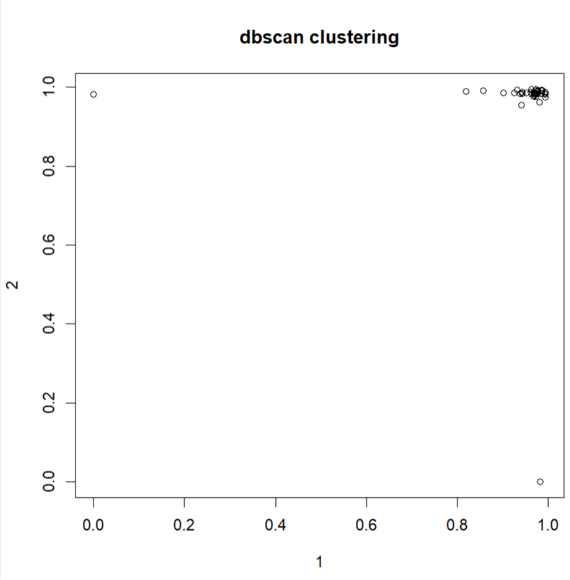


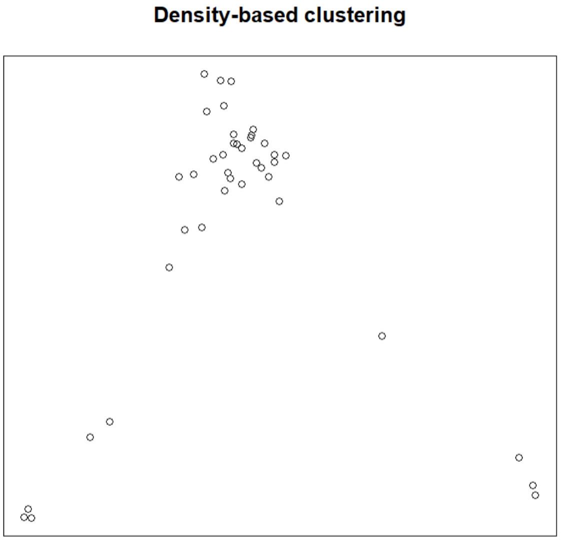
```
slave.hierarchical
 1  2  3
32  5  3
```

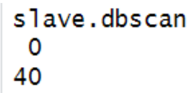
From the hierarchical clustering, out of 40 documents, a total of 32 documents are grouped into cluster-1, 3 documents are grouped into cluster-2, and 5 documents are grouped into cluster-3.

As opposed to k-means clustering, hierarchical clustering shows the structure of similarity between documents prior to grouping into a cluster rather than having a cluster centroid representing the predetermined number of clusters.

## Density-based clustering







```
slave.dbscan
 0
40
```

DBSCAN clustering groups the documents based on their density in the feature space.

From the left plot, we observe that there might be a cluster as the density of the document is situated at coordinate (1.0, 1.0).

However, from the right plot, we observe that all of the documents are classified as noises and do not belong to any cluster group.

The difference observed between the plots is that the left plot is a hierarchical DBSCAN while the right plot is a DBSCAN plot. The hierarchical DBSCAN captures the hierarchical structure from the DBSCAN algorithm. Therefore, it considers not only the dense region but also the intermediate regions and outliers of the data. That is why we observe that the DBSCAN plot identified no specific cluster, but the HDBSCAN includes the similarity approximation feature in plotting the density of the documents.

## Part 2

### Task 3: sentiment analysis

### Dataset background

The dataset used for sentiment analysis is a subset of dataset of Book reviews from Amazon Kindle Store from May 1996 to July 2014.

The dataset contains a total of 12,000 entries. The variables involved in the dataset are:

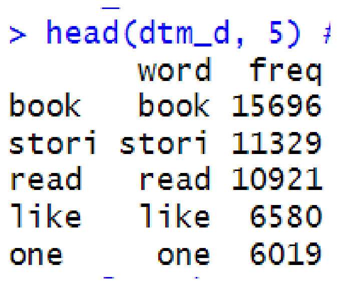
-overall - rating of the product.

-reviewText - text of the review (heading).

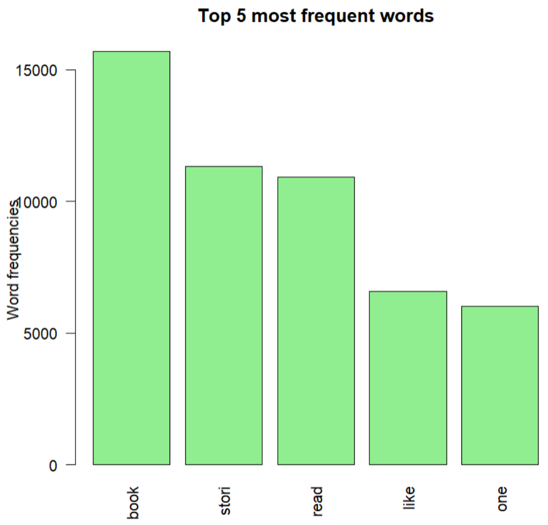
-summary - summary of the review (description).

Lastly, this dataset is taken from Amazon product data, Julian McAuley, UCSD website. http://jmcauley.ucsd.edu/data/amazon/ . License to the data files belong to them.

### Sentiment analysis discussion



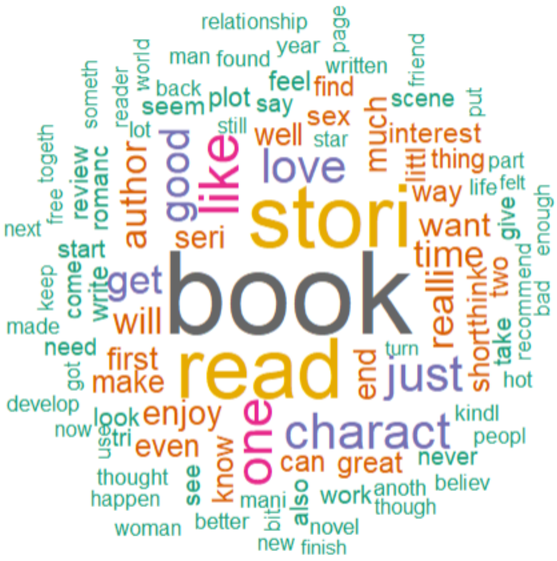
```
> head(dtm_d, 5) #
       word  freq
book   book 15696
stori stori 11329
read   read 10921
like   like  6580
one     one  6019
```



From the dataset, we observe that the word “book” has the highest frequency at 15,696, followed by “stori” and “read” with small difference in frequency for 2nd and 3rd highest at 11,329 and 10,921 respectively. Furthermore, the 4th and 5th most frequent words are also close between each other at 6,580 and 6,019 respectively.

These top most frequent words reflects the kindle dataset as it revolves around words closely related to book reviews alike.

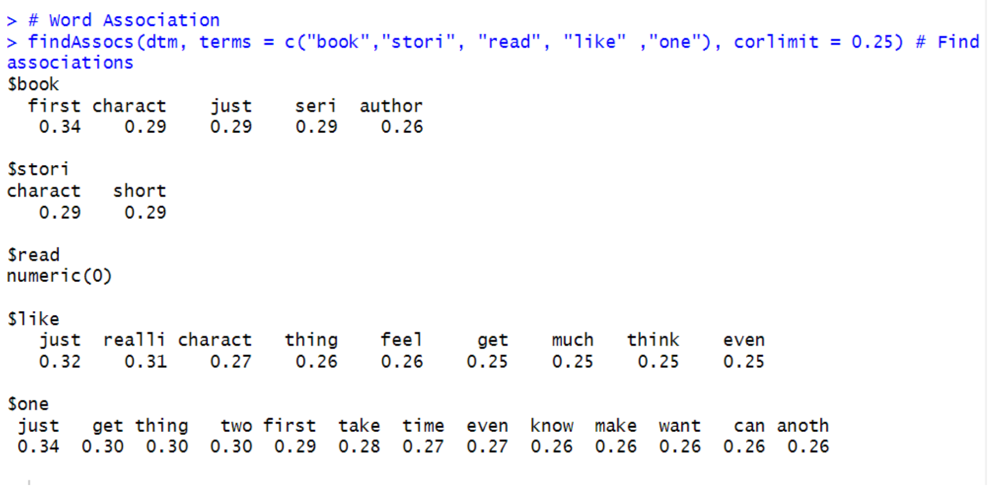
**Wordcloud**



From the wordcloud, it can be observed that the word “book” is being the center of the wordcloud due to its significant frequency in relative to other words.

Hence, we can observe there are multiple words associated to the word “book”. Therefore, we can infer that the words surrounding the word “book” are the possible words when searching the word “book” in the kindle bookstore. Plus, using the associated words surrounding in the wordcloud, it can be included in the strategy to include the books with these associated words so that it would enhance the readers’ experience when using kindle.

## Word association



```
> # Word Association
> findAssocs(dtm, terms = c("book","stori", "read", "like" ,"one"), corlimit = 0.25) # Find
associations
$book
  first charact    just    seri  author 
   0.34    0.29    0.29    0.29    0.26 

$stori
charact   short 
   0.29    0.29 

$read
numeric(0)

$like
   just  realli charact   thing    feel     get    much   think    even 
   0.32    0.31    0.27    0.26    0.26    0.25    0.25    0.25    0.25 

$one
 just   get thing   two first  take  time  even  know  make  want   can anoth 
 0.34  0.30  0.30  0.30  0.29  0.28  0.27  0.27  0.26  0.26  0.26  0.26  0.26 
```

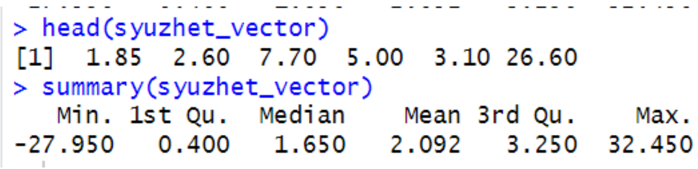
As these words are the top most frequent words in sentiment analysis, the overall associated words for books are similar. This indicates that there are not specific words that are highly associated to book, however the highest word association to the word “book” is “first”.

Similarly to the word “stori”, “like”, and “one”, the correlation coefficient is averagely range around 0.25 to 0.34.

It can be inferred that the words revolving around the top most frequent word do not have any significant correlation, but is well distributed equally in average correlation value. Thus, the words have closely associated between each other. For example, when observing the word “one”, the associated words from the diagram above are equally correlated and all of those associated words do not have a  significant relationship with the word “one” as compared to the other associated words.

This applies to all of the top most frequent words from the kindle review dataset.

### Sentiment scores: Syuzhet analysis



```
> head(syuzhet_vector)
[1]  1.85  2.60  7.70  5.00  3.10 26.60
> summary(syuzhet_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-27.950   0.400   1.650   2.092   3.250  32.450 
```

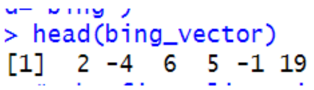
From the Syuzhet vector, we can observe the sentiment scores for the first 6 lines.

Syuzhet scoring ranges between -1 to 1, where -1 indicates a negative word score and 1 indicates a positive word score.

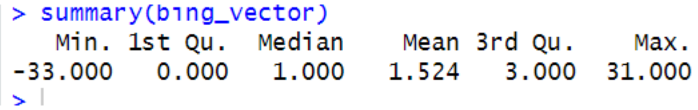
In overall, the first 6 lines in the dataset has a positive score. Although the score range is between -1 to 1, the exceeded value is the total score value in the line, meaning that the words in line 1 has a total sentiment score of 1.85, and line 2 has a total sentiment score of 2.60 and so on.

Furthermore, in average, the dataset has a positive score at 2.092, and the score is not significantly different from the mean at a score of 1.650.

### Sentiment scores: bing analysis



```
> head(bing_vector)
[1]  2 -4  6  5 -1 19
```



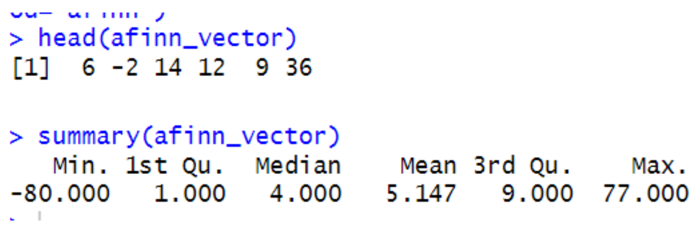
```
> summary(bing_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-33.000   0.000   1.000   1.524   3.000  31.000 
> 
```

As for Bing sentiment analysis, we can observe that there are a mixture of scores for the dataset. Line 1, 3, 4, and 6 have positive scores while line 2 and 5 have negative scores.

As compared to Syuzhet, bing provides a binary sentiment classification, and the scores are derived from a set of manually labeled words from social media. Whereas for Syuzhet, the scores are based on a collection of novels, and other literary narratives.

Furthermore, the scoring in Bing is discreet and does not consider decimal values, this explains the high positive score for line 6.

### Sentiment scores: afinn analysis



```
> head(afinn_vector)
[1]  6 -2 14 12  9 36

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

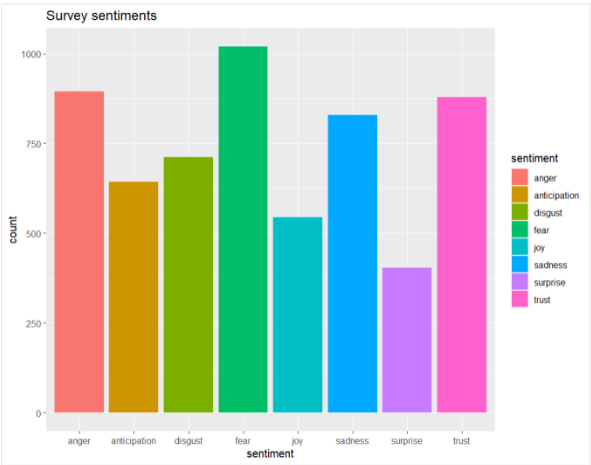
As for AFINN analysis, we can observe only line 2 has a total negative score, and the line with highest positive score among the six line is the 6th line, scoring at 36. Similarly to Bing and Syuzhet, all three analysis shows that line 6 has the highest positive score.

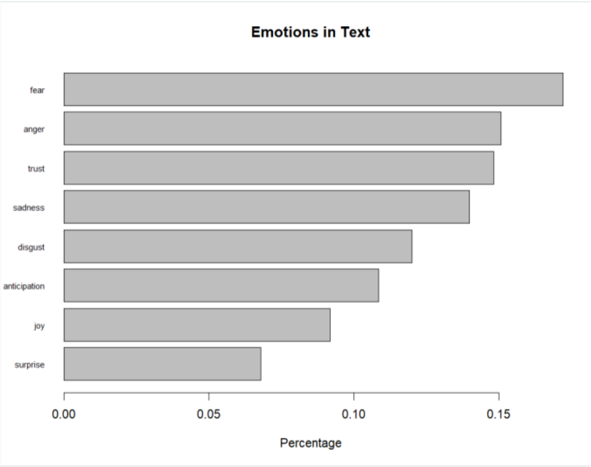
Thus, we can infer that line 6 is the most positive among the other lines observed.

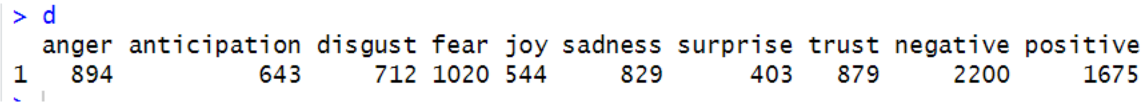
The variance for mean and median is relatively large but the difference is not too significant to consider that the dataset is not symmetrical.

### NRC sentiment analysis

As compared to Syuzhet, Bing and AFINN sentiment analysis, National Research Council (NRC) sentiment analysis provides a list of words manually annotated with sentiment categories. The words annotated includes, but not limited to, positive, negative, anger, joy and others.







```
> d
  anger anticipation disgust fear joy sadness surprise trust negative positive
1   894          643     712 1020 544     829      403   879     2200     1675
```

From the plot, we infer that the dataset has a highest score in fear at 1,020, and thus shows the negative score at 2,200. This is different as compared to the above three sentiment analysis as Syuzhet, Bing and AFINN analysis each shows an average of positive review in the kindle dataset.

The positive scores consist of only a total of 1,675, and the highest positive sentiment category is trust with only a score of 879.

However, NRC sentiment analysis focused primarily on individual words and assigning it to their respective categories, it does not consider the overall context of the text as compared to Bing, Syuzhet and AFINN. Bing is commonly used for social media context sentiment analysis, while Syuzhet considers more on the literary context related sentiment analysis, and AFINN is commonly used in the context of social media and informal text.

Therefore, the contextual consideration is important when performing sentiment analysis, and in this dataset, the more suitable analysis would either be Bing and AFINN as the kindle dataset is closely related to social media context.

Thus, in reference to Bing and AFINN, we can consider that this dataset provides a positive review in average.

## Conclusion

As a whole, the text analysis in regard to the text documents shows that optimally, it can be considered to be clustered at most 3 clusters for hierarchical clustering based on the data structure. However, it is reasonable to infer that 2 clusters are recommended under k-means clustering due to the calculated distance is not significantly large when observing k=3 clusters. Plus, we should also take into account that the behaviour of the text documents are not densely distributed in the data point under DBSCAN clustering, meaning that DBSCAN does not consider all of the text document to be classed into a single cluster as they are not densely related to one another.

As for sentiment analysis, contextual consideration is important when performing the analysis of the dataset, and the limitation of each sentiment analysis techniques should be considered to minimize any analysis error. Therefore, for kindle review dataset, the review are related to social media context, thus, AFINN and Bing sentiment analysis is best to represent and decide the positive and negativity behaviour of the kindle review dataset.