STQD6114 Task 3

PANZHANGYU P136922

Data: Starbucks Reviews Dataset (kaggle.com)

This dataset contains a comprehensive collection of consumer reviews and ratings for Starbucks, a renowned coffeehouse chain. The data was collected through web scraping and includes textual reviews, star ratings, location information, and image links from multiple pages on the ConsumerAffairs website. It offers valuable insights into customer sentiment and feedback about Starbucks locations.

About this file：

File Name: "reviews_data.csv"

File Format: CSV (Comma-Separated Values)

File Size: Approximately 10 MB

Data Columns:

name: Name of the reviewer (text)

location: Location or city of the reviewer (text)

Date: Date when the review was posted (text)

Rating: Star rating given by the reviewer (1 to 5)

Review: Textual content of the review (text)

= Image_Links: Links to associated images (text)

Missing Data: Some reviews may have missing values, indicated as "NaN."

Encoding: UTF-8

Data Delimiters: Comma-separated values

I will use the data from the Review column.

First, I preprocessed the data, including cleaning the data, removing meaningless words, and stemming.

I calculated sentiment scores for the reviews using four different sentiment lexicons (syuzhet, bing, afinn, and nrc). Here are the results for each method:

Syuzhet：

```
> head(syuzhet_vector, 10)
 [1]  3.65  6.00  1.05  3.65  6.25 -0.35 -1.25  3.10 -1.50 -2.40
```

Bing:

```
> head(bing_vector, 10)
 [1]  4  8  3  5 10  0 -2  2 -2 -5
```

Afinn:

```
> head(afinn_vector, 10)
 [1]  4 11  3 14 14 -6 -2  6 -4 -5
```

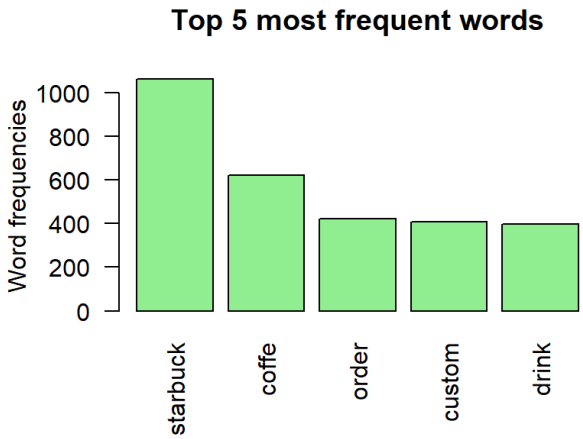
Nrc:

```
> head(nrc_vector, 10)
 [1]  2  8 -2  1  2  0 -2  3 -2 -4
```

I extracted the most common words by building a word frequency matrix. Here are the top five most common words:

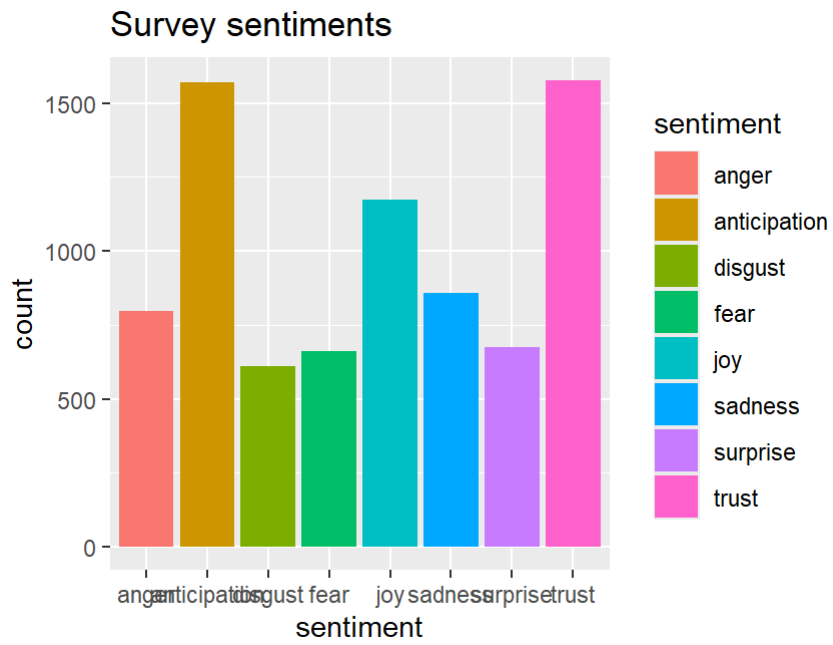
```
> head(dtm_d, 5)
           word freq
starbuck starbuck 1062
coffe      coffe  620
order      order  423
custom    custom  409
drink      drink  398
```

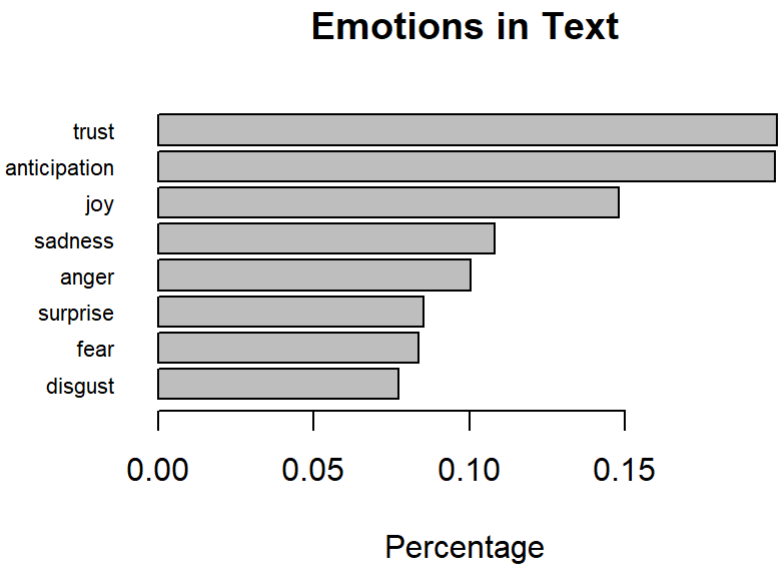
I used word clouds and bar charts to show the word frequency distribution.



Example of sentiment classification results:

```
> head(nrc_sentiment, 10)
   anger anticipation disgust fear joy sadness surprise trust negative
 positive
1      0            0       0    0   1       0        1     1        0
 2
2      1            5       2    2   4       1        2     3        1
 9
3      1            0       1    1   1       2        1     1        3
 1
4      0            0       1    0   2       0        0     1        2
 3
5      2            1       0    0   0       0        0     2        2
 4
6      0            0       0    0   0       0        0     0        1
 1
7      0            1       2    1   0       1        0     1        3
 1
8      1            2       1    0   2       2        2     3        2
 5
9      1            0       0    0   0       1        0     0        2
 0
10     3            3       3    3   2       5        2     2        7
 3
```


Survey sentiments
count
1500
1000
500
0
anger
anticipation
disgust
fear
joy
sadness
surprise
trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
sadness
surprise
trust


Emotions in Text
trust
anticipation
joy
sadness
anger
surprise
fear
disgust
0.00
0.05
0.10
0.15
Percentage

Through sentiment analysis of the review data, we found that most reviews are positive, with the main sentiments being Trust and Anticipation. Common positive words include "great," "product," "love," "good," and "excellent." Common sentiments in negative reviews include Anger, Fear, and Disgust. Our analysis results can help companies understand users' overall sentiment toward their products and make corresponding adjustments in future marketing strategies.

In the future, we can consider the following points for further in-depth analysis:

Combining the time information of the comments to analyze the trend of sentiment changes.

Performing topic analysis on the comments to understand the specific issues that users are concerned about.

Combining the sentiment analysis results with user ratings to further verify the accuracy of the sentiment analysis.