Amazon review topic and sentiment modeling

Qianhui Yang

10/6/2020

Introduction

The purpose of this project is to discover if there is strong sentimental difference between the positive and negative reviews. And if we can use machine learning to let the computer to detect the topic of reviews.

We found there are significant sentimental difference between the negative and positive review, and the machine help us to detect the topic 4 in 5 time correct. Not bad!

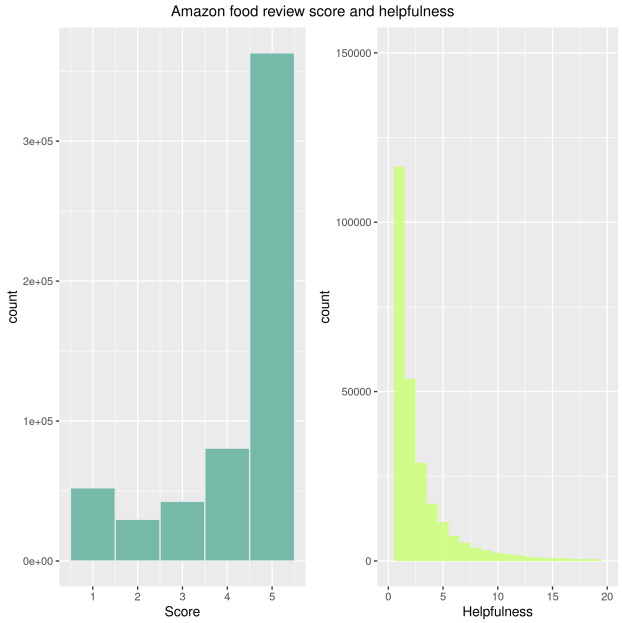
Dataset

The data set comes from Kaggle: https://www.kaggle.com/snap/amazon-fine-food-reviews This dataset contains Amazon fine food review, including around 500,000 reviews from Amazon Oct 1999 - Oct 2012

and UCSD: http://jmcauley.ucsd.edu/data/amazon/ The dataset contains product review and metadata from Amazon May 1996 - July 2014.

Data Overview

The food review overview shows the distribution of food review score and helpfulness score. The food review have mainly positive review but less help in general.



Review Length

Top 20 review length and summary. It is very surprising that so many summary has length over 1000

Top 20 amazon food reviews with Highest Word Count

id	summary	score	num_words
290808	an okay filtered municipal tap water with slight "chalky" aftertaste but the misleading health claims and hype are quackery	3	3483
455394	an okay filtered municipal tap water with slight "chalky" aftertaste but the misleading health claims and hype are quackery	3	3483
496754	an okay filtered municipal tap water with slight "chalky" aftertaste but the misleading health claims and hype are quackery	3	3483
68701	searching for a pet appetite enhancer?	5	2628
541159	not funny / update: second gift basket i received	1	2232
175973	a fine new option for a low calorie sweetner	4	2137
331319	disagree with negative reviews	5	2130
346184	tea antioxidants	5	2120
407776	the real black pearl: an adventure tale	4	1946
175185	most dog foods are not human grade. there simply is no better dog food. you will see it in your dog's coat.	5	1924
269917	sets the bar in dog food most are not human grade and most are highly processed	5	1924
209089	do not follow the directions	5	1910
97611	saving whisper's life	5	1866
248852	weight loss benefits of green tea	5	1810
539894	weight loss benefits of green tea	5	1810
483160	family saga of the indoor kitties versus the feral felines	4	1797
10005	constipation	1	1763
253601	spookylicious pop tarts: a cautionary tale.	5	1684
247127	how to grow dragonfruit	5	1675
541047	really disappointed	2	1666
276020	works great you have to follow the instructions read on	5	1650
230036	green tea ingredient slows breast cancer antioxidant in green tea may stop breast cancer growth	5	1641
137989	some education about vegan cats	1	1637

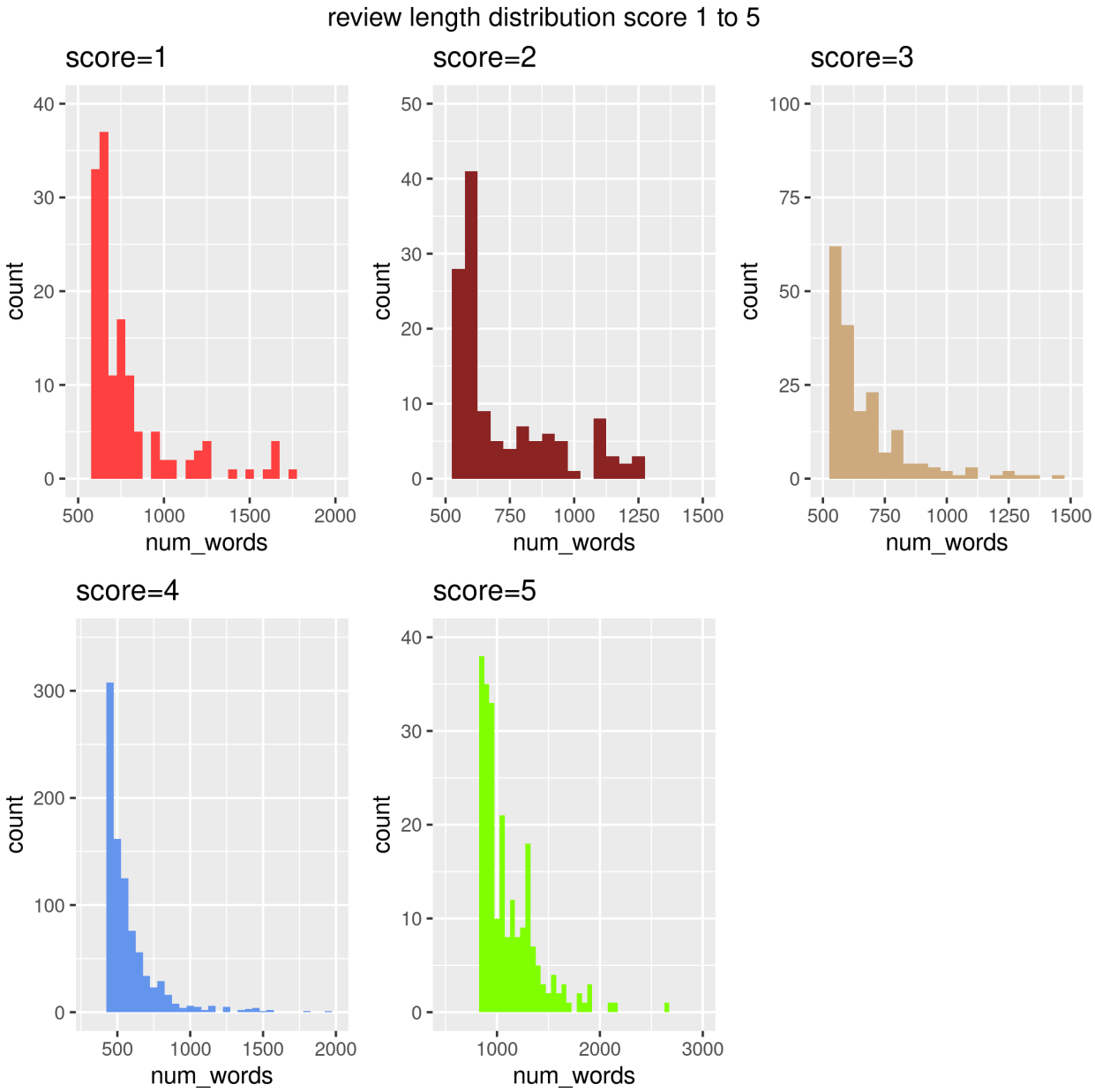
Least 20 Review Length and Summary from Amazon food review. The number 61 comes form the review limitation.

Last 20 amazon food reviews with Highest Word Count

id	summary	score	num_words
338140	yummy	5	61
338177	"dust to dust" no actual leaves in this tea	1	61
338256	pretty good taste; goes a long way	4	61
338436	5 for flavor 2 for price	5	61
338503	salba	5	61
338625	this is what converted me	5	61
338990	this product works	5	61
339145	7.99 shipping???	3	61
339168	grrr	2	61
339181	these are good but were broken.	4	61
339266	sorry this drink is really bad	1	61
339291	tasty light alternative to a real cocktail	4	61
339461	ugh it does not taste good	1	61
339517	it must be the box?	3	61
339890	french candy	5	61
340015	kids love them	5	61
340049	great quality	4	61
340372	i love lipton iced tea drink mix but....	5	61
340683	sweet deal	5	61
340726	ok?? : nutritional info.? ; salt/sodium content???	3	61
340747	love these super cookies chocolate coconut without the guilt	5	61

Review length distribution by score

The higher the score, the longer the review could be.



Do does the length relate to the score in the statistical sense?We conduct ANOVA here and found significant association between these two (p-value: $< 2.2e\text{-}16$)!

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
full_word_count<-read_csv("derived_data/full_word_count.csv")
```

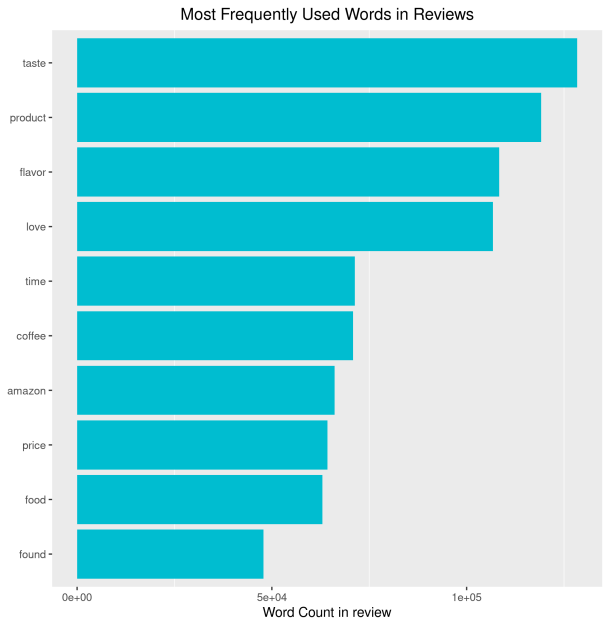
```
##
## -- Column specification --------------------------------------------------------
## cols(
##   id = col_double(),
##   summary = col_character(),
##   score = col_double(),
##   num_words = col_double()
## )
```

```
anova<-lm(score~num_words,data = full_word_count)
summary(anova)
```

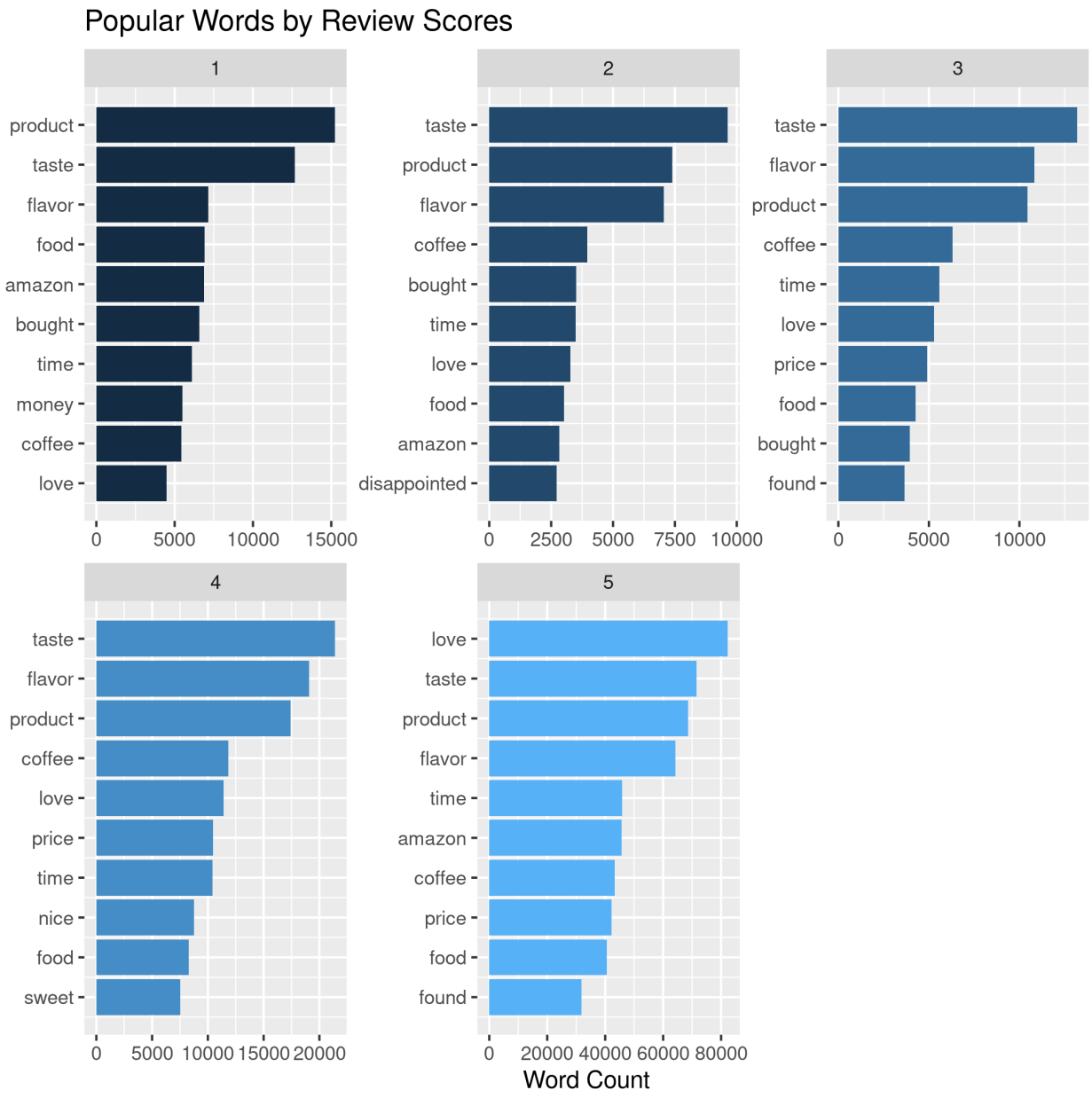
```
##
## Call:
## lm(formula = score ~ num_words, data = full_word_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2808 -0.2427  0.7524  0.7966  3.9419
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.284e+00  2.453e-03  1746.3   <2e-16 ***
## num_words   -1.228e-03  2.106e-05   -58.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.307 on 568452 degrees of freedom
## Multiple R-squared:  0.005944,   Adjusted R-squared:  0.005942
## F-statistic:  3399 on 1 and 568452 DF,  p-value: < 2.2e-16
```

Popular Words in the reviews

The most 10 popular words in the all reviews are below:we can found many words used to describe the food such as flavor,taste!



By scores,the populor words are not quiet different from each other. The word :taste,flavor, amazon, food are very common in all the review score.



Sentiment analysis

Here start the fun staff, the question is: can we tell the sentimental difference between the negative and positive review?

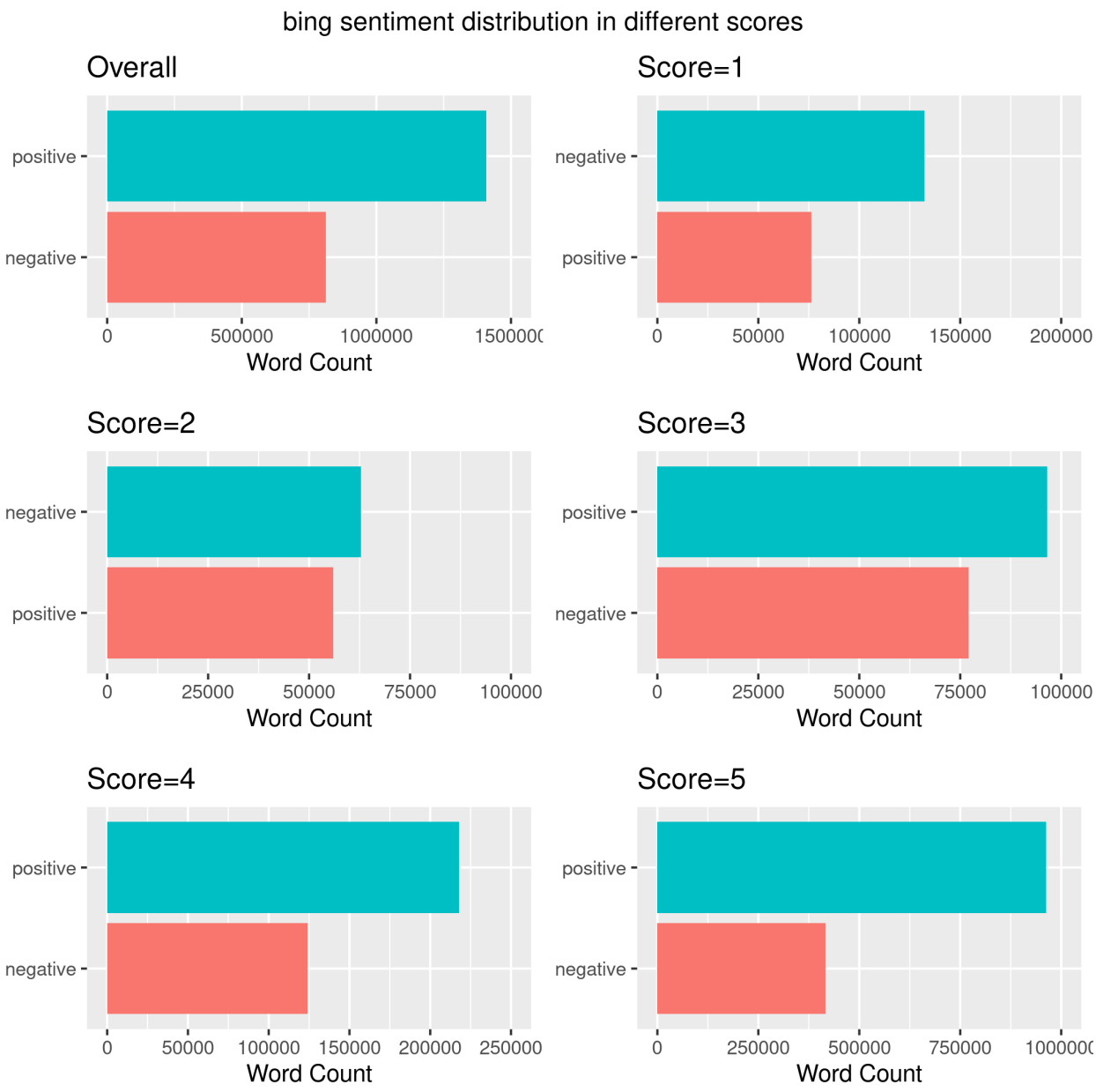
R has three sentimental database: Bing,nrc and afinn, they are very different on their own way of defining sentiment. The Bings dataset split the words into negative and positive category. The NRC datase split the words into more detailes like angry, upset, fear... The afinn dataset have a numeric database to decribe the positiveness and negativeness.

Here are the overlap between our data and three database we mentioned.

Reviews Found In Lexicons

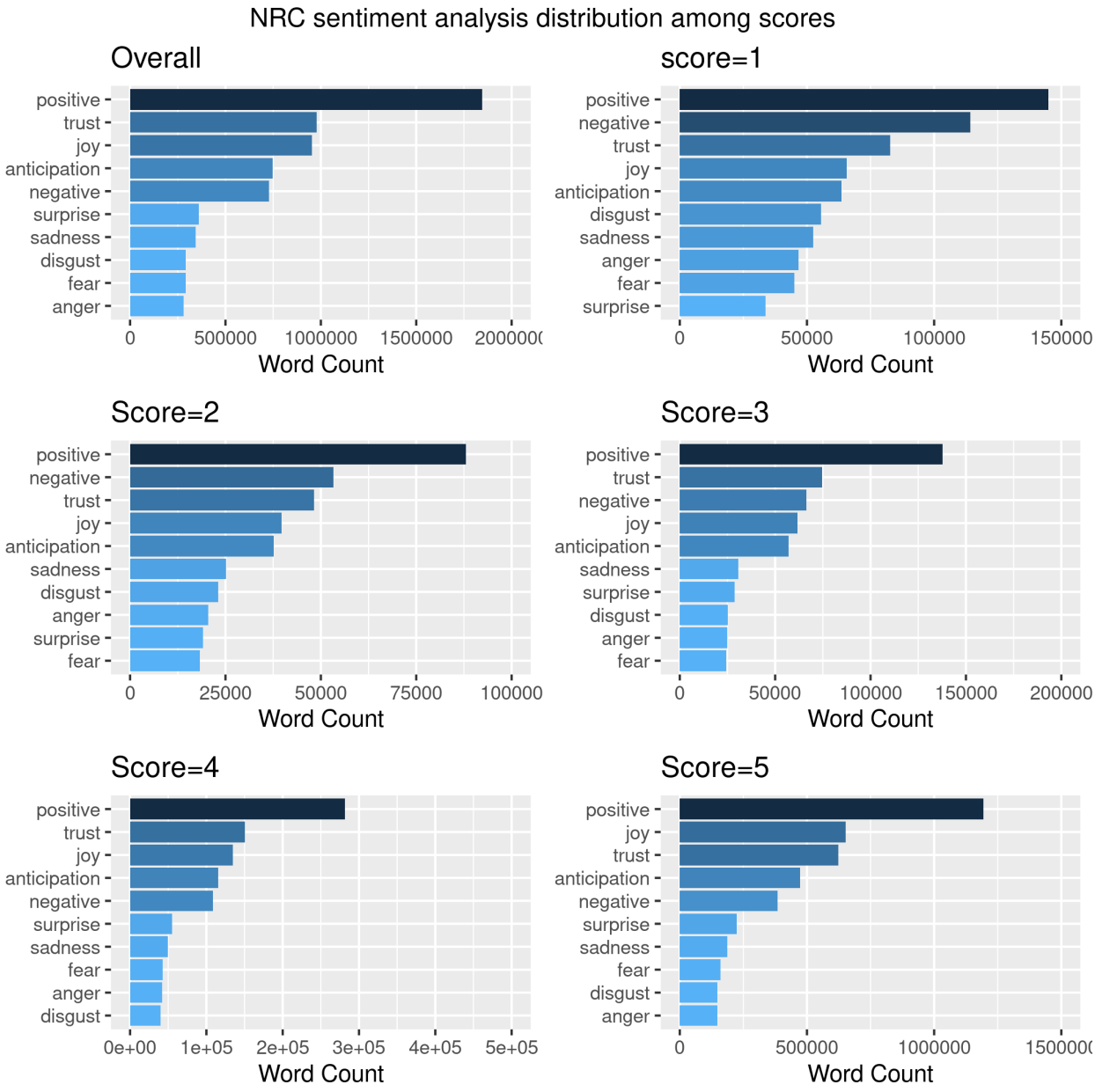
lexicon	lex_match_words	words_in_review	match_ratio
afinn	2087	131929	0.0158191
bing	5014	131929	0.0380053
nrc	5478	131929	0.0415223

The NRC and bing database has more than five thousand overlap and afinn has over 2000 overlap. From the Bing sentiment database, we get the following result.



The result shows that positive review has more positive words than negative words. While the negative review has more negative words than positive.

From the nrc database analysis, we get:



It is little strange to see positive sentiment in the negative report.But we can see many negative sentiment are higher in the lower score compared to the score=5.

Topic analysis

This last part are a little train to see if the computer can tell the topic of reviews.Adding to the origin food review, we use the review from Beauty, Outdoor, movieTV, VideoGame. The analysis we use is called LDA and k-means, they are widely used in the natural language processing. LDA is a short for latent Dirichlet allocation, if observations are words collected into documents, it posits each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

By LDA and k-means topic modeling, top five topic we can distinguish the 4 out of 5!

k-means is a clustering method used to partition observations into k clusters, it is hard for us to tell the accuracy.

Future interests

In the future, we are interested in trying other sentimental analysis dictionary and add some of our customized dictionary for the food reviews. In addition, we are interested in trying to make a overlap of the topic and sentiment analysis to develop a recommendation system for the customers by their personal preference topic.