Text Analytics and Natural Language Processing - DAT-5317 – FMBAN2

AssigmentsA3: Business Insight Report

Diana Aycachi Mamani

Diana Aycachi

Business Insight Report

Peruvian Restaurants in the Bay Area

PART I

I arrived in the US in 2019, and since I arrived here, I realized that many Peruvian Restaurants have started to open, and most of them with good acceptance from the customers, I want to know which are the main success factors for Peruvian restaurants here in the Bay Area.

1. Collection of data

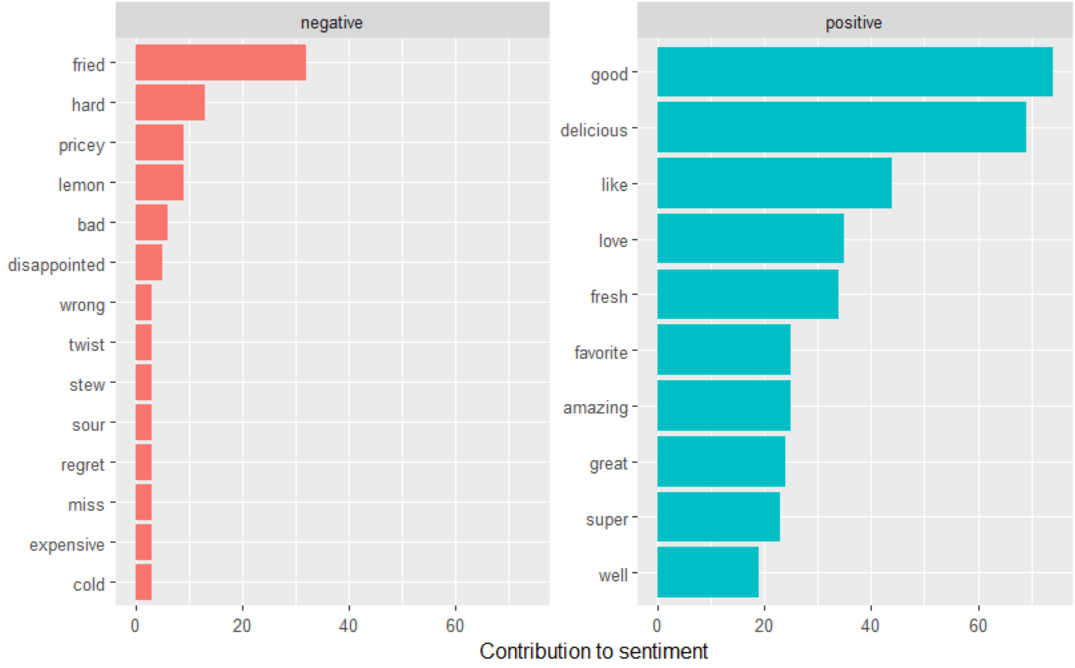
I did web scraping on the reviews from Yelp, with the help of "Selector Gadget" (Which I just installed in the Chrome Extension) and created a data frame of one single column in R Studio. For these I used the library rvest.

The restaurants that I analyzed were Jora Peruvian food located in San Jose, MR Kano located in Santa Clara, and Emelina Restaurant located in San Carlos. I have chosen these restaurants because they have more than 10 review pages in Yelp and at least 4 stars.

2. Sentiment Analysis

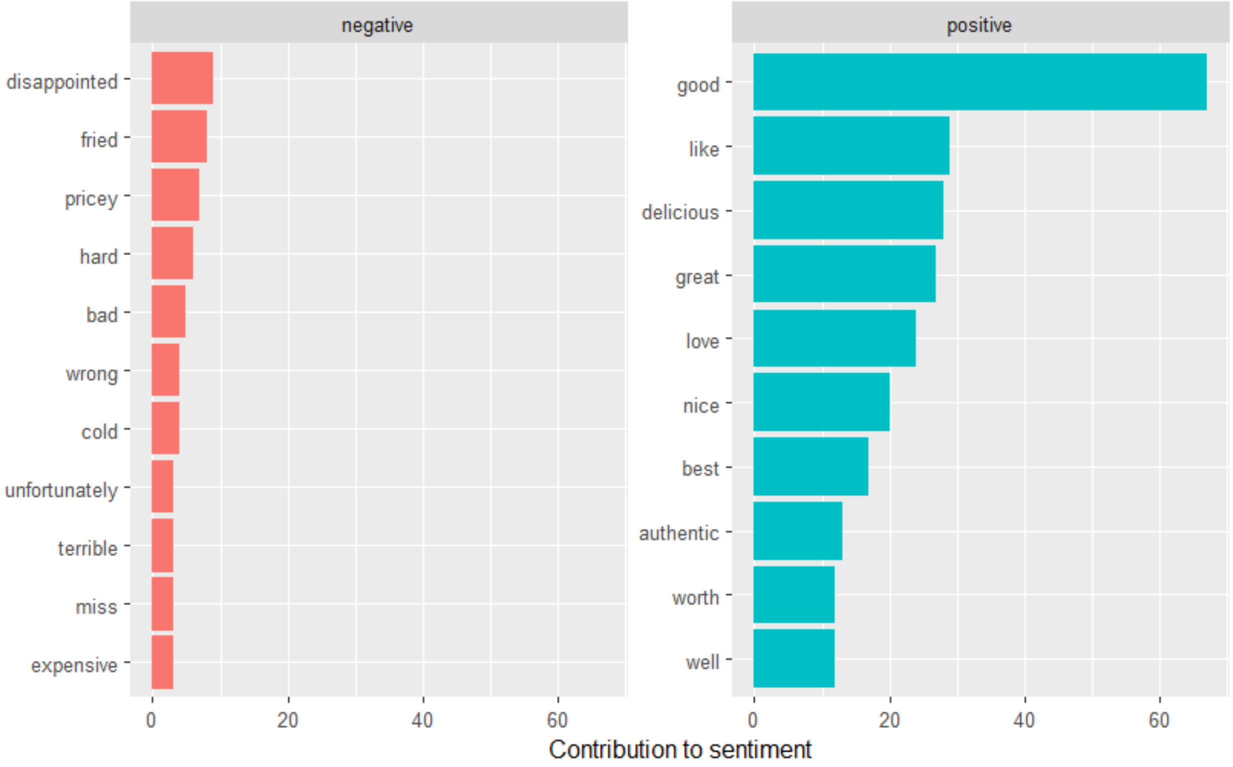
2.1 Jora Peruvian Food

Using the library Bing which will give us a positive or negative impression from the clients we can say about Jora that the food is expensive. We can not take the other words as negative because they are not necessarily negative under the restaurant context.
About positive sentiment we can observe that the customers consider the food from Joras as delicious, made with love, good and fresh.



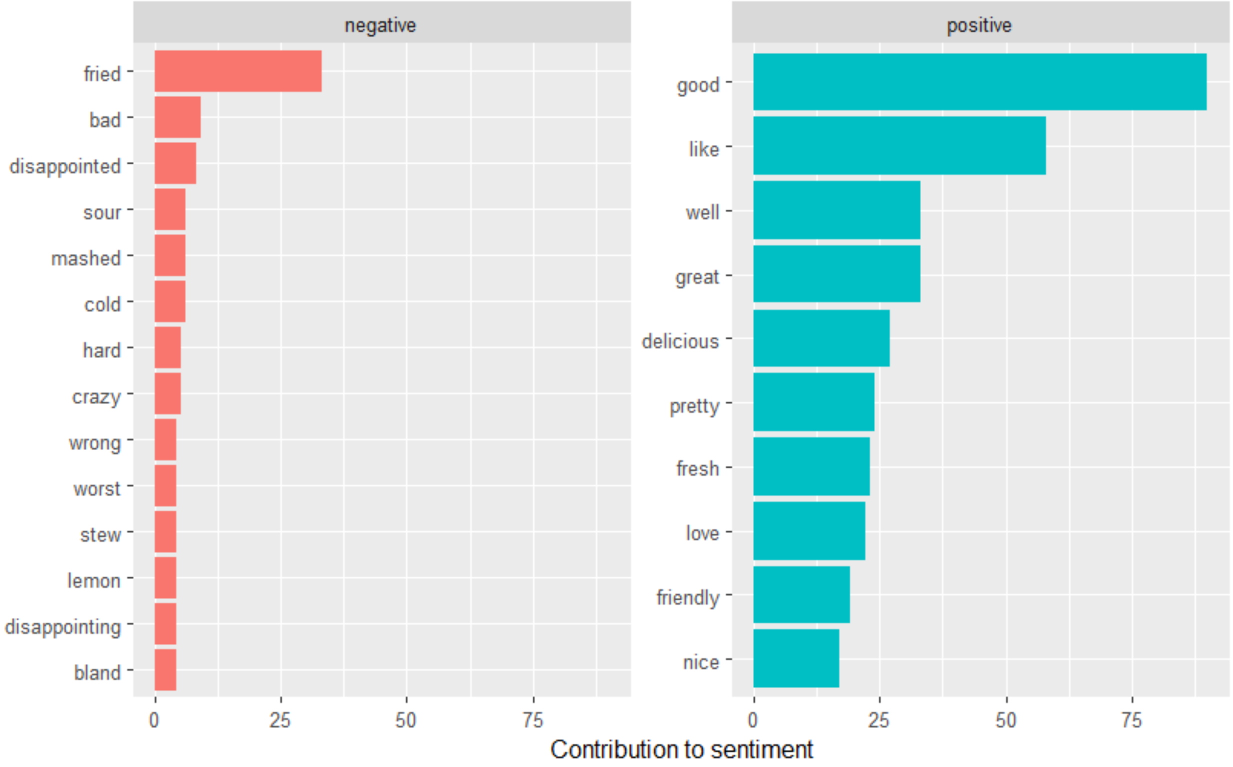
2.2 MR Kano Restaurant

About Mr Kano Restaurant we can observe some similarities with Jora, in the graphic bar for the negative sentiment we can observe the same words: fried, pricey, hard. And for positive sentiment we can observe also another similar positive tokens as: good, delicious, love, like



2.3 Emelina Restaurant

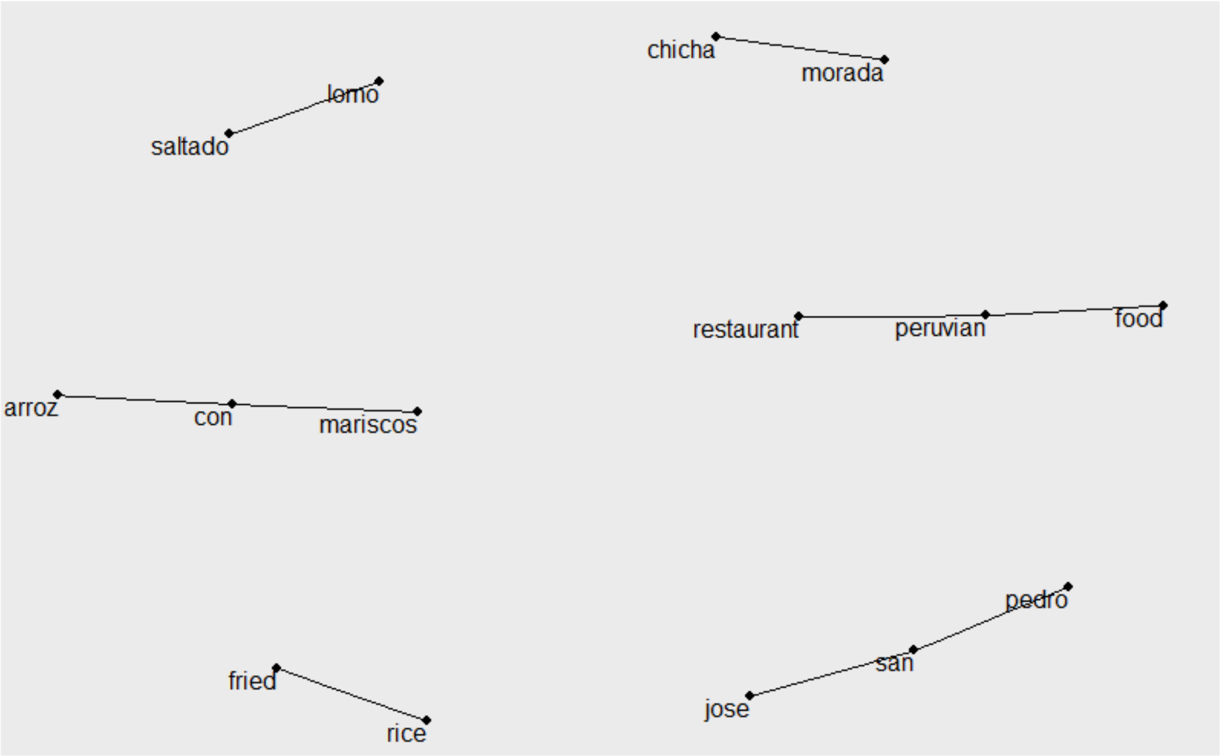
For Emelina Restaurant we can observe as a negative token the fried token again, hard. And as a positive sentiment we can observe the tokens as good, like, well, great, fresh and love.



Comparing the three sentiment analysis based on the Bing library we can say that peruvian food in general had positive attributes as fresh, delicious, made with love, and good. On the other hand, we can say about the peruvian that it is expensive.

3. N-Grams Analysis

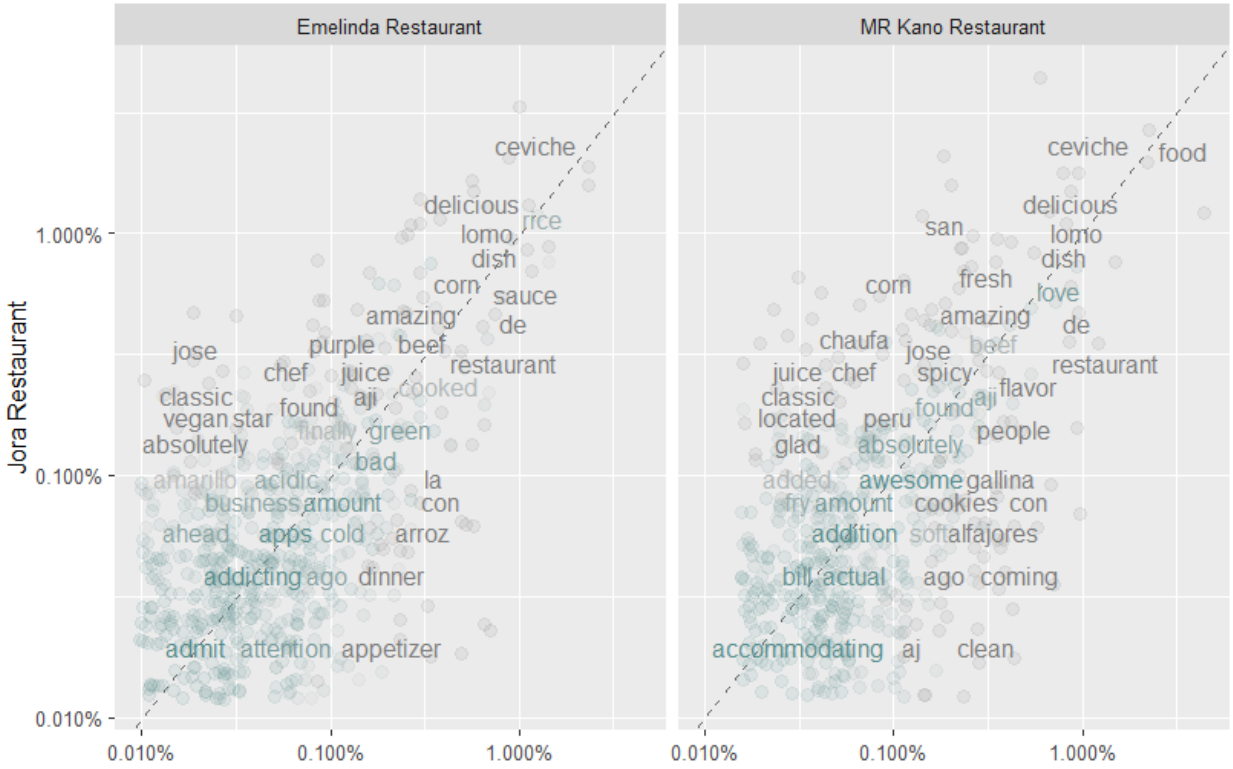
For the N-Grams analysis I have merged the 3 data frames from the 3 restaurants, as a result we have 6 sets of bigrams, which give us information about three main products “lomo saltado”, “arroz con mariscos” and “chicha morada”. This information can help us to open a new peruvian restaurant or guide us to create new dishes with a variation in these three products.



4. Correlograms

Doing an analysis of the three texts we can observe that Emelina Restaurant and Jora Peruvian Food have in common “apps”, “addicting”, “attention”, “admit”, “acidic”. Jora, in comparison with Emelina, offers juice and vegan products. And Emelina in comparison to Jora offers “dinners” and appetizer products.

Now comparing Jora and Mr Kano, some similarities that we can find is that products or flavors are awesome, made with love. On the other hand, some differences about Jora respect to MR Kano are about juice, fresh, chaufa. And about MR Kano respect to Jora some tokens are clean, alfajores and gallina.



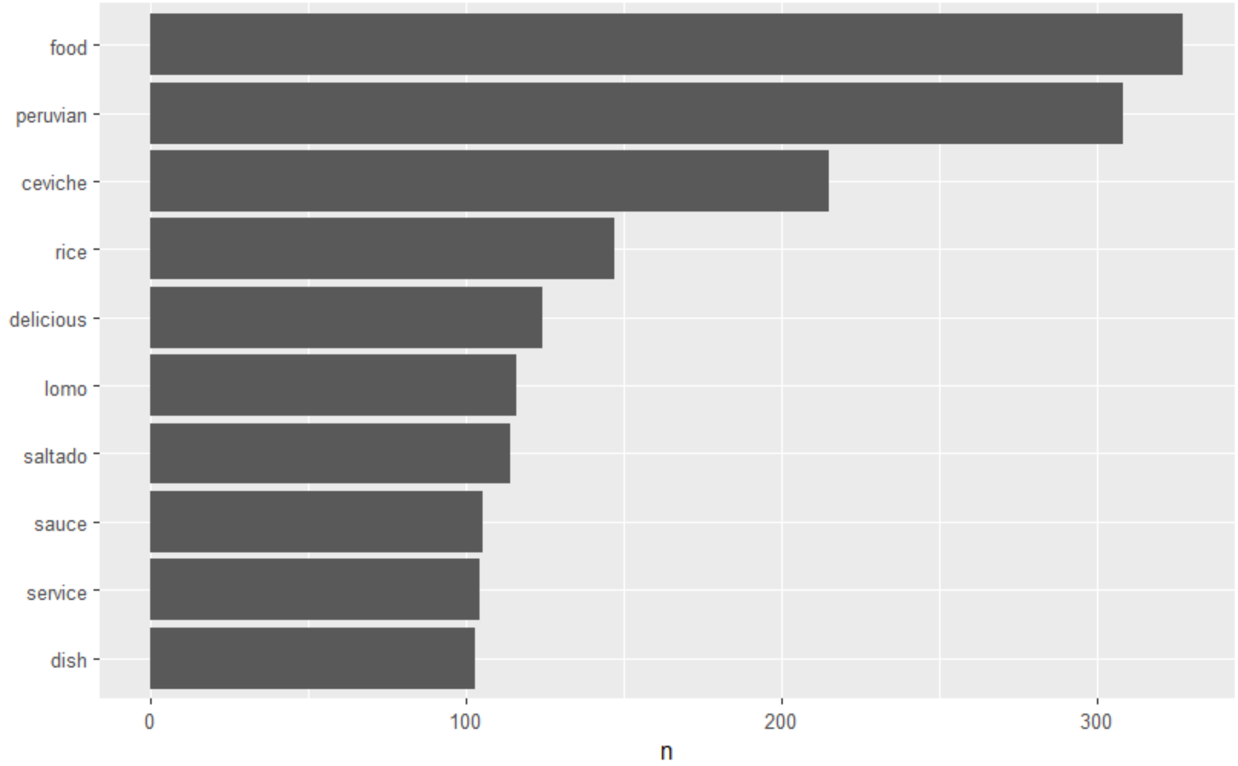
5. Correlation Test

Jora - MR Kano	Jora- Emelina
Pearson's product-moment correlation data: proportion and Jora Restaurant t = 28.862, df = 505, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.7537366 0.8197841 sample estimates: cor	Pearson's product-moment correlation data: proportion and Jora Restaurant t = 40.551, df = 596, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.8338178 0.8766795 sample estimates: cor 0.8567207

At 95 % confidence we can observe that there is a high correlation between the data of these three restaurants because it is over 0.70, which means the products and service offered are very similar.

6. Token Frequency with no stop words

Doing an analysis of the frequency of the three restaurants and filtering the tokens equal or over 100 repetition we can observe the words “ceviche”, “rice”, “delicious”,”lomo”, “saltado”,”sauce”, and “service” as more frequent.



7. Conclusion

After completing the analysis, and combining the different data frame analysis, N grams size 2, Sentiment Analysis “Bing”, and Correlograms we can say about peruvian food that although is expensive is worth because it is delicious, some of the most popular dishes are the ‘Ceviche’, ‘Lomo Saltado’, ‘Fried Rice”, “Chicha morada”. About the service we can say that it is good since the word “love” appears many times. Also, it might be something related with the sauces because it is repeated at least 100 times. And, in general lines we can say that the Peruvian Food is delicious but perceived as expensive.

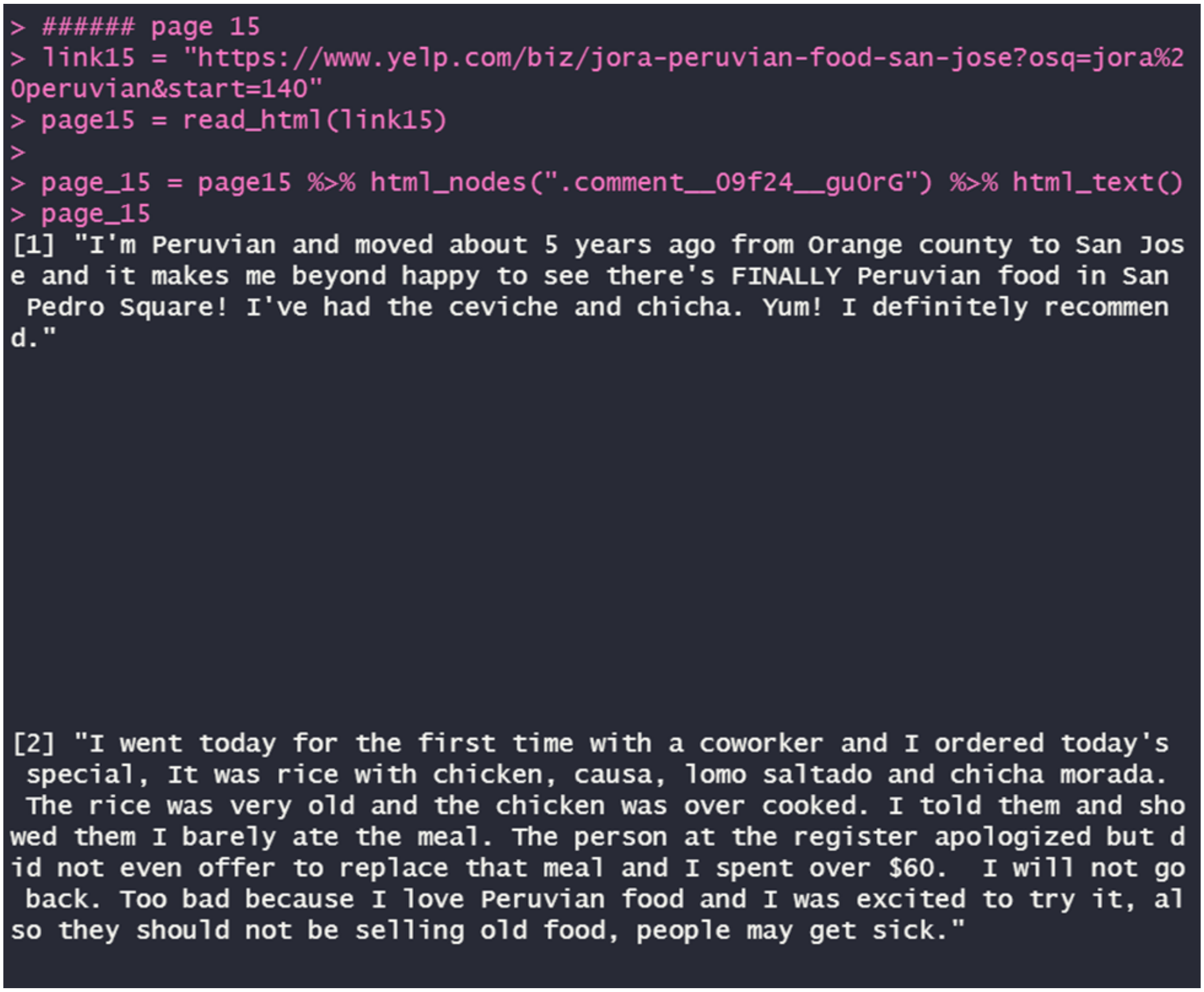
PART II

1.1 Collecting the data for Jora

```
library(rvest)
library(dplyr)

############################################################################
####
############################################################################
####

###############    JORA PERUVIAN FOOD

############################################################################
####
############################################################################
####

link1 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian"
page1 = read_html(link1)

page_1 = page1 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_1


####### page 2
link2 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=10"
page2 = read_html(link2)

page_2 = page2 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_2

###### page 3
link3 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=20"
page3 = read_html(link3)

page_3 = page3 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_3
```

```
####### page 4
link4 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=30"
page4 = read_html(link4)

page_4 = page4 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_4

###### page 5
link5 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=40"
page5 = read_html(link3)

page_5 = page5 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_5

####### page 6
link6 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=50"
page6 = read_html(link6)

page_6 = page6 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_6

###### page 7
link7 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=60"
page7 = read_html(link7)

page_7 = page7 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_7

####### page 8
link8 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=70"
page8 = read_html(link8)

page_8 = page8 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_8

###### page 9
link9 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=80"
page9 = read_html(link9)

page_9 = page9 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()

####### page 10
link10 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%20peruvian&start=90"
page10 = read_html(link10)

page_10 = page10 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_10

###### page 11
```

```
link11 = "https://www.yelp.com/biz/jora-peruvian-food-san-
jose?osq=jora%20peruvian&start=100"
page11= read_html(link11)

page_11 = page11 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_11

####### page 12
link12 = "https://www.yelp.com/biz/jora-peruvian-food-san-
jose?osq=jora%20peruvian&start=110"
page12= read_html(link12)

page_12 = page12 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_12

###### page 13
link13 = "https://www.yelp.com/biz/jora-peruvian-food-san-
jose?osq=jora%20peruvian&start=120"
page13 = read_html(link13)

page_13 = page13 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_13

####### page 14
link14 = "https://www.yelp.com/biz/jora-peruvian-food-san-
jose?osq=jora%20peruvian&start=130"
page14 = read_html(link14)

page_14 = page14 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_14

###### page 15
link15 = "https://www.yelp.com/biz/jora-peruvian-food-san-
jose?osq=jora%20peruvian&start=140"
page15 = read_html(link15)

page_15 = page15 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
page_15
```



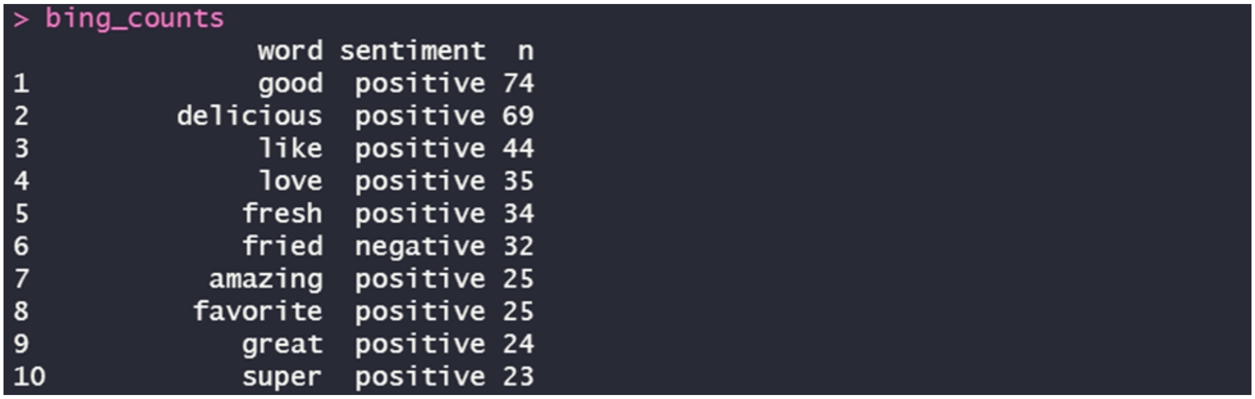
```
> ###### page 15
> link15 = "https://www.yelp.com/biz/jora-peruvian-food-san-jose?osq=jora%2
0peruvian&start=140"
> page15 = read_html(link15)
>
> page_15 = page15 %>% html_nodes(".comment__09f24__gu0rG") %>% html_text()
> page_15
[1] "I'm Peruvian and moved about 5 years ago from Orange county to San Jos
e and it makes me beyond happy to see there's FINALLY Peruvian food in San
 Pedro Square! I've had the ceviche and chicha. Yum! I definitely recommen
d."











[2] "I went today for the first time with a coworker and I ordered today's
 special, It was rice with chicken, causa, lomo saltado and chicha morada.
 The rice was very old and the chicken was over cooked. I told them and sho
wed them I barely ate the meal. The person at the register apologized but d
id not even offer to replace that meal and I spent over $60.  I will not go
 back. Too bad because I love Peruvian food and I was excited to try it, al
so they should not be selling old food, people may get sick."
```

1.2 Structuring the data and tokenizing it

```
library(tidytext)

list_pages =
c(page_1,page_2,page_3,page_5,page_6,page_7,page_8,page_9,page_10,page_11,page_12,
page_13,page_14,page_15)
df <- as.data.frame(list_pages)

colnames(df)[1] <- "text"

jora_token <- df %>%
  unnest_tokens(word, text)
```

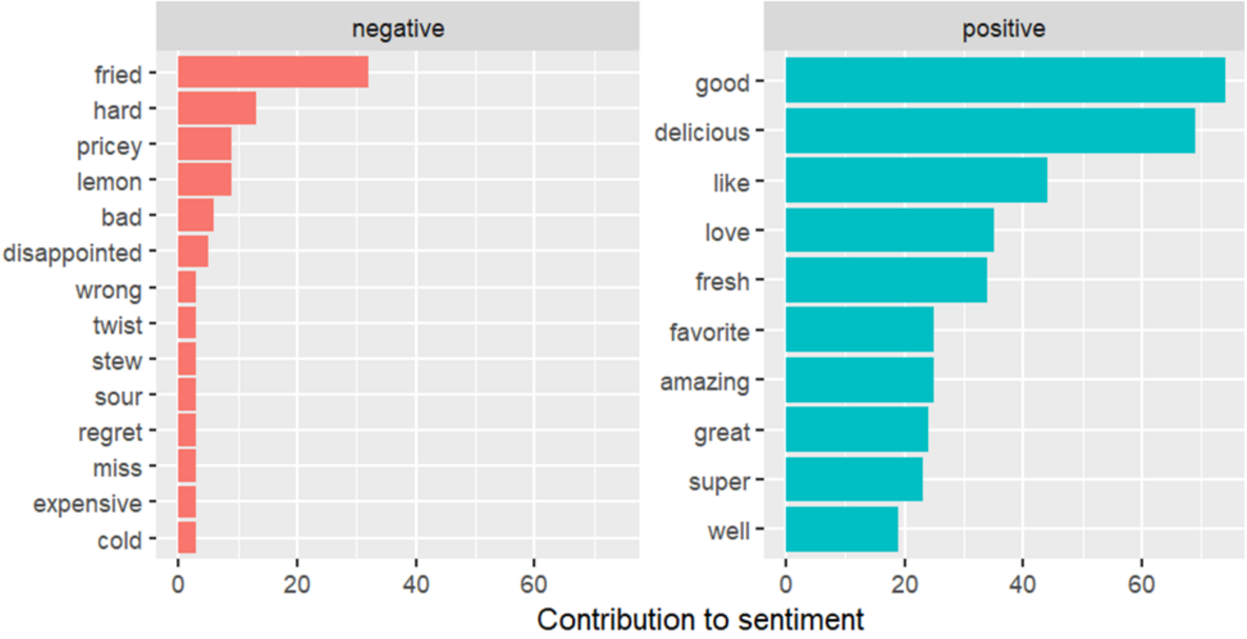
1.3 Sentiment Analysis - Library Bing

```
bing_counts <- jora_token %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_counts
```



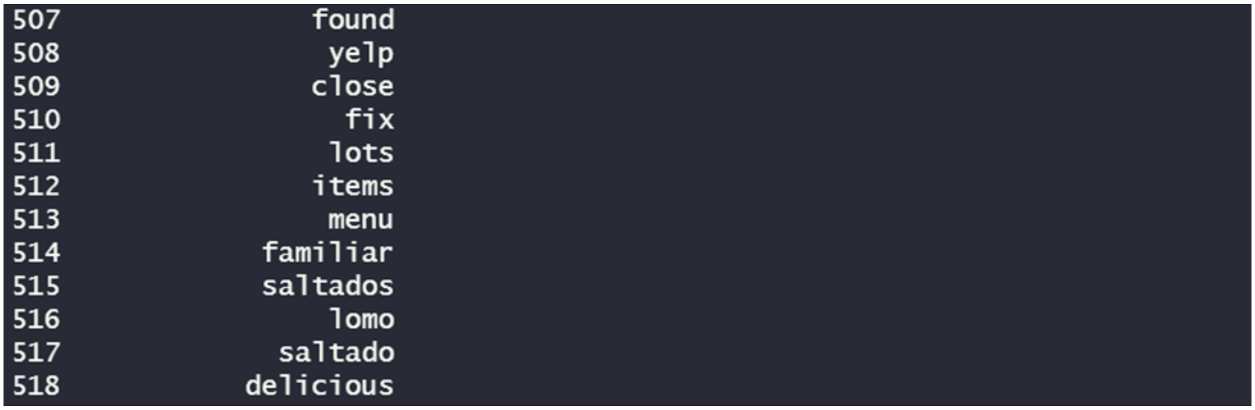
```
> bing_counts
             word sentiment  n
1            good  positive 74
2       delicious  positive 69
3            like  positive 44
4            love  positive 35
5           fresh  positive 34
6           fried  negative 32
7         amazing  positive 25
8        favorite  positive 25
9           great  positive 24
10          super  positive 23
```

```
library(ggplot2)
bing_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()

bing_counts
```



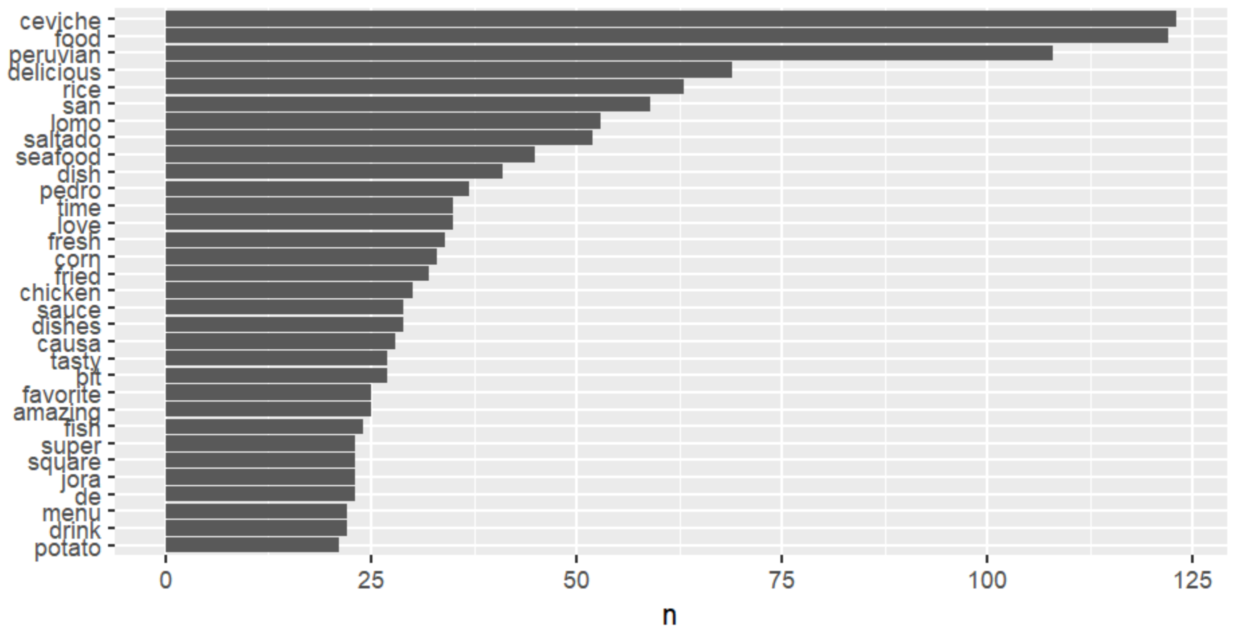
1.4 Tokens frequency of Jora with no stop words

```
library(tidytext)
tidy_jora <- df %>%
  unnest_tokens(word, text)
print(tidy_jora)
#removing stop words
data(stop_words)

jora_no_stop <- tidy_jora %>%
  anti_join(stop_words)
print(jora_no_stop)
#printing the count frequencies for each token without stop words
jora_no_stop %>%
  count(word, sort=TRUE)
```



```
507               found
508                yelp
509               close
510                 fix
511                lots
512               items
513                menu
514            familiar
515            saltados
516                lomo
517             saltado
518           delicious
```

```
library(ggplot2)
freq_hist <-jora_no_stop %>%
  count(word, sort=TRUE) %>%
  filter(n>20) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  xlab(NULL)+
  coord_flip()
print(freq_hist)
```

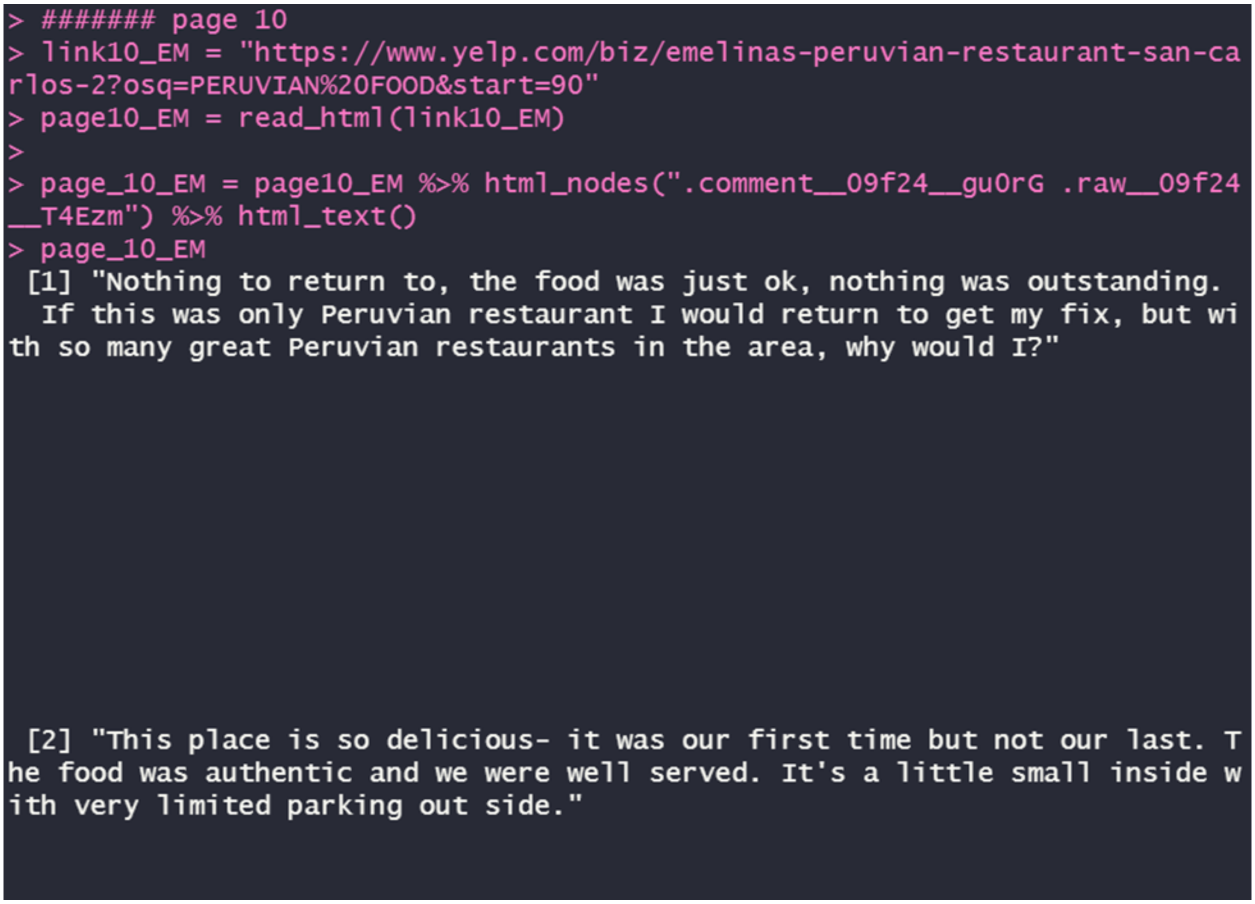


2.1 Collecting data from Emelina Restaurant

```
############################################################################
####
############################################################################
####

###############    EMELINA'S PERUVIAN RESTAURANT

############################################################################
####
############################################################################
####


link1_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD"
page1_EM = read_html(link1_EM)

page_1_EM = page1_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_1_EM

####### page 2
link2_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=10"
page2_EM = read_html(link2_EM)

page_2_EM = page2_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_2_EM
```

```
###### page 3
link3_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=20"
page3_EM = read_html(link3_EM)

page_3_EM = page3_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_3_EM

####### page 4
link4_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=30"
page4_EM = read_html(link4_EM)

page_4_EM = page4_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_4_EM
###### page 5
link5_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=40"
page5_EM = read_html(link5_EM)

page_5_EM = page5_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_5_EM

####### page 6
link6_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=50"
page6_EM = read_html(link6_EM)

page_6_EM = page6_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_6_EM

###### page 7
link7_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=60"
page7_EM = read_html(link7_EM)

page_7_EM = page7_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_7_EM

####### page 8
link8_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=70"
page8_EM = read_html(link8_EM)
```

```
page_8_EM = page8_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_8_EM

###### page 9
link9_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=80"
page9_EM = read_html(link9_EM)

page_9_EM = page9_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_9_EM
####### page 10
link10_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-carlos-
2?osq=PERUVIAN%20FOOD&start=90"
page10_EM = read_html(link10_EM)

page_10_EM = page10_EM %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_10_EM
```



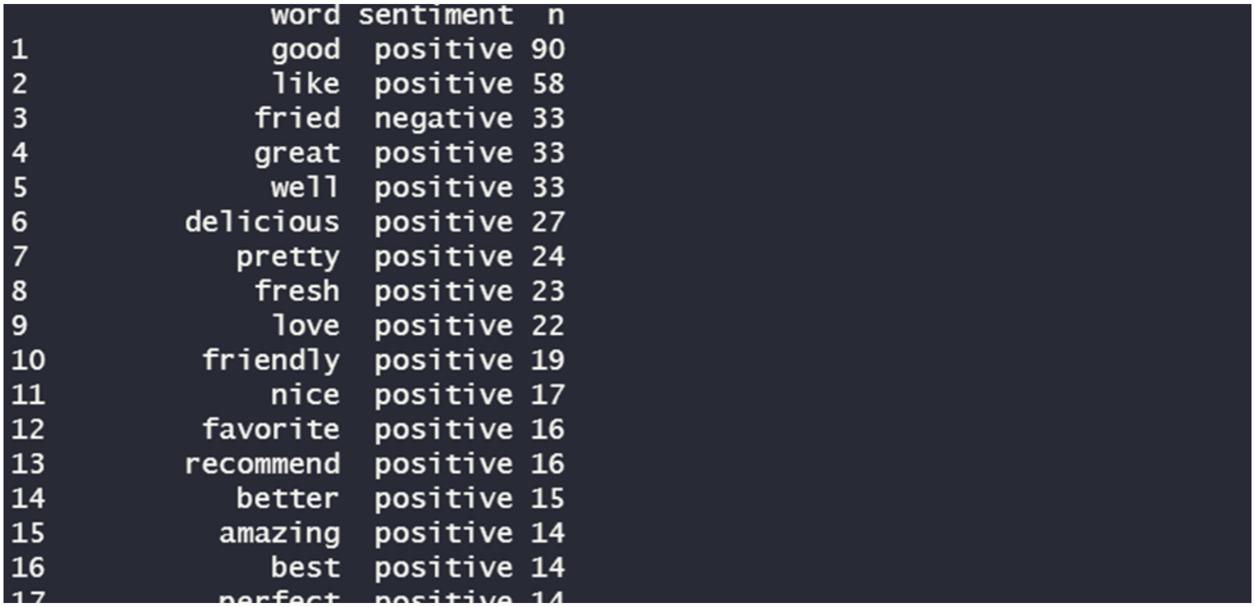
```
> ####### page 10
> link10_EM = "https://www.yelp.com/biz/emelinas-peruvian-restaurant-san-ca
rlos-2?osq=PERUVIAN%20FOOD&start=90"
> page10_EM = read_html(link10_EM)
>
> page_10_EM = page10_EM %>% html_nodes(".comment__09f24__gu0rG .raw__09f24
__T4Ezm") %>% html_text()
> page_10_EM
 [1] "Nothing to return to, the food was just ok, nothing was outstanding.
  If this was only Peruvian restaurant I would return to get my fix, but wi
th so many great Peruvian restaurants in the area, why would I?"












 [2] "This place is so delicious- it was our first time but not our last. T
he food was authentic and we were well served. It's a little small inside w
ith very limited parking out side."
```

2.2 Structuring the data and tokenizing it

```
list_pages_EM =
c(page_1_EM,page_2_EM,page_3_EM,page_4_EM,page_5_EM,page_6_EM,page_7_EM,pag
e_8_EM,page_9_EM,page_10_EM)
df_EM <- as.data.frame(list_pages_EM)

colnames(df_EM)[1] <- "text"


EM_token <- df_EM %>%
  unnest_tokens(word, text)
```

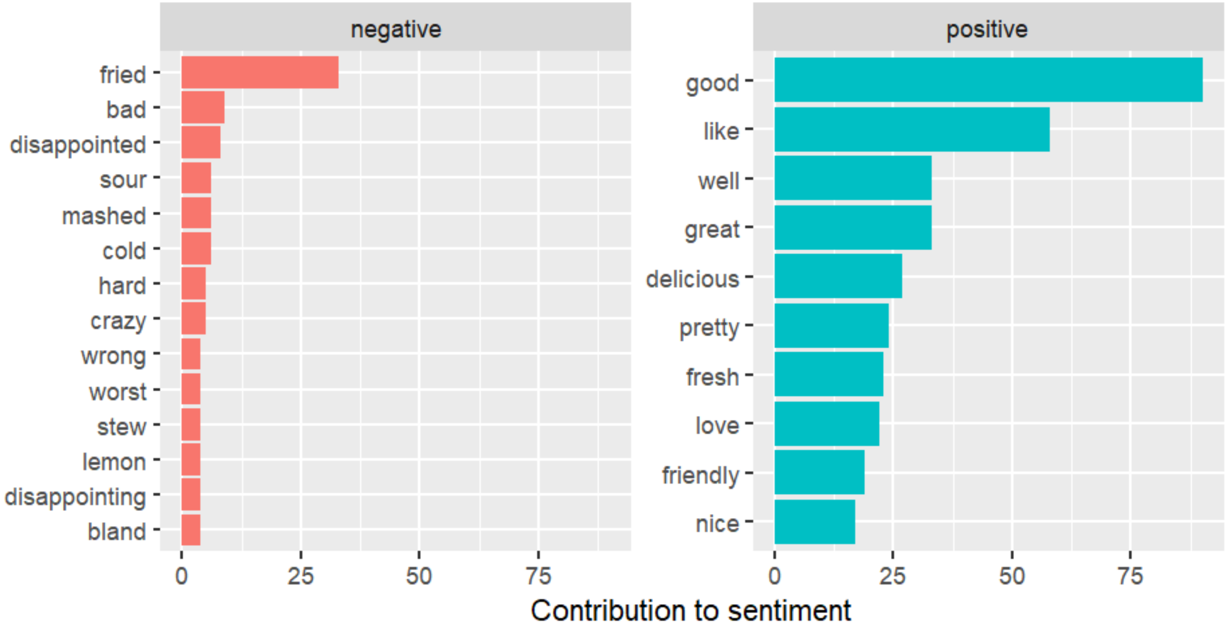
2.3 Sentiment Analysis - Library Bing

```
bing_counts_EM <- EM_token %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_counts_EM
```



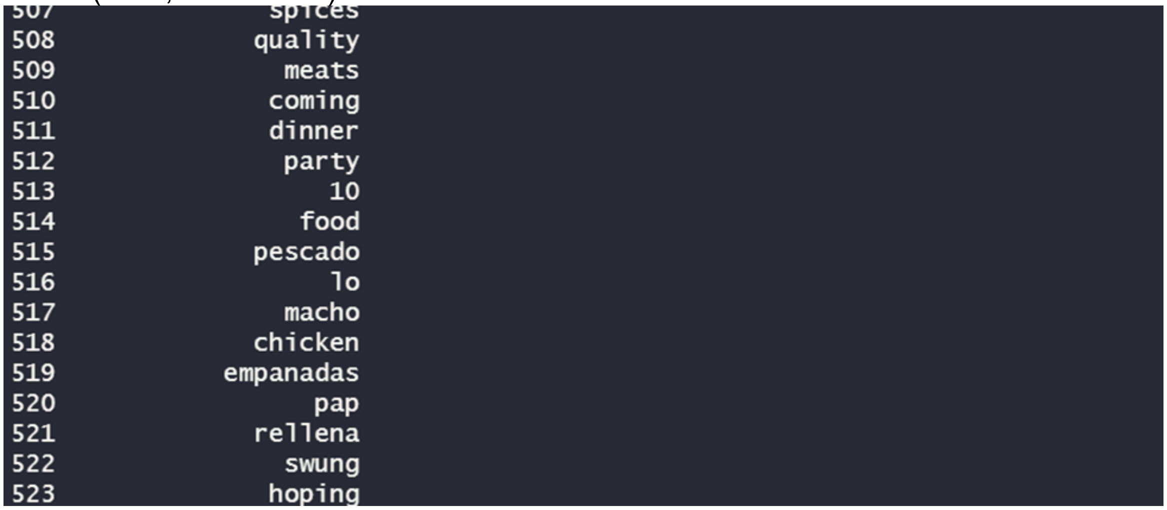
```
                word sentiment  n
1               good  positive 90
2               like  positive 58
3              fried  negative 33
4              great  positive 33
5               well  positive 33
6          delicious  positive 27
7             pretty  positive 24
8              fresh  positive 23
9               love  positive 22
10          friendly  positive 19
11              nice  positive 17
12          favorite  positive 16
13         recommend  positive 16
14            better  positive 15
15           amazing  positive 14
16              best  positive 14
17           perfect  positive 14
```

```
bing_counts_EM %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()
```

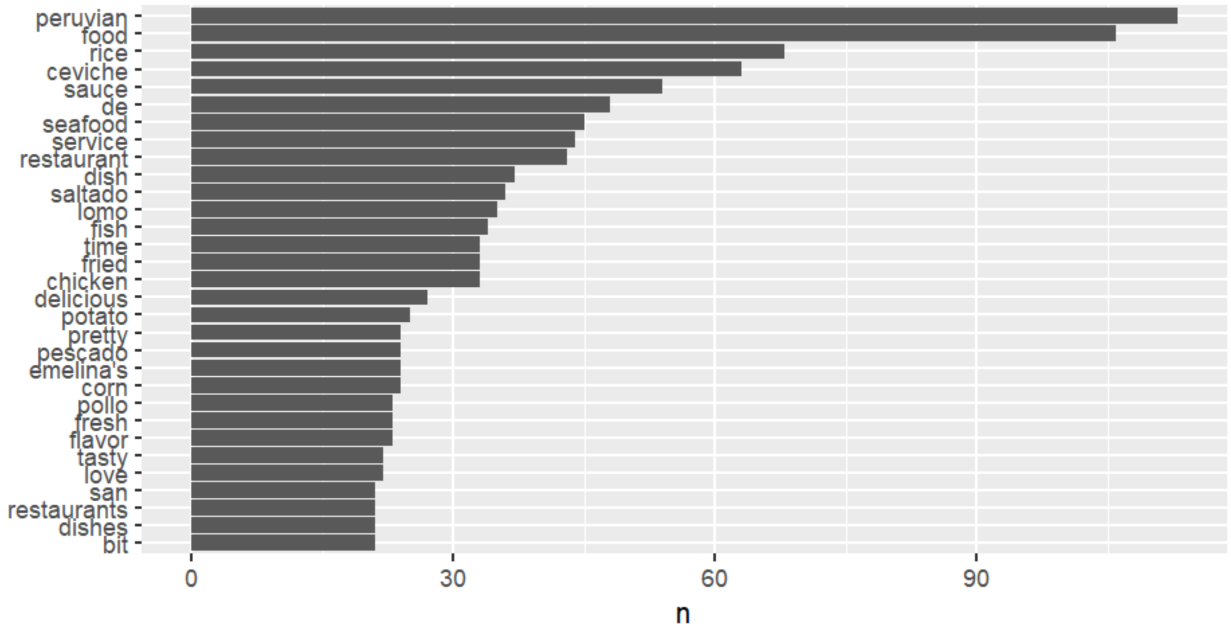


2.4 Tokens frequency of MR Kano with no stop words

```
library(tidytext)
tidy_EM <- df_EM %>%
  unnest_tokens(word, text)
print(tidy_EM)
#removing stop words
data(stop_words)

EM_no_stop <- tidy_EM %>%
  anti_join(stop_words)
print(EM_no_stop)
#printing the count frequencies for each token without stop words
EM_no_stop %>%
  count(word, sort=TRUE)
```


```
507                spices
508               quality
509                 meats
510                coming
511                dinner
512                 party
513                    10
514                  food
515               pescado
516                    lo
517                 macho
518               chicken
519             empanadas
520                   pap
521               rellena
522                 swung
523                hoping
```

```
#plotting the token frequencies:
library(ggplot2)
freq_hist_EM <-EM_no_stop %>%
  count(word, sort=TRUE) %>%
  filter(n>20) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  xlab(NULL)+
  coord_flip()
print(freq_hist_EM)
```

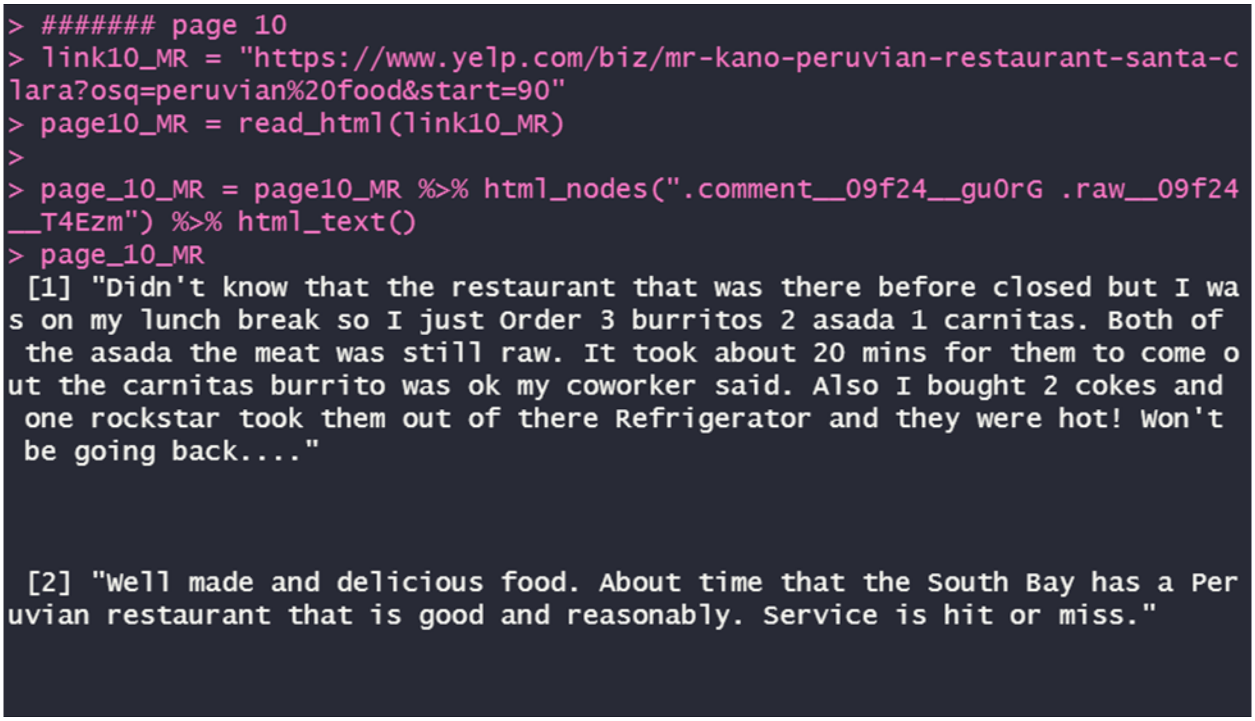


2.1 Collecting data from MR KANO Restaurant

```
##############################################################################
####
##############################################################################
####

###############    MR KANO RESTAURANT

##############################################################################
####
##############################################################################
####

link1_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian+food"
page1_MR = read_html(link1_MR)

page_1_MR = page1_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
```

```
page_1_MR

####### page 2
link2_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=10"
page2_MR = read_html(link2_MR)

page_2_MR = page2_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_2_MR

###### page 3
link3_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=20"
page3_MR = read_html(link3_MR)

page_3_MR = page3_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_3_MR

####### page 4
link4_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=30"
page4_MR = read_html(link4_MR)

page_4_MR = page4_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_4_MR
###### page 5
link5_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=40"
page5_MR = read_html(link5_MR)

page_5_MR = page5_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_5_MR

####### page 6
link6_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=50"
page6_MR = read_html(link6_MR)

page_6_MR = page6_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_6_MR

###### page 7
link7_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=60"
page7_MR = read_html(link7_MR)
```

```
page_7_MR = page7_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_7_MR

####### page 8
link8_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=70"
page8_MR = read_html(link8_MR)

page_8_MR = page8_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_8_MR

###### page 9
link9_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=80"
page9_MR = read_html(link9_MR)

page_9_MR = page9_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_9_MR
####### page 10
link10_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-
clara?osq=peruvian%20food&start=90"
page10_MR = read_html(link10_MR)

page_10_MR = page10_MR %>% html_nodes(".comment__09f24__gu0rG
.raw__09f24__T4Ezm") %>% html_text()
page_10_MR
```



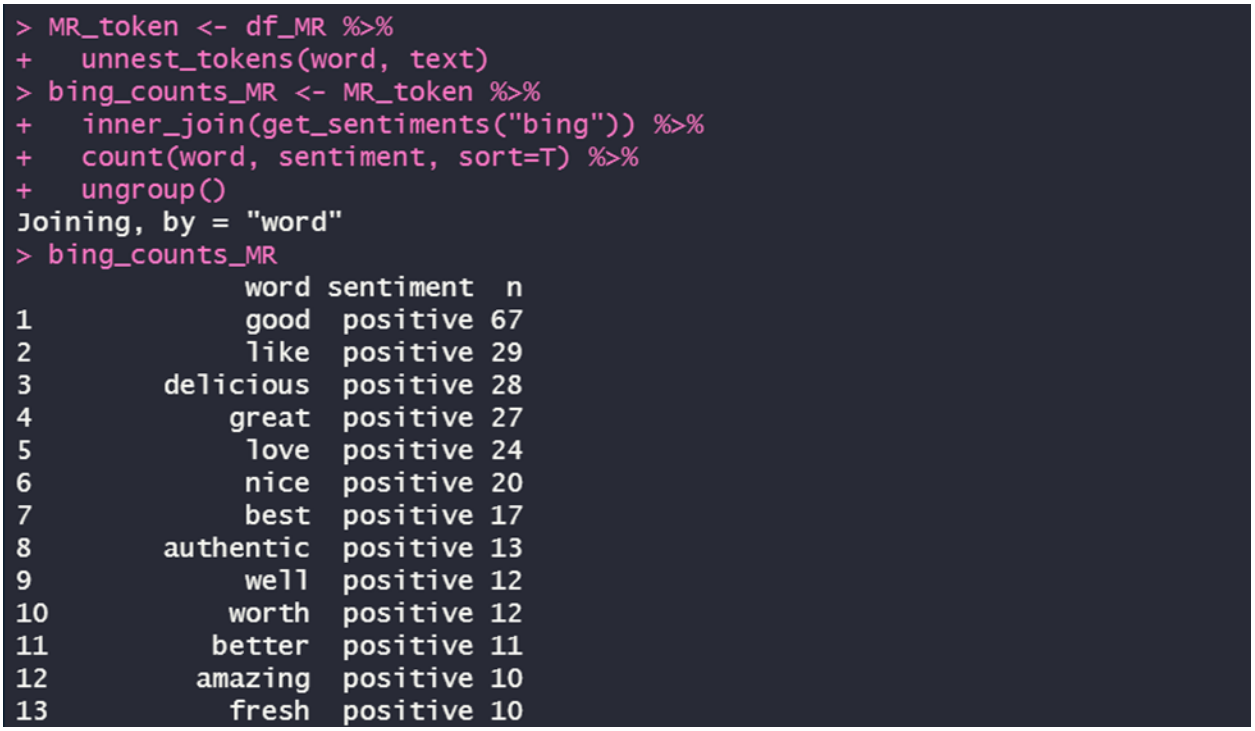
```
> ####### page 10
> link10_MR = "https://www.yelp.com/biz/mr-kano-peruvian-restaurant-santa-c
lara?osq=peruvian%20food&start=90"
> page10_MR = read_html(link10_MR)
>
> page_10_MR = page10_MR %>% html_nodes(".comment__09f24__gu0rG .raw__09f24
__T4Ezm") %>% html_text()
> page_10_MR
 [1] "Didn't know that the restaurant that was there before closed but I wa
s on my lunch break so I just Order 3 burritos 2 asada 1 carnitas. Both of
 the asada the meat was still raw. It took about 20 mins for them to come o
ut the carnitas burrito was ok my coworker said. Also I bought 2 cokes and
 one rockstar took them out of there Refrigerator and they were hot! Won't
 be going back...."



 [2] "Well made and delicious food. About time that the South Bay has a Per
uvian restaurant that is good and reasonably. Service is hit or miss."
```

2.2 Structuring the data and tokenizing it

```
list_pages_MR =
c(page_1_MR,page_2_MR,page_3_MR,page_4_MR,page_5_MR,page_6_MR,page_7_MR,page_8_MR,page_9_MR,page_10_MR)
df_MR <- as.data.frame(list_pages_MR)

colnames(df_MR)[1] <- "text"


MR_token <- df_MR %>%
  unnest_tokens(word, text)
```

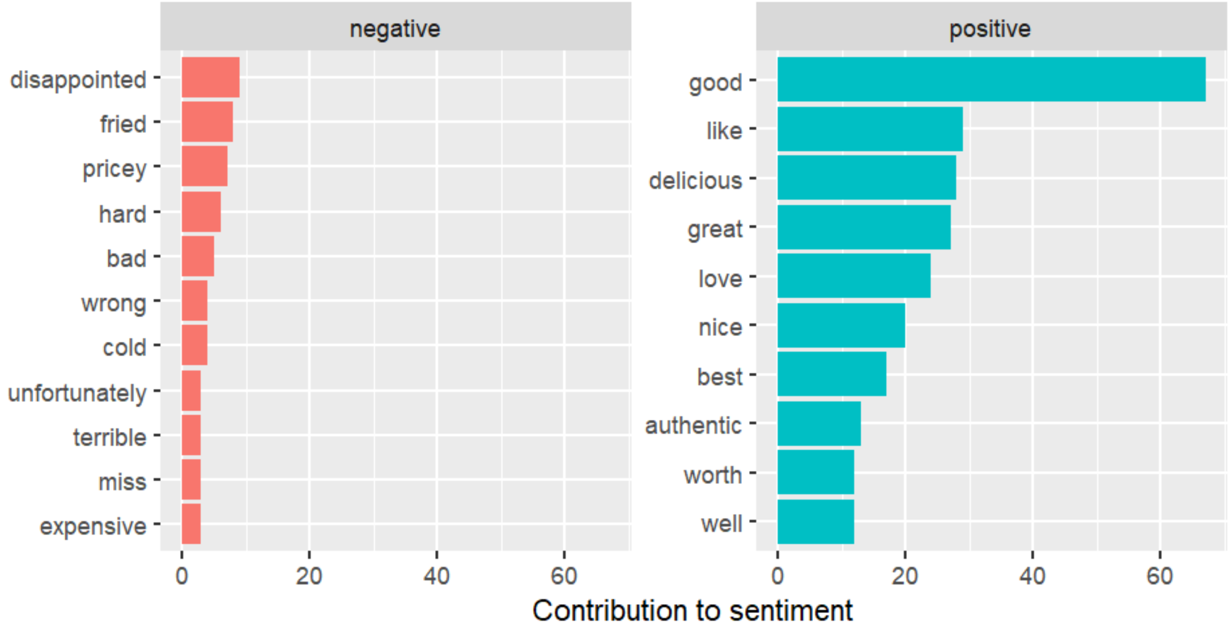
2.3 Sentiment Analysis - Library Bing

```
bing_counts_MR <- MR_token %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_counts_MR
```



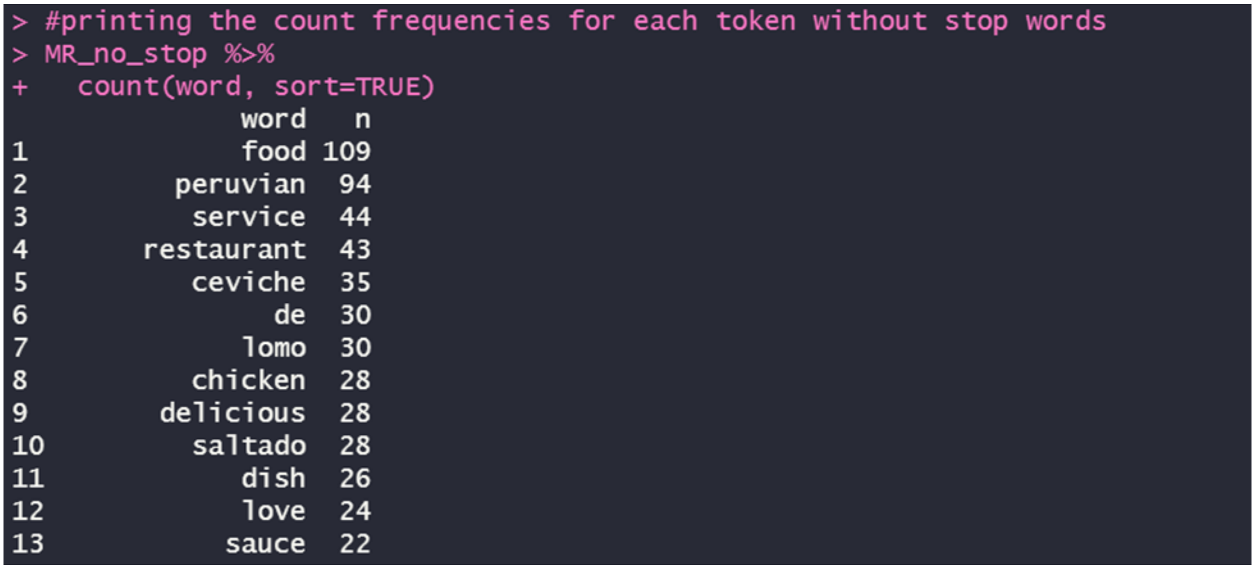
```
> MR_token <- df_MR %>%
+   unnest_tokens(word, text)
> bing_counts_MR <- MR_token %>%
+   inner_join(get_sentiments("bing")) %>%
+   count(word, sentiment, sort=T) %>%
+   ungroup()
Joining, by = "word"
> bing_counts_MR
            word sentiment  n
1           good  positive 67
2           like  positive 29
3      delicious  positive 28
4          great  positive 27
5           love  positive 24
6           nice  positive 20
7           best  positive 17
8      authentic  positive 13
9           well  positive 12
10         worth  positive 12
11        better  positive 11
12       amazing  positive 10
13         fresh  positive 10
```

```
bing_counts_MR %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()
```



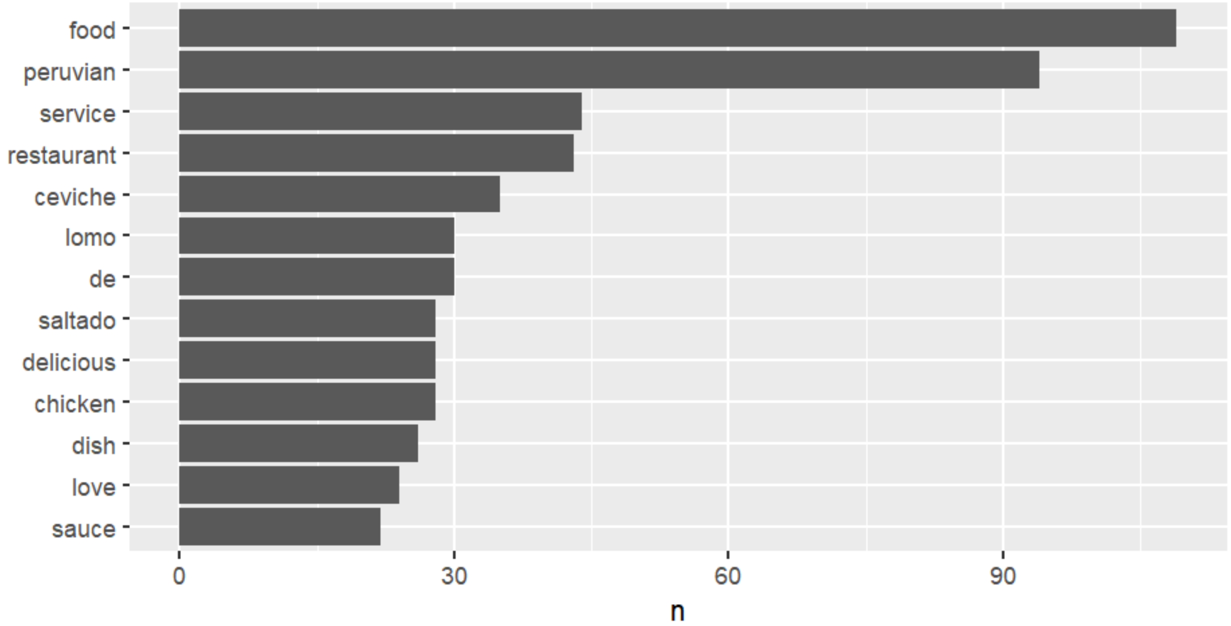
3.4 Tokens frequency of MR Kano with no stop words

```
library(tidytext)
tidy_MR <- df_MR %>%
  unnest_tokens(word, text)
print(tidy_MR)
#removing stop words
data(stop_words)

MR_no_stop <- tidy_MR %>%
  anti_join(stop_words)
print(MR_no_stop)
#printing the count frequencies for each token without stop words
MR_no_stop %>%
  count(word, sort=TRUE)
```



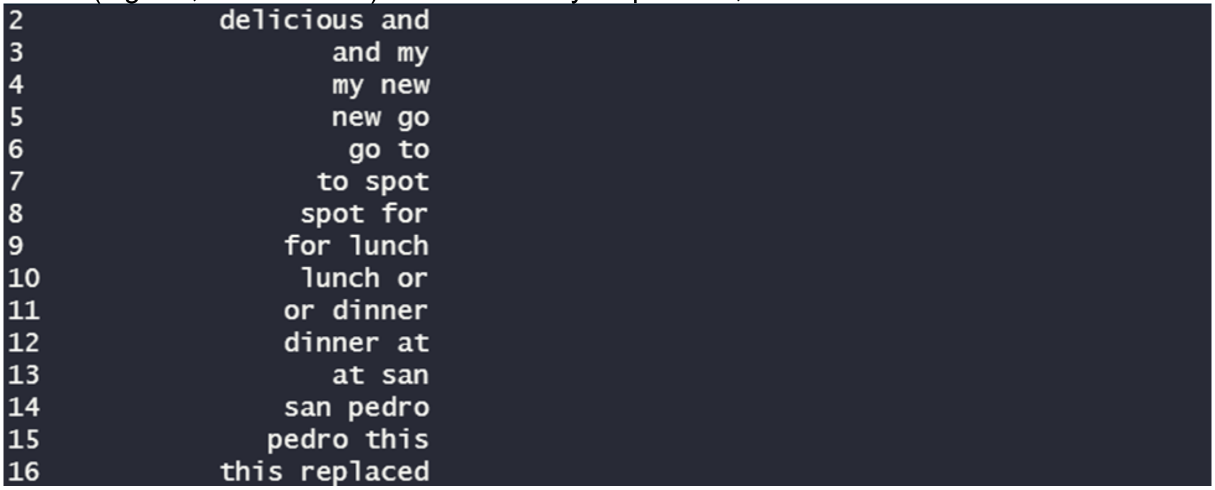
```
> #printing the count frequencies for each token without stop words
> MR_no_stop %>%
+   count(word, sort=TRUE)
              word   n
1             food 109
2         peruvian  94
3          service  44
4       restaurant  43
5          ceviche  35
6               de  30
7             lomo  30
8          chicken  28
9        delicious  28
10         saltado  28
11            dish  26
12            love  24
13           sauce  22
```

```
#plotting the token frequencies:
library(ggplot2)
freq_hist_MR <-MR_no_stop %>%
  count(word, sort=TRUE) %>%
  filter(n>20) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  xlab(NULL)+
  coord_flip()
print(freq_hist_MR)
```



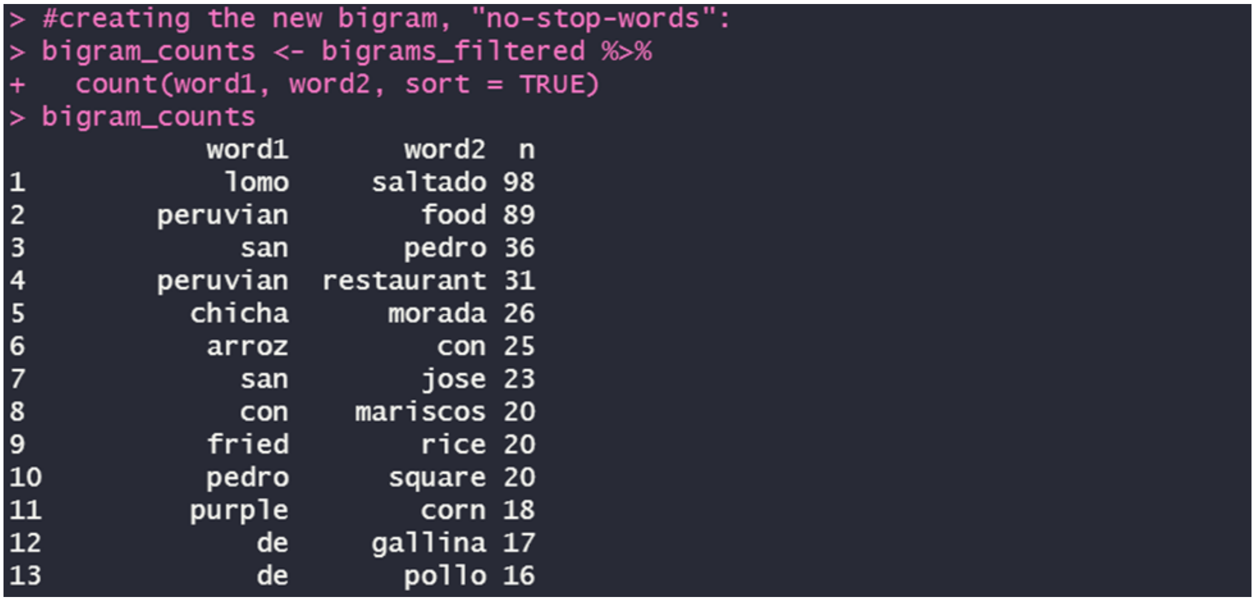
4. Analysis N-Grams, size 2

```
####Merging data of 3 restaurants

df_res <- rbind(df, df_MR, df_EM)

#We will tokenize the data by ngram by ngram no but word
res_bigrams <- df_res %>%
  unnest_tokens(bigram, text, token = 'ngrams', n=2)
# The location information is book
# bigram = we have now pair of tokens

res_bigrams #We want to see the bigrams (words that appear together, "pairs")

res_bigrams %>%
  count(bigram, sort = TRUE) #this has many stop words, need to remove them
```



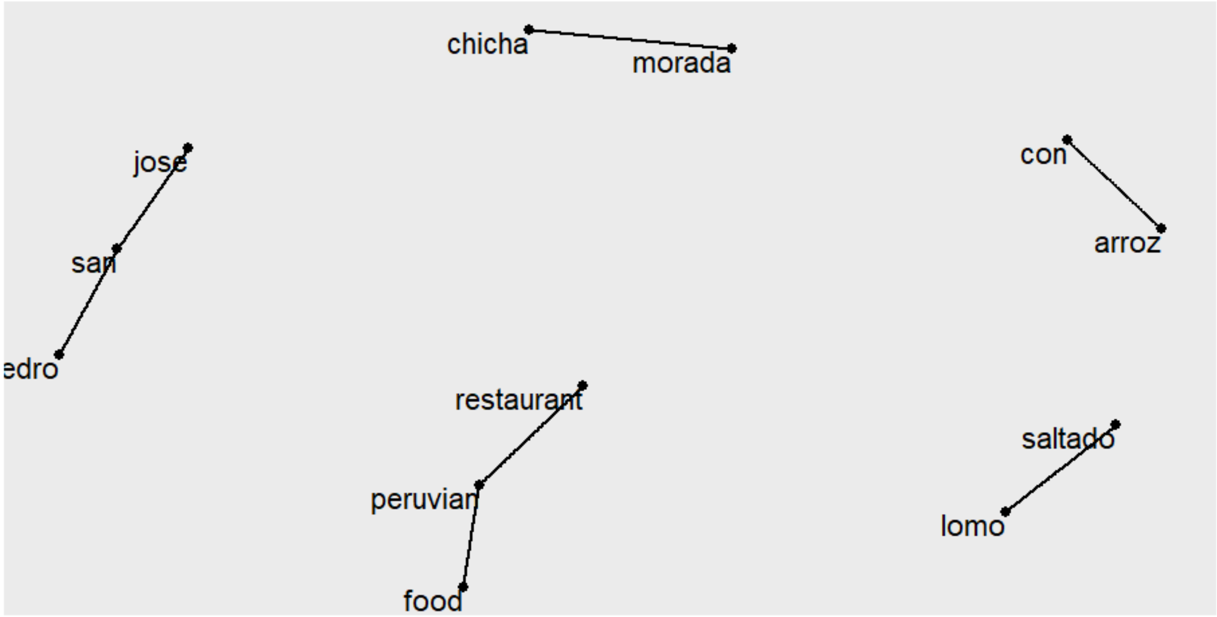
```
2           delicious and
3                  and my
4                  my new
5                  new go
6                   go to
7                 to spot
8                spot for
9               for lunch
10               lunch or
11              or dinner
12              dinner at
13                 at san
14              san pedro
15             pedro this
16          this replaced
```

```
library(tidyr)
bigrams_separated <- res_bigrams %>%
  separate(bigram, c("word1","word2"), sep = " ") # Separating the bigrams into 2 tokens per
observation
# the output is 2 separate tokens

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>% #! exclamation sign removes
  filter(!word2 %in% stop_words$word)

#creating the new bigram, "no-stop-words":
bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

bigram_counts
```



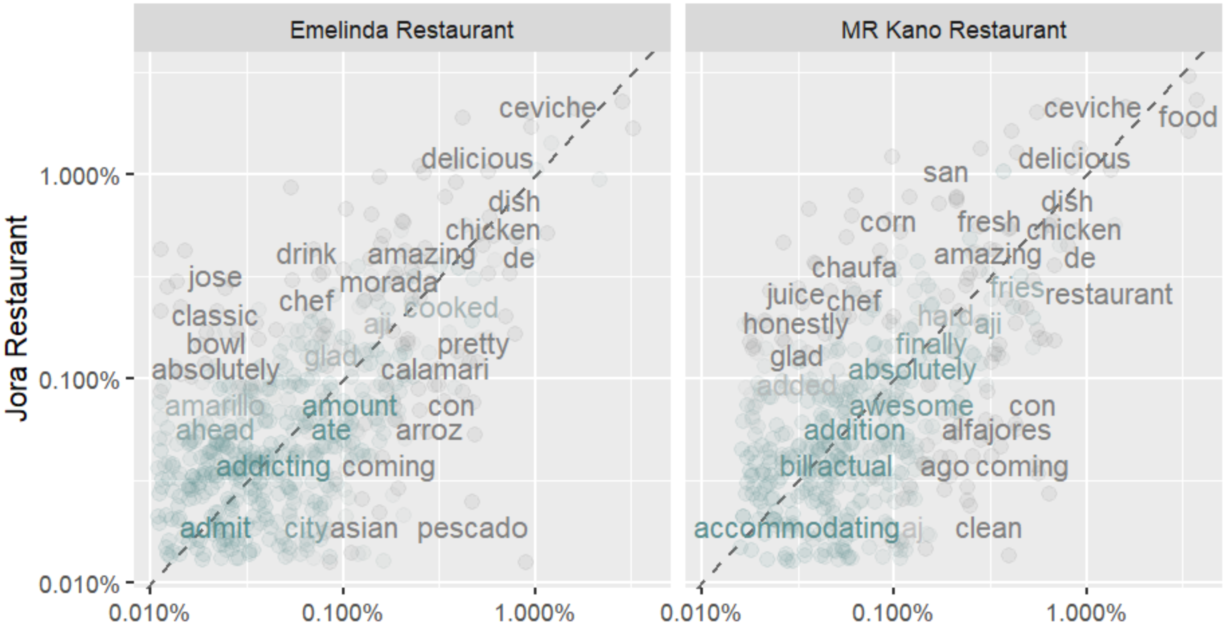
```
> #creating the new bigram, "no-stop-words":
> bigram_counts <- bigrams_filtered %>%
+   count(word1, word2, sort = TRUE)
> bigram_counts
         word1      word2  n
1         lomo    saltado 98
2     peruvian       food 89
3          san      pedro 36
4     peruvian restaurant 31
5       chicha     morada 26
6        arroz        con 25
7          san       jose 23
8          con   mariscos 20
9        fried       rice 20
10       pedro     square 20
11      purple       corn 18
12          de    gallina 17
13          de      pollo 16
```

```
#####################################################
####### VISUALISING A BIGRAM NETWORK #################
#####################################################

#install.packages("igraph")
library(igraph)
bigram_graph <- bigram_counts %>%
  filter(n>20) %>% #for our own project n small
  graph_from_data_frame()

bigram_graph

#install.packages("ggraph")
library(ggraph)

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link()+    #we have 2 geometrics edge and node
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)
```



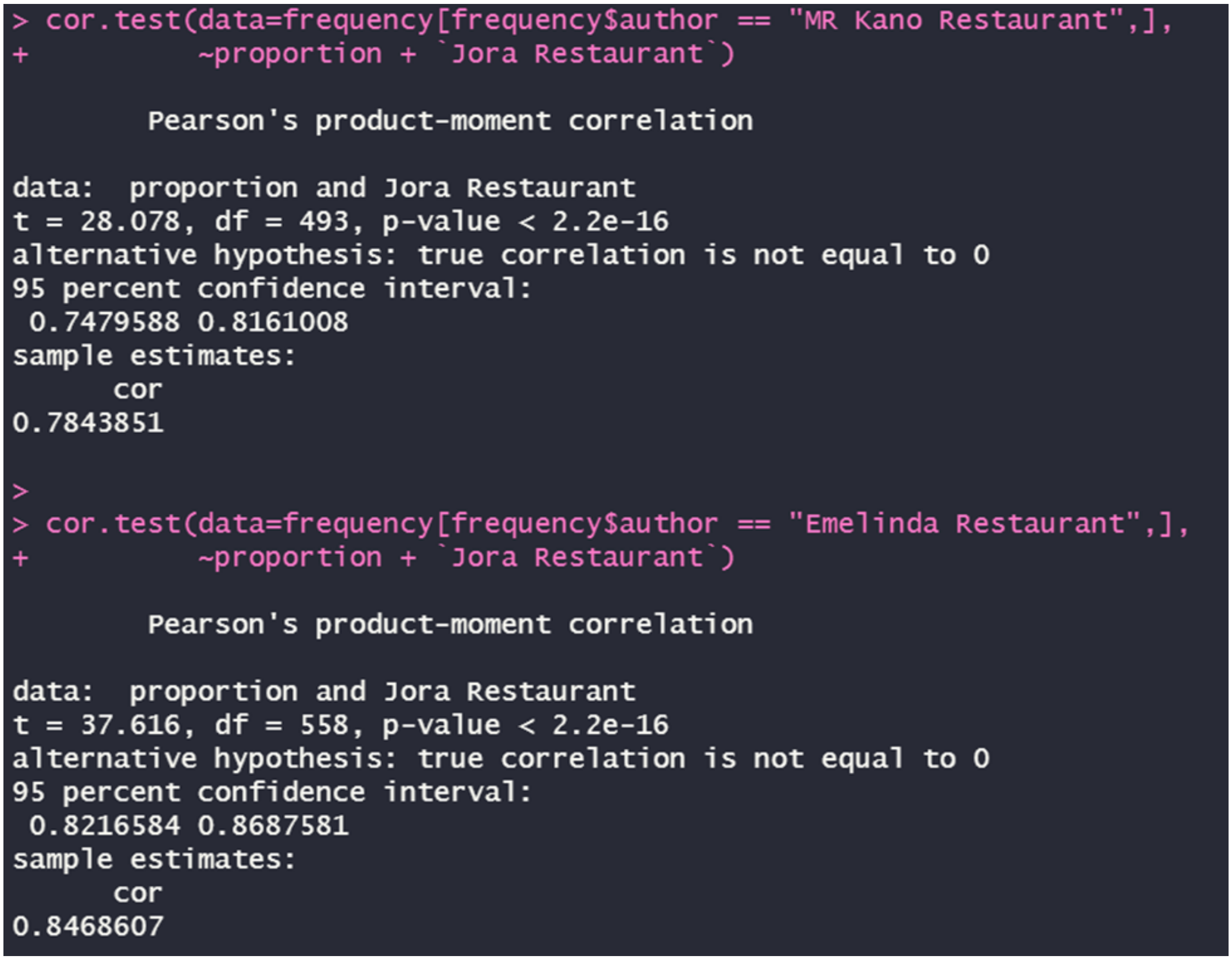
5. Correlograms

```
library(tidyr)
library(stringr)
frequency <- bind_rows(mutate(jora_no_stop, author="Jora Restaurant"),
                       mutate(MR_no_stop, author= "MR Kano Restaurant"),
                       mutate(EM_no_stop, author="Emelinda Restaurant")
)%>%#closing bind_rows
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(author, proportion) %>%
  gather(author, proportion, `MR Kano Restaurant`, `Emelinda Restaurant`)

#let's plot the correlograms:
library(scales)
ggplot(frequency, aes(x=proportion, y=`Jora Restaurant`,
                      color = abs(`Jora Restaurant`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~author, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "Jora Restaurant", x=NULL)
```


Emelinda Restaurant
MR Kano Restaurant
Jora Restaurant
1.000%
0.100%
0.010%
0.010%
0.100%
1.000%
0.010%
0.100%
1.000%
ceviche
delicious
dish
chicken
drink
amazing
de
jose
morada
chef
cooked
classic
aji
bowl
pretty
glad
absolutely
calamari
amarillo
amount
con
ahead
ate
arroz
addicting
coming
admit
city
asian
pescado
ceviche
food
delicious
san
dish
corn
fresh
chicken
amazing
de
chaufa
fries
juice
chef
restaurant
hard
aji
honestly
finally
glad
absolutely
added
awesome
con
addition
alfajores
bilractual
ago
coming
accommodating
aj
clean

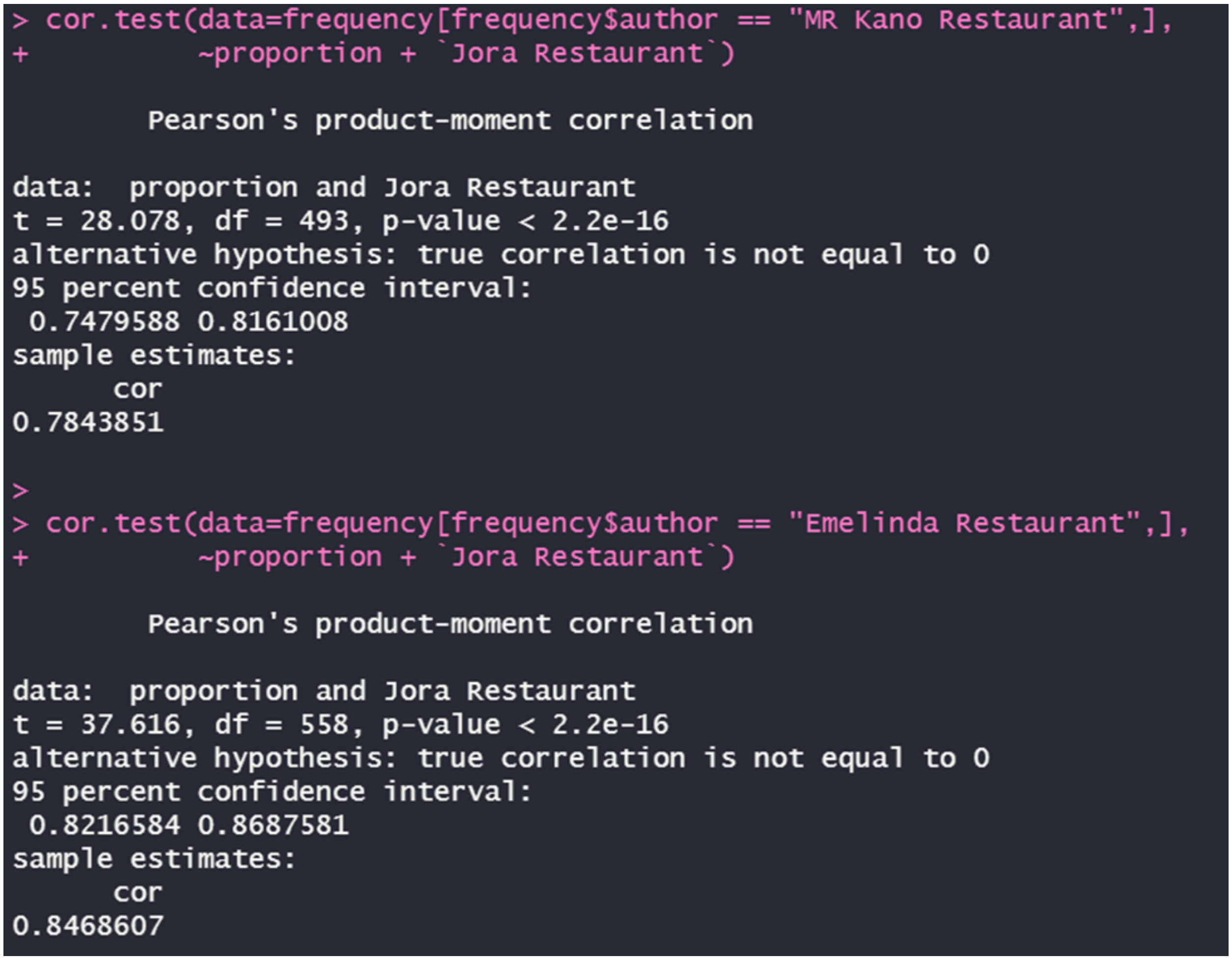
6. Correlation test

```
cor.test(data=frequency[frequency$author == "MR Kano Restaurant",],
        ~proportion + `Jora Restaurant`)

cor.test(data=frequency[frequency$author == "Emelinda Restaurant",],
        ~proportion + `Jora Restaurant`)
```

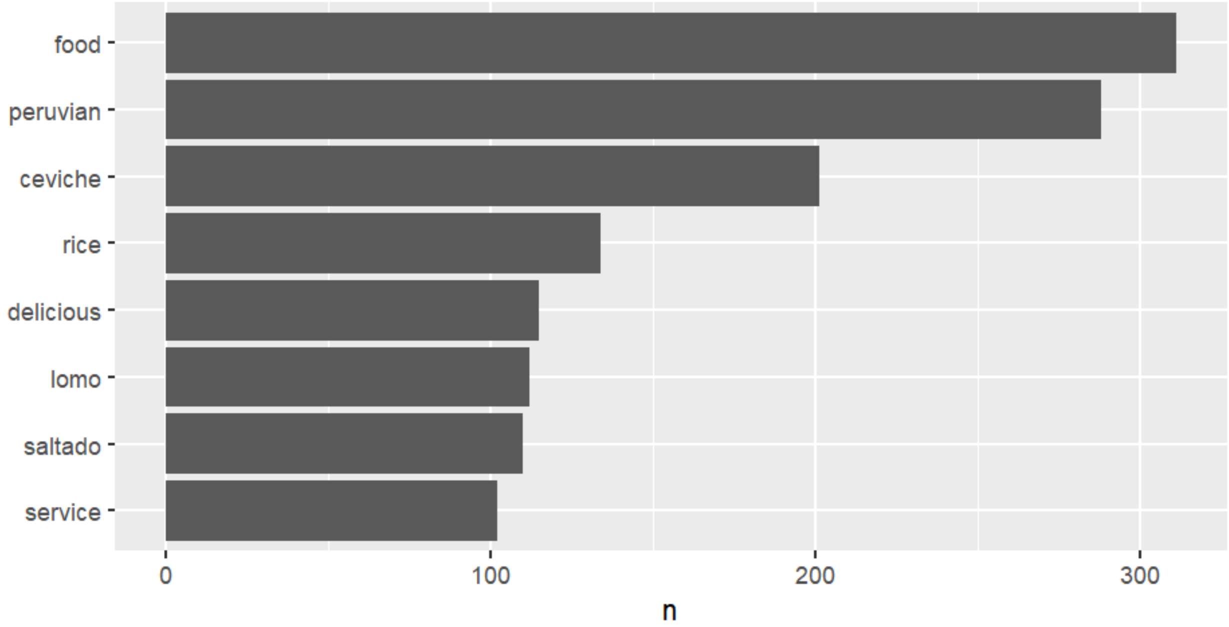


```
> cor.test(data=frequency[frequency$author == "MR Kano Restaurant",],
+           ~proportion + `Jora Restaurant`)

        Pearson's product-moment correlation

data:  proportion and Jora Restaurant
t = 28.078, df = 493, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7479588 0.8161008
sample estimates:
      cor 
0.7843851 

>
> cor.test(data=frequency[frequency$author == "Emelinda Restaurant",],
+           ~proportion + `Jora Restaurant`)

        Pearson's product-moment correlation

data:  proportion and Jora Restaurant
t = 37.616, df = 558, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8216584 0.8687581
sample estimates:
      cor 
0.8468607 
```

7. Frequency Tokens all restaurants



```
> cor.test(data=frequency[frequency$author == "MR Kano Restaurant",],
+           ~proportion + `Jora Restaurant`)

        Pearson's product-moment correlation

data:  proportion and Jora Restaurant
t = 28.078, df = 493, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7479588 0.8161008
sample estimates:
      cor
0.7843851

>
> cor.test(data=frequency[frequency$author == "Emelinda Restaurant",],
+           ~proportion + `Jora Restaurant`)

        Pearson's product-moment correlation

data:  proportion and Jora Restaurant
t = 37.616, df = 558, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8216584 0.8687581
sample estimates:
      cor
0.8468607
```


food
peruvian
ceviche
rice
delicious
lomo
saltado
service
n