ADS Final Report

Peter Kolodziej, Tianyi Li, Hanao Li, Fanyi Yang

5/5/2019

```
## Loading required package: pacman
```

Introduction

Yelp, founded in 2004, publishes crowd-sources reviews about businesses (Wikipedia), helping people locate local restaurants based on star-rating and reviews. As more and more customers rely on Yelp for food hunting, the review on Yelp has become a critical index for restaurants.

In this project, we focused on the analysis of the text reviews and star prediction. We are interested in this topic for two reasons. First of all, rating acts as an identifier for discriminating positive or negative sentiment, which is an interesting feature that directly ties to business quality. Second, rating is intangible, and thus difficult to quantify in an exact way. Therefore, building a model to predict rating accurately based on comments would be useful. With such a model, we can access unlabeled text. For example, we can look at unlabeled reviews and assign score to it. The predictive model may also enable us to monitor the social presence of a business on social media and other communication venues. An important feature of this model is that it can be used to address the misclassification of reviews. Whenever people write a really negative reviews, and submitted it, and mistakenly put a five star on it, the model can identify such error and fix it. In that way, our model functions as a fake review filter that can improve the efficiency of Yelp's rating.

Source Data

The dataset was downloaded from the Yelp Dataset Challenge, consisting of five files, including yelp_user, yelp_review, yelp_checkin, yelp_tip and yelp_business, in json format. For our analysis, we focused on rating and reviews for restaurants and used the customer reviews and business attributes data. We extracted restaurants from all business, and there werenearly 5,000,000 customer reviews collected from approximately 75,000 restaurants.

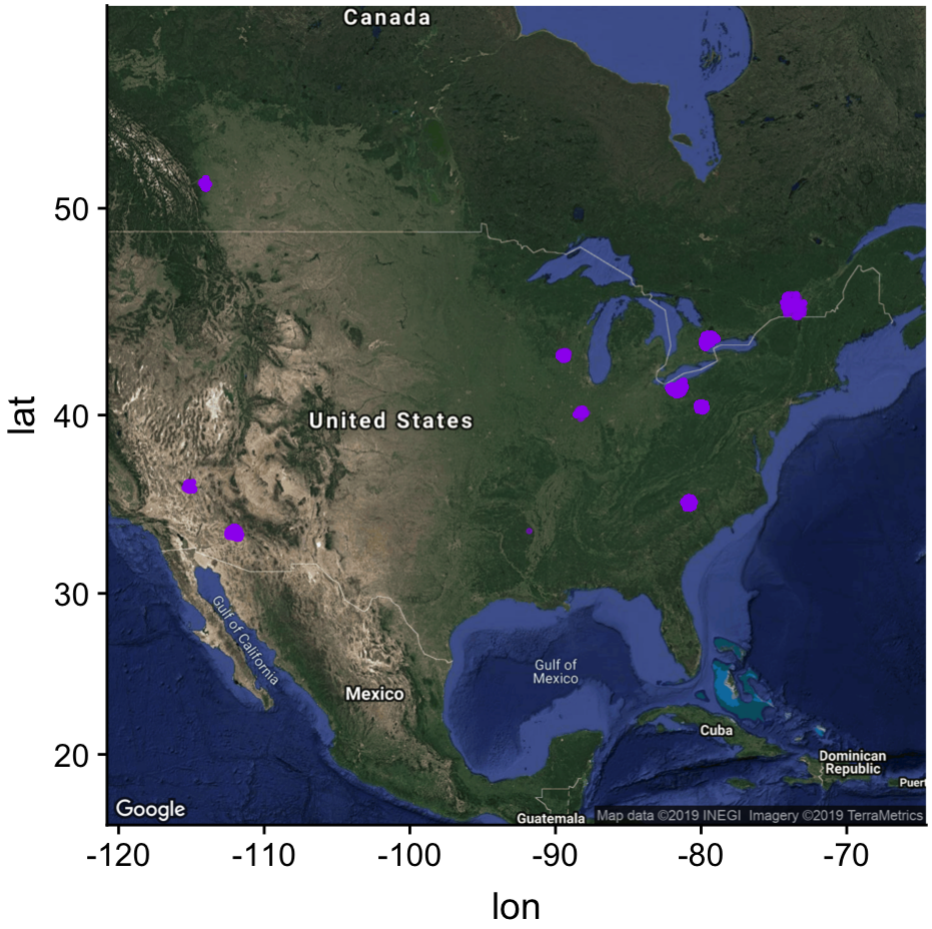
Business data is about75,000 observations, with 58 variables like business id, business name, cities, state, zip code, attributes like the environment of the restaurants (the presence of wifi, parking lots or not etc.), rating, review counts and etc. The review data set is larger, which is of 2GB text review data. Attributes in review data include 9 variables, such as business id, star rating, review content and people's opinion on reviews (funny, cool and useful) and etc.

Examination

In this section, we will investigate the distribution of the business and review data. First we will look at the distribution of stars for the review dataset. We will then define reviews with one to three stars as negative reviews and reviews with four and five stars as positive reviews.We will check the distribution of stars and groups Positive or Negative) for each group using 1% of the data. Then, we will use natural language processing for our sample dataset. We tokenized the words into unigram and bigram tokens, and then removed stop words and punctuations. We will check the most frequent words used in each group and created a 2-d frequency graph to visualze these

words. Wordclouds for unigram and bigrams and bigram network graph can also be plotted to see the relationship between the words used in the reviews, although they provide duplicated information and so will not be used in the report. Please see the Reporting Engine for this information.

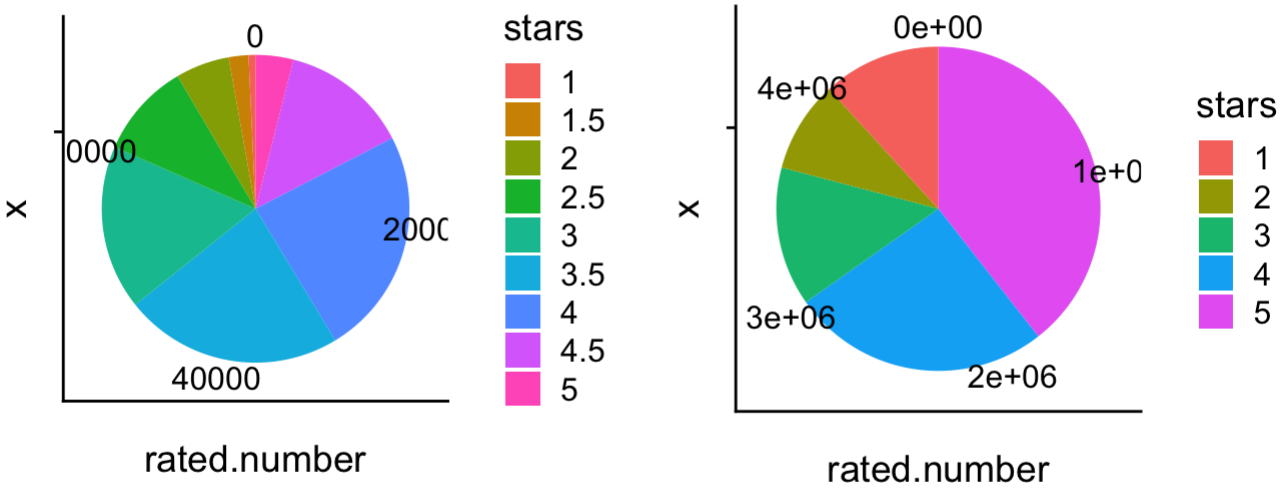
```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.023099,-92.71177&zoom=4&size=640x640&scale=2&maptype=hybrid&key=xxx
```



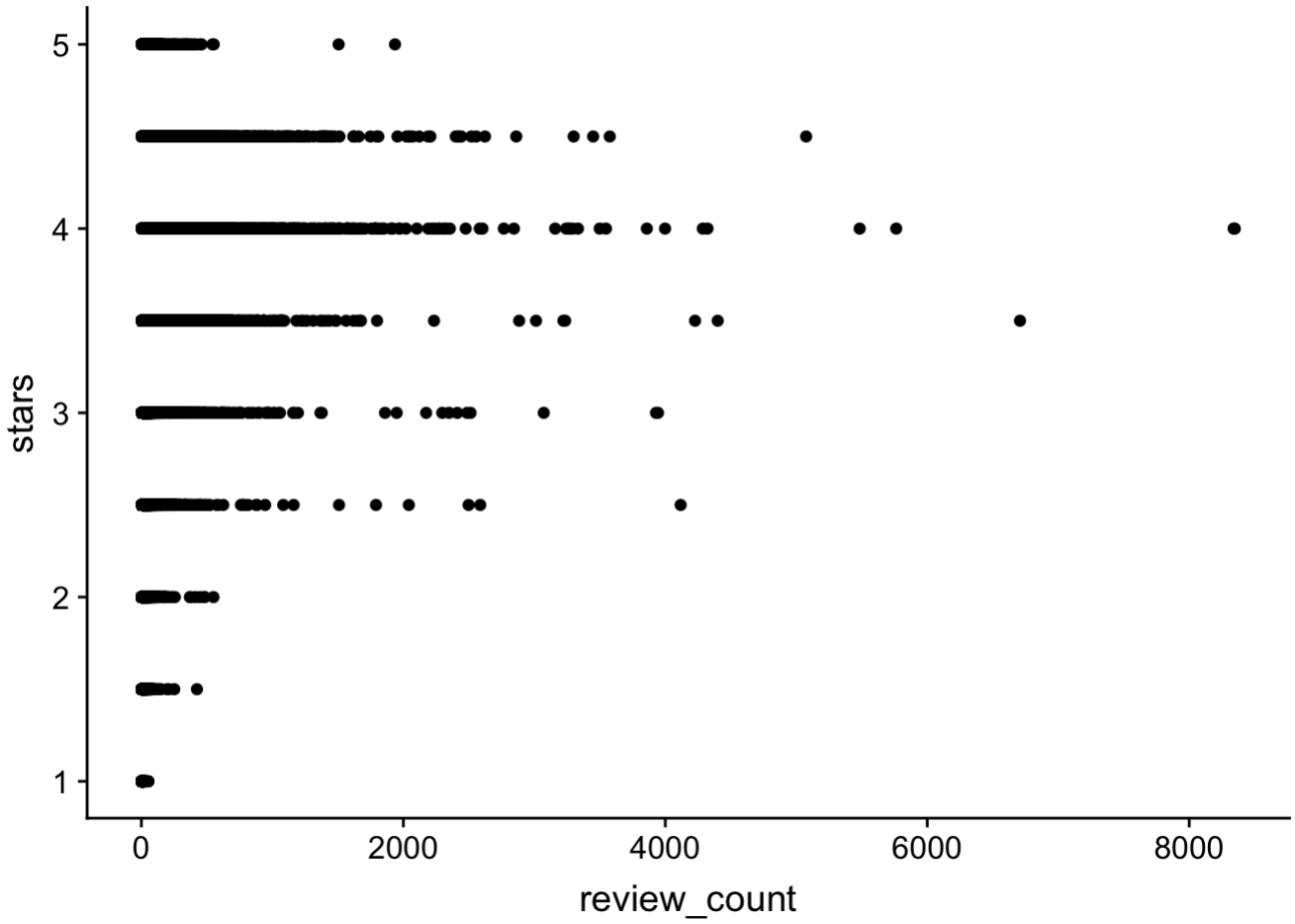

```
##             V1     N
##  1:     Toronto 10094
##  2:   Las Vegas  8290
##  3:     Phoenix  5135
##  4:    Montréal  4540
##  5:     Calgary  3694
##  6:   Charlotte  3491
##  7:  Pittsburgh  3124
##  8:  Scottsdale  2004
##  9:   Cleveland  1813
## 10: Mississauga  1732
```

Next, we want to investigate how the restaurants cluster on the map. By using the leaflet and ggmap packages, based on the restaurants' distances and at a given zoom level, we can see how the restaurants cluster in each specific area and we also can see restaurants name on the map. From our shiny app initial clustering map, there are five primary clusters: one in the west coast of United States, one in the east coast of United States, one in the cities which contain the most business data. The top ten are displayed.

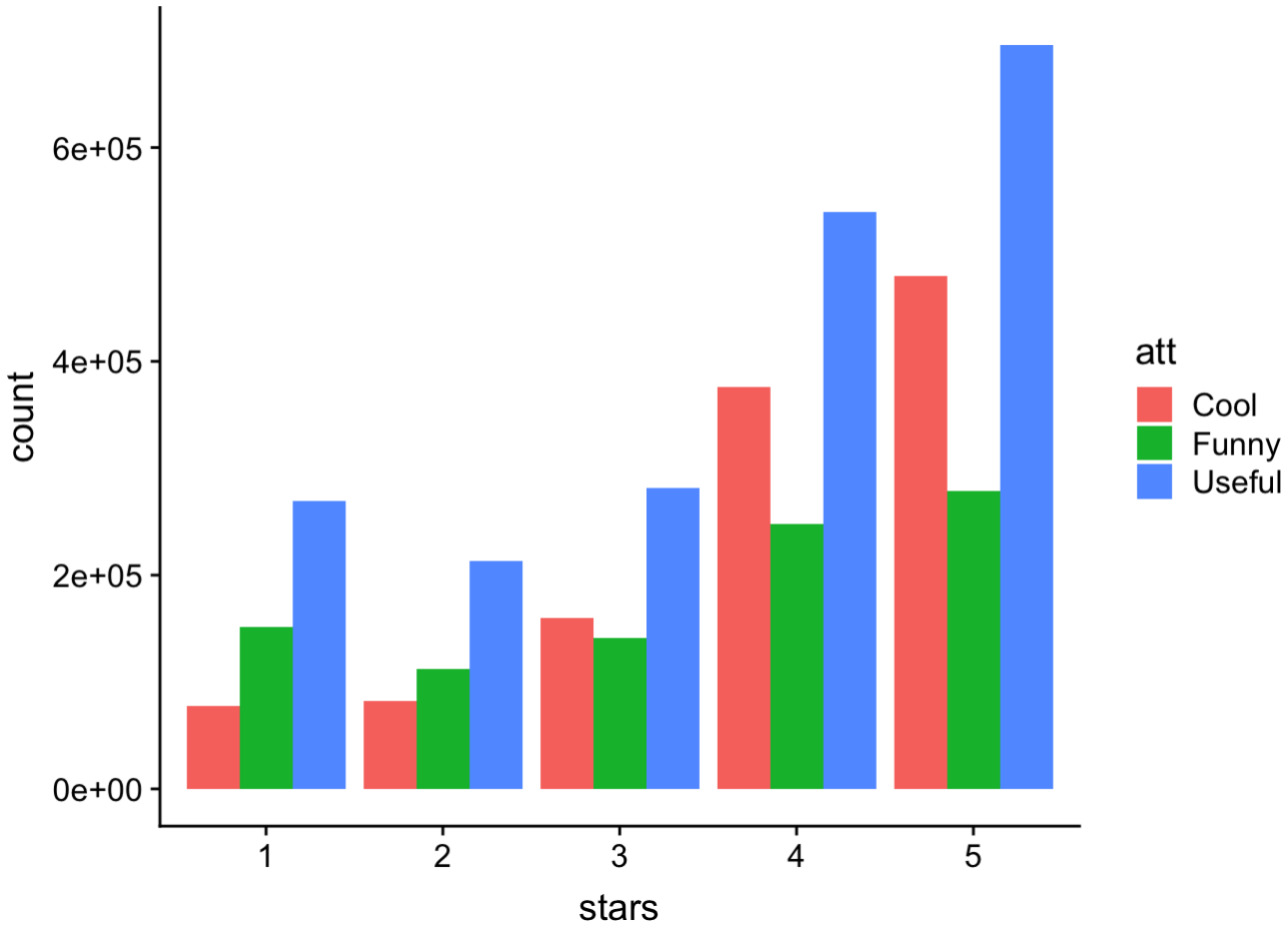
In our Reporting Engine we want to investigate the business level of each cuisine type by showing restaurants' star rating distribution of each type of cuisine in a specific city. What we built in the shiny app was the selection of city sorted by the number of restaurants they have and the thirteen types of cuisine. There are two check boxes, one is whether the star rating is in sorted order and another one is whether the star rating is showing percentage values. When we select one specific city, it will show what percent of each star rating of the cuisine we select in that city. For example, when we choose Las Vegas with French cuisine, we can see in Las Vegas, there is 38.36 percent of French cuisine is 4.5-star, and 2.74 percent of French cuisine is 5-star. From the star rating distribution, we can conclude 4-star and 4.5-star rating restaurants dominate the French restaurant's business in Las Vegas.



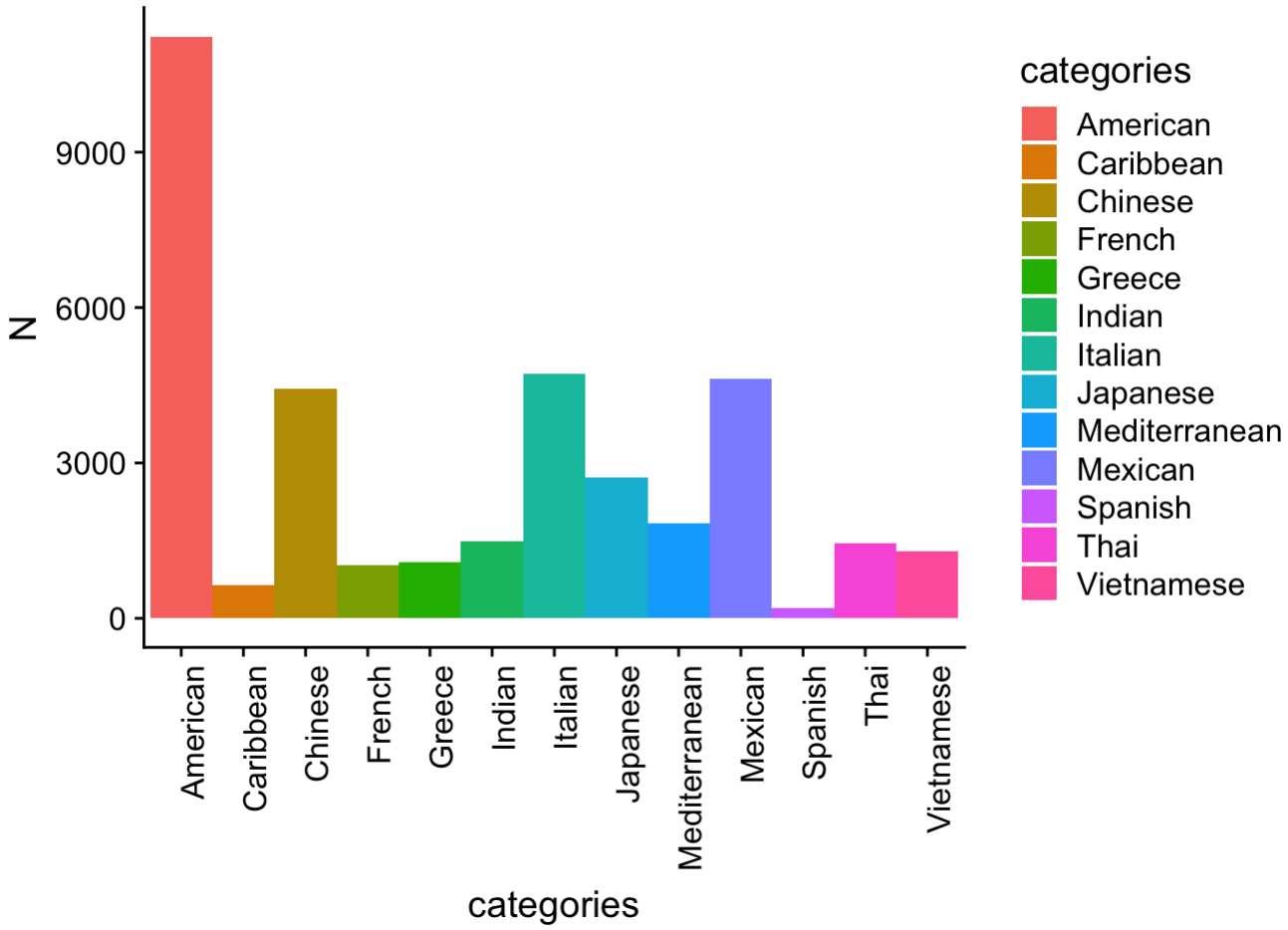
Now, we look at the business ratings and individual reviews, and the distribution is different. In business ratings, 3.5 and 4 stars are the most dominant, while for individual users, they tend to rate restaurants as 4 or 5 stars more often. The average individual rating is higher than the average business stars. Part of the reason may be that in business data, rating can have half scores, while in individual data, scores can only be whole numbers, and customers tend to rated a businee higher in such integer setting. However, such difference is how the rating of business works: individuals' low scores, such as 1 and 2s, averaged out the 4 and 5s, and thus lowered the overall industry mean to 3s and 4s.



Next, we investigated the relationship between number of reviews a restaurant received and its rating on the Yelp. The scatterplot shows no clear pattern that indicates a positive or negative relationship between review counts and star rating. However, there are a great amount of reviews clustered around 3 and 4 stars restaurants. In addition, for restaurants that have extremely high review counts, they are generally 3 or 4 stars' restaurants.

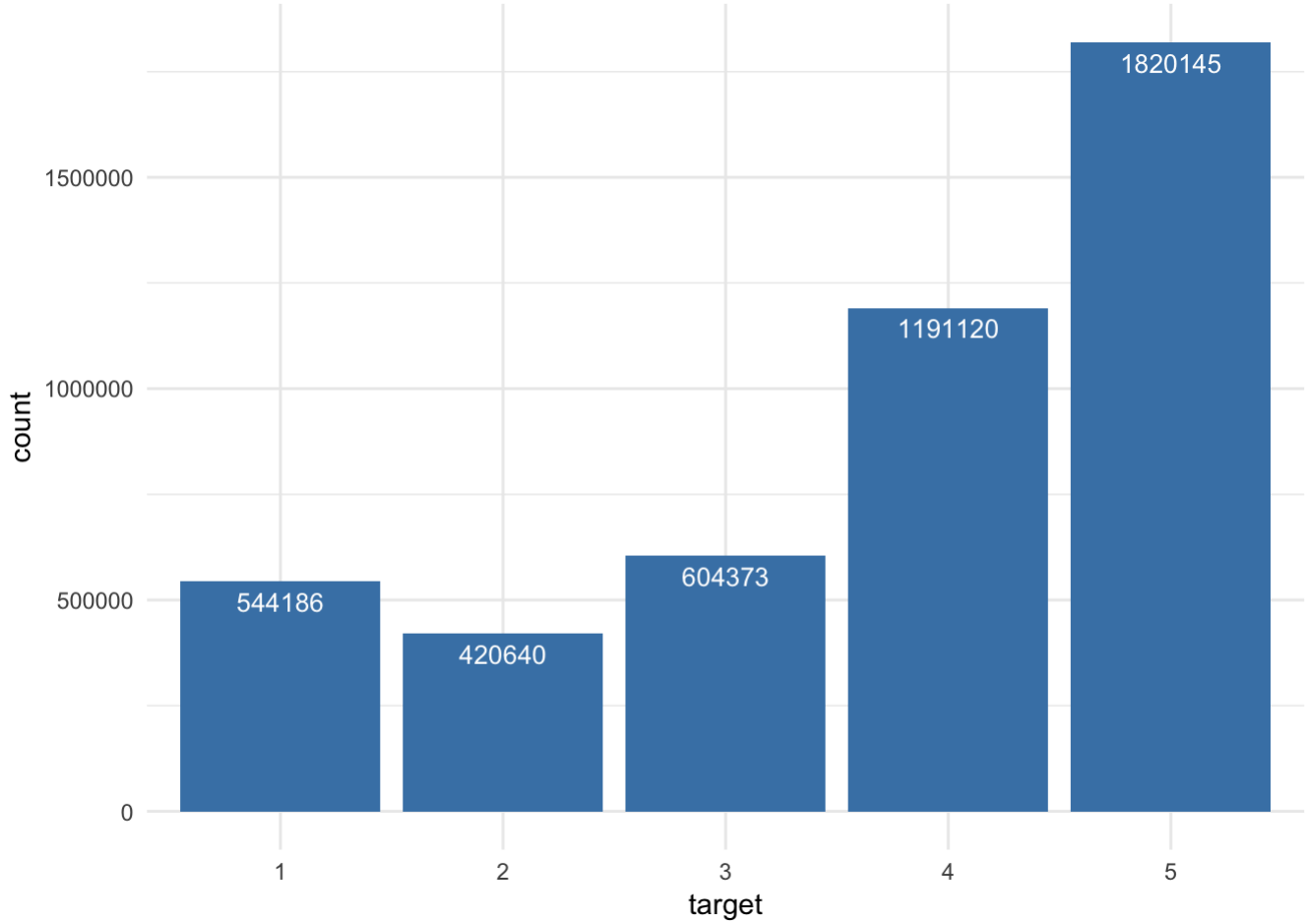


When we examine the review attributes, we found that positive reviews are more useful than negative reviews, because there is more usefulness as rating increases. Beyond that, people find that reviews for higher rating restaurants are cooler, but they think all reviews are of the same level of funny.

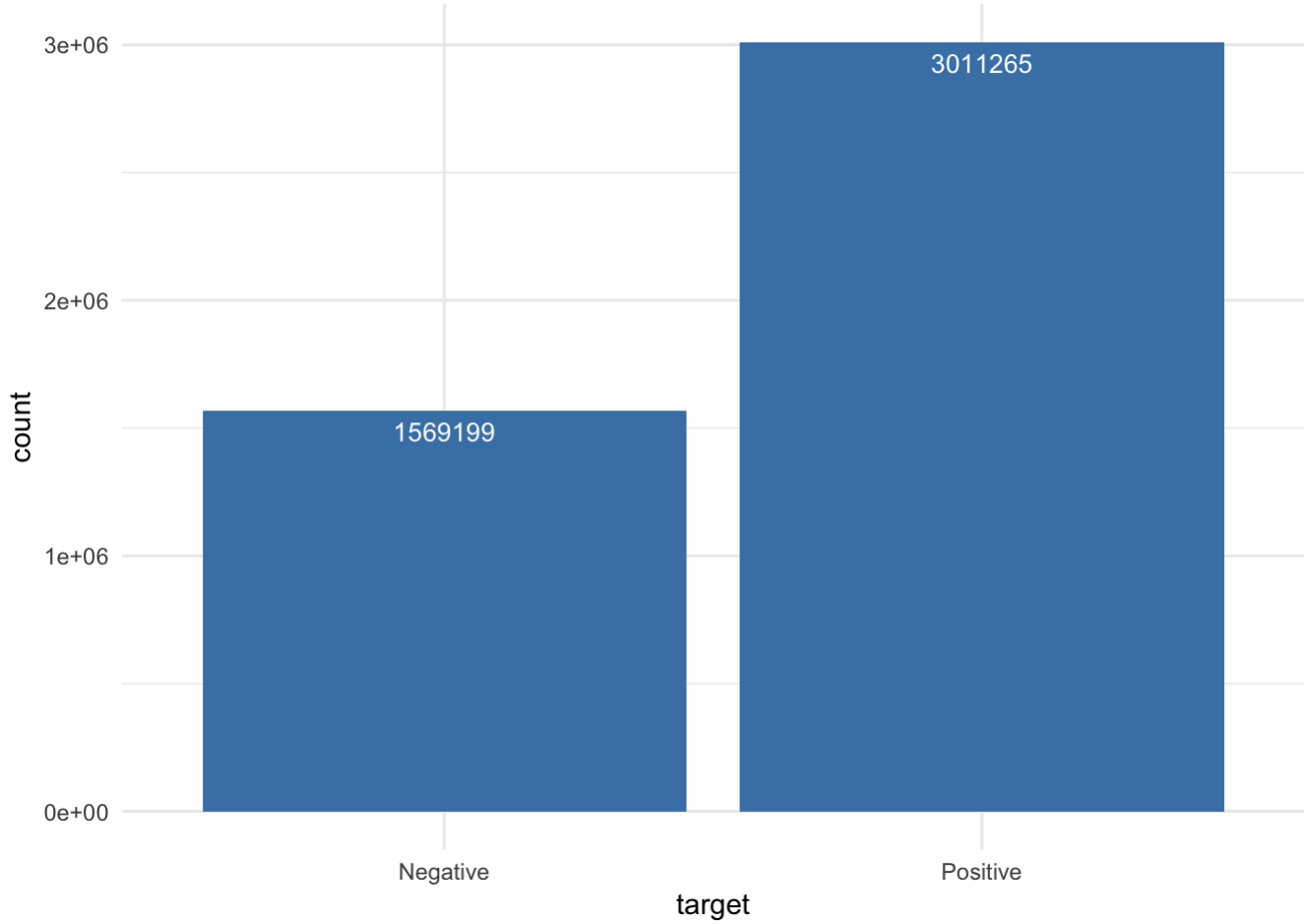


After reviewing the restaurants' data, we decide to find each restaurants' categories by their cuisine type. To do that, we conducted research from the CNN and BBC published articles, and find that there are thirteen most popular cuisine types, generally speaking. After we use regular expressions to grep each cuisine name in the restaurants' dataset, some of them are under several cuisine types, and so the categories were named separately into several rows and concatenated with the rest of dataset. Finally, we want to investigate the top restaurants of each cuisine type in one specific city. Our top restaurants' recommendation is based on our business restaurant dataset's star rating and the amount of review. By selecting the particular city with one cuisine category, we slide the slider bar to display the number of top restaurants with their information, such as name, address, city, state, stars, review counts.

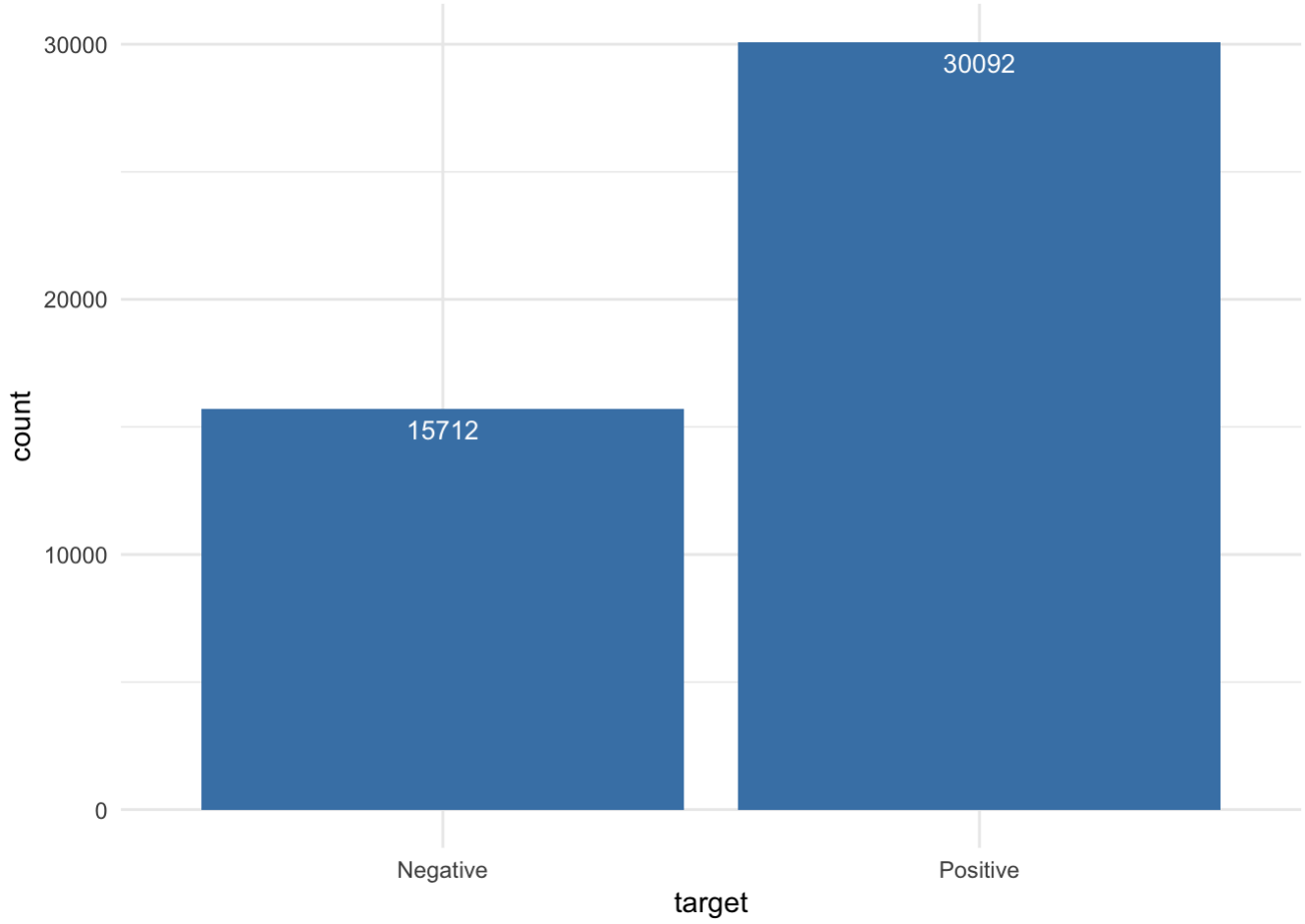
All of this information can additionally be found within the Reporting Engine.



Here we can see the distribution of review ratings grouped by stars. The most important thing to note is the majority of Fours and Fives within the set. There is also a significant number of Ones. However Twos and Threes are underrepresented in this dataset.



Here you can see the distribution of positive to negative reviews, concentrated only on restaurant reviews. As before, there are more positive than negative reviews.



Here is a 1% sample of the restaurant reviews dataset. This subsampling is done to facilitate other operations within the report. Clearly the distribution of reviews has been largely unaffected by the subsampling.

Show 10 **entries** **Search:**

	word	n
1	the	259889
2	and	171341
3	a	125987
4	i	123027
5	to	109593
6	was	93786
7	of	73172
8	it	66080
9	is	63031

	word	n
10	for	58230

Showing 1 to 10 of 10 entries Previous 1 Next

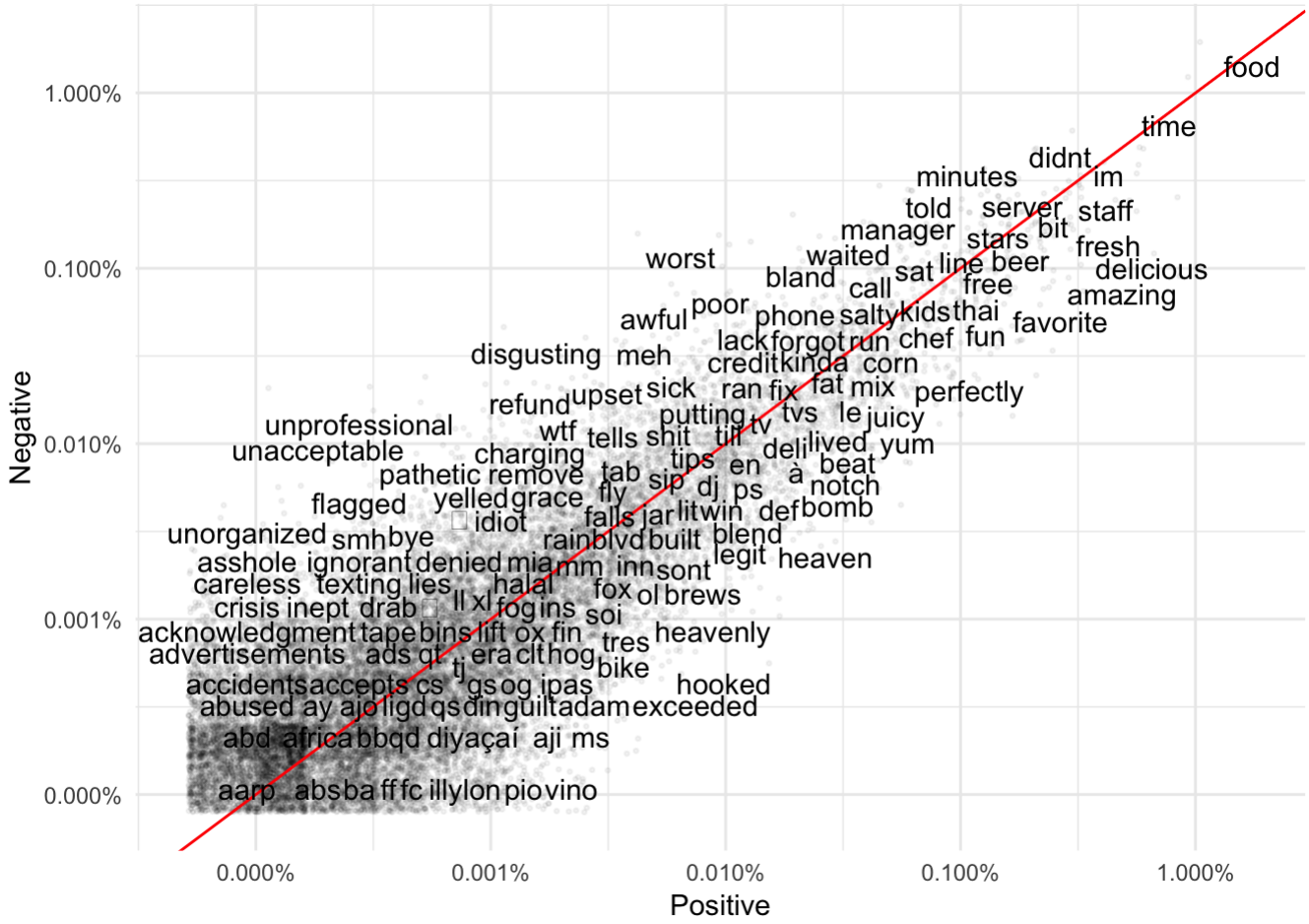
The data has been initially tokenized. Clearly more work needs to be done. These words contain no information with regards to the content of the review. More preprocessing is required.

Show 10 entries **Search:**

	word	n
1	food	32216
2	service	17698
3	time	14523
4	restaurant	9540
5	chicken	8888
6	nice	8814
7	dont	8538
8	delicious	8025
9	menu	7978
10	love	7869

Showing 1 to 10 of 10 entries Previous 1 Next

Now common stopwords have been removed. We are getting closer to the important information contained in the reviews.



Here we can see words which occur most frequently in both positive and negative reviews. Words closest to the red line appear equally in positive and negative reviews, while words farthest from the red line appear primarily in positive and negative reviews.

Show 10 ▾ **entries** **Search:**

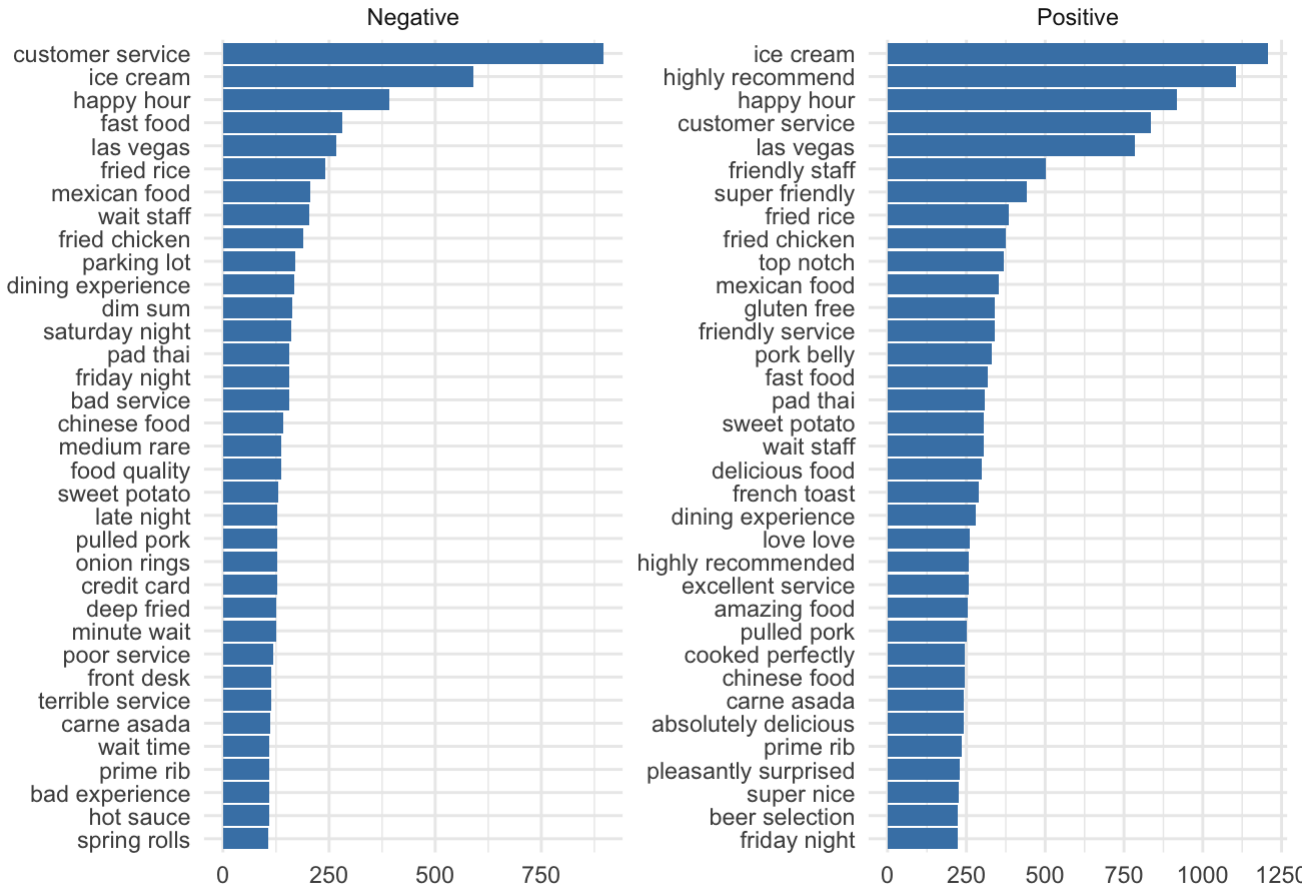
	bigram	n
1	ice cream	1799
2	customer service	1732
3	happy hour	1310
4	highly recommend	1162
5	las vegas	1049
6	fried rice	624
7	fast food	599
8	fried chicken	564
9	friendly staff	562

	bigram	n
10	mexican food	557

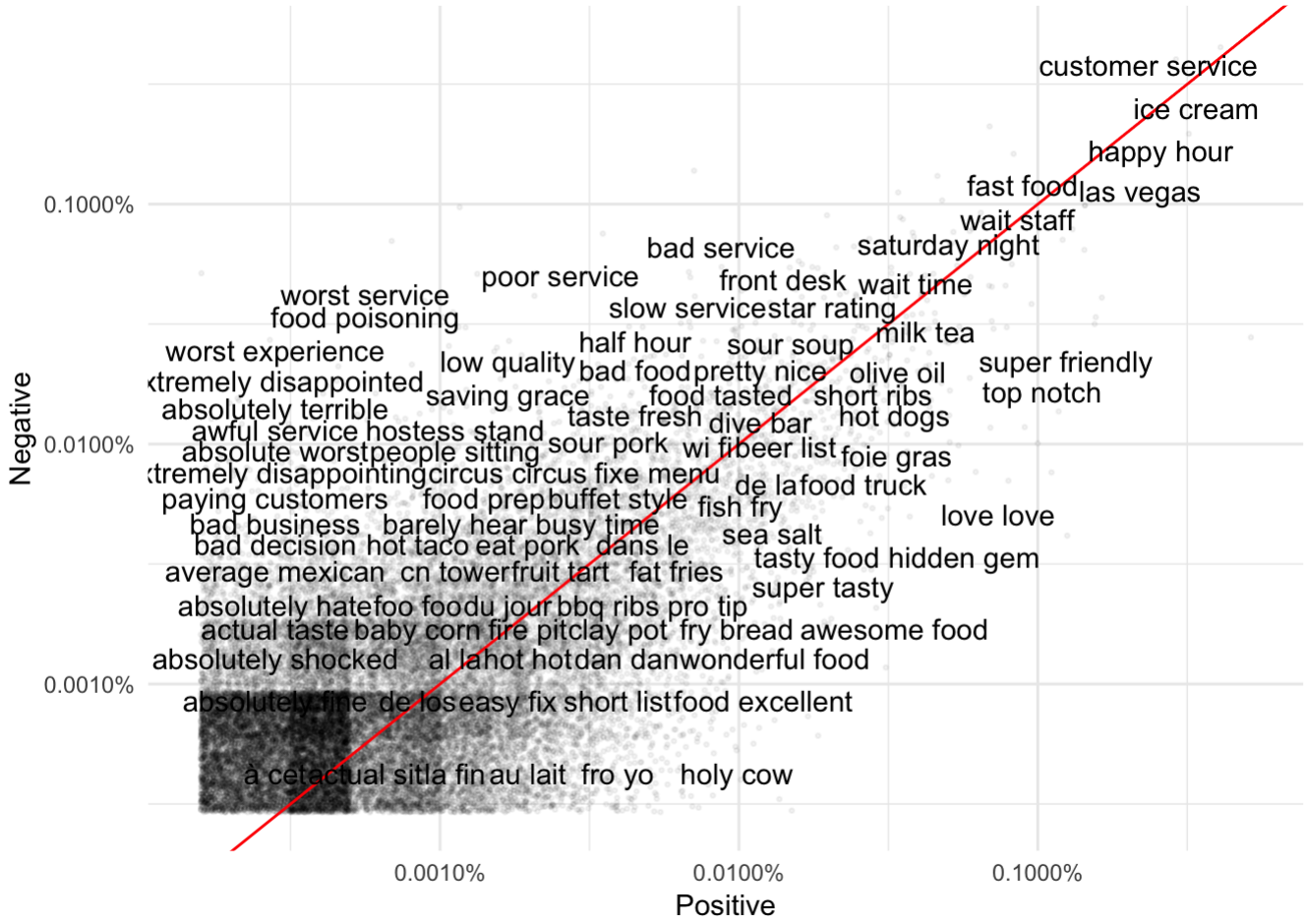
Showing 1 to 10 of 10 entries

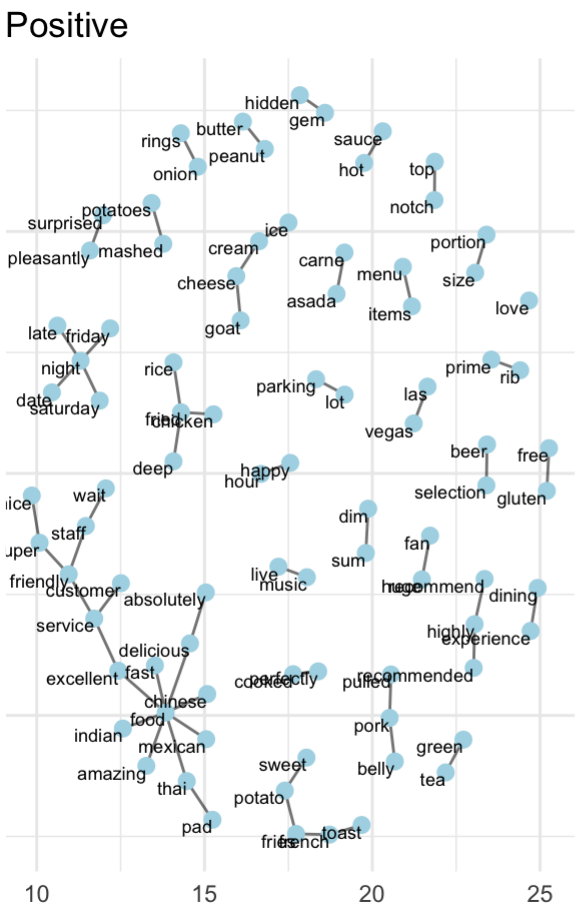
Previous 1 Next

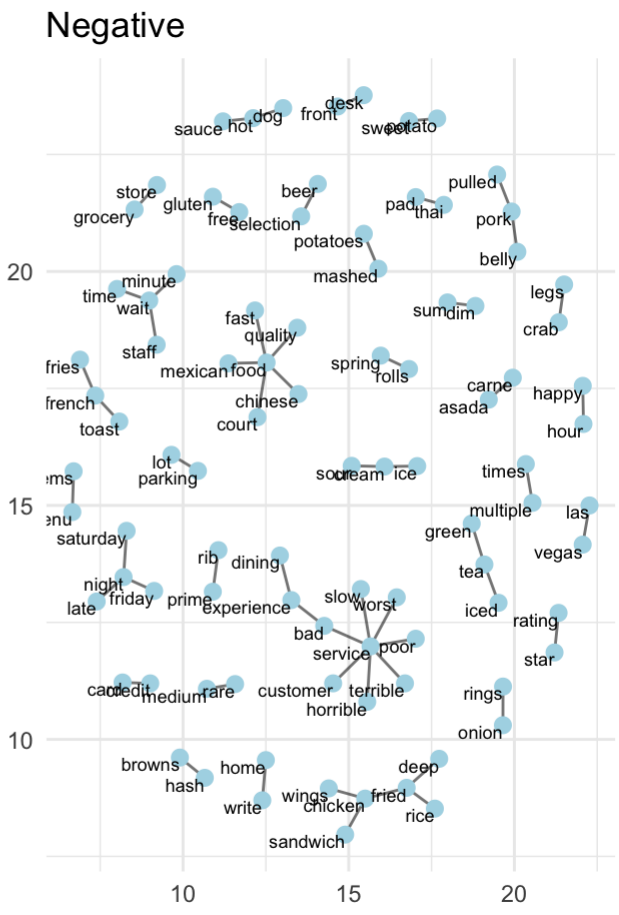
Here are some bigrams after removing stopwords. Note that by using n-grams we can begin to estimate the sequential relationships between words in a review. This will be explored further in the modeling section of the report.



Again, here are bi-gram tokens which occur most frequently in Positive and Negative reviews. There is intuitive meaning behind these words, as the positive column clearly contains words we naturally associate with positivity. The same is true for the negative column.


customer service
ice cream
happy hour
fast food
las vegas
wait staff
saturday night
bad service
poor service
front desk
wait time
worst service
food poisoning
slow service
star rating
milk tea
worst experience
half hour
sour soup
low quality
bad food
pretty nice
olive oil
super friendly
xtremely disappointed
saving grace
food tasted
short ribs
top notch
absolutely terrible
taste fresh
dive bar
hot dogs
awful service hostess stand
absolute worst
people sitting
sour pork
wi fi
beer list
foie gras
xtremely disappointing
circus circus fixe menu
de la
food truck
paying customers
food prep
buffet style
fish fry
love love
bad business
barely hear
busy time
sea salt
bad decision
hot taco
eat pork
dans le
tasty food hidden gem
average mexican
cn tower
fruit tart
fat fries
super tasty
absolutely hate
foo foo
du jour
bbq ribs
pro tip
actual taste
baby corn
fire pit
clay pot
fry bread
awesome food
absolutely shocked
al la
hot hot
dan dan
wonderful food
absolutely fine
de los
easy fix
short list
food excellent
à cet
actual sit
la fin
au lait
fro yo
holy cow
0.1000%
0.0100%
0.0010%
Negative
0.0010%
0.0100%
0.1000%
Positive


Positive
hidden
gem
rings
butter
sauce
peanut
onion
hot
top
notch
surprised
potatoes
pleasantly
mashed
ice
cream
portion
carne
menu
size
cheese
asada
items
love
goat
late
friday
night
rice
parking
prime
rib
date
saturday
lot
las
vegas
deep
happy
hour
beer
free
selection
gluten
wait
staff
dim
fan
sum
friendly
live
music
customer
absolutely
service
delicious
excellent
fast
chinese
food
indian
mexican
amazing
thai
pad
sweet
potato
belly
pork
green
tea
toast
10
15
20
25


Negative
sauce
hot
dog
front
desk
potato
store
gluten
free
grocery
beer
selection
pulled
pad
thai
pork
belly
potatoes
mashed
minute
time
wait
fast
quality
staff
mexican
food
fries
spring
rolls
chinese
court
french
toast
carne
asada
happy
hour
lot
parking
cream
ice
times
multiple
las
vegas
green
tea
iced
rating
star
saturday
rib
dining
night
late
friday
prime
experience
slow
worst
bad
service
poor
terrible
horrible
customer
rare
medium
rings
onion
browns
hash
home
write
deep
fried
chicken
wings
rice
sandwich
10
15
20

Here are network graphs which display common bigram connections between both positive and negative terms. Similar conclusions can be drawn from this graph, as it also displays partially sequential information regarding the reviews.

```
## Using TensorFlow backend.
```

Investigations

To answer the question of whether a business stars can be predicted to a satisfactory level of accuracy, we investigated several models. The final model will be explored in detail, and the unsuccessful attempts will be summarized in the coming paragraphs.

First, a predictive model was fit based on business attributes. Given the size and diversity of review data, it seemed at first desirable to fit a model based on simple features. The chosen attributes are displayed in the business data table at the head of the report. They include categorical variables, such as price range and ambience. They also include binary variables like delivery and wi-fi availability.

There were two problems with this approach. First, because the original dataset was stored in JSON files, many variables were still nested within the characteristic JSON brackets. These needed to be manually extracted using regular expression functions. Second, the distribution of these attributes made them uniquely unsuitable for modeling purposes. There were varying categorical levels with similar meanings, such as “no”, ”NO”, and “ “. Again, this needed to be manually remedied. Second, the levels of missingness far exceeded levels appropriate for missing data imputation. This is possibly due to the mixed nature of the dataset. For example, a hair salon will never have a liquor license or offer delivery. However, even after filtering the dataset down to only restaurants, the levels of missingness were still exceedingly high. To solve this problem, an additional factor was added to every variable to connote missing information.

After all of this, a Naïve Bayes Classifier was fit on the data, which only achieved a 47% accuracy. We concluded that due to the incomplete nature of the data, a different approach was necessary.

Next, we turned to the reviews dataset. Due to the size of the dataset (~5,000,00 reviews) sub-sampling was necessary to reduce training size. We also deemed it wise to restrict our training and test sets to restaurants. This will simplify the information contained in the reviews by reducing the complexity of the vocabulary. This is a limitation, however it can be removed whenever this issue is revisited with more computing power. To sub-sample our data, we imposed the restaurant restriction, and additionally only sampled restaurants from Las Vegas. This will be explored in more detail in the limitations section. From this subset of the data, we drew ten percent to serve as our training and test set.

The next step is the pre-processing of the text data. This will be explained briefly, as there could be many reports which only focus on these steps. First, it is necessary to convert our text to vector representation. There are various methods for dealing with this, but for simplicity we opt to use Stanford NLP’s Global Vectors for Word Representation. According to their website “GloVe is an unsupervised learning algorithm for obtaining vector representations for words”. This is useful for two reasons. First, it expresses text in a lower dimension than the number of unique words. Second, through these representations word similiarity can be assessed.

Our text was first converted to vector form, with a maximum unique word count of 20,000. These will be the features of our training set. Then, each review was expressed as a sequence of values, with each value corresponding to a unique word. Then, reviews shorter than 200 words were “padded” with zeroes, to ensure uniformity of all observations.

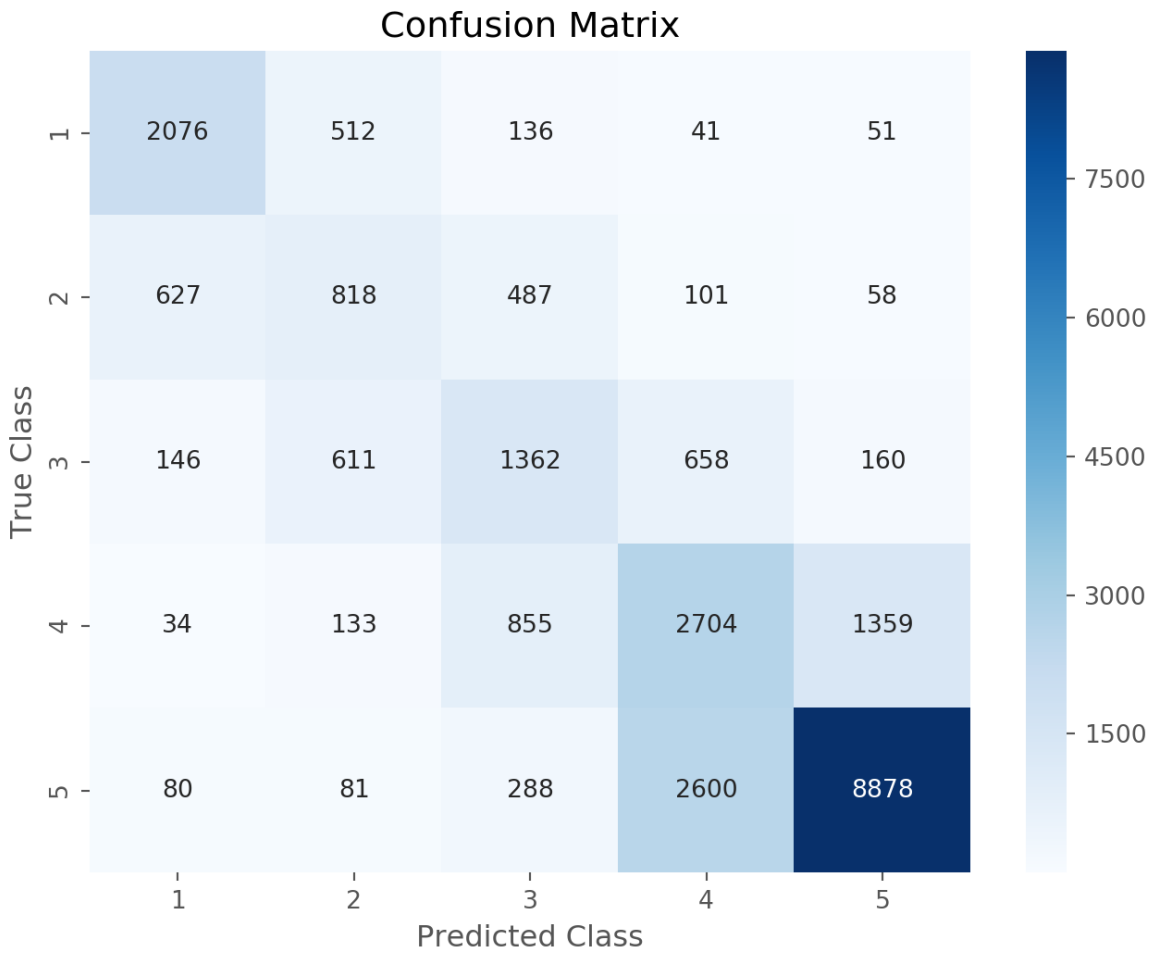
Three models were explored when fitting the model on text data . The results will be shown for our superior model, but I will explain all three. First, we attempted once again to fit a naïve Bayes classifier. However, since the ability of this classifier to utilize sequential data is incomplete at best, single words were considered as inputs. This model achieved an accuracy of 17%, which amounts to worse than a random guess. This model will not be explored further. Next we attempted to implement a Support Vector Classifier with similar inputs. This model achieved an accuracy of 58%, which is certainly an improvement. While acceptable, we still desired to utilize the information contained within the sequential portion of the data.

To do so, we utilized a shallow Recurrent Neural Network. RNN's are well suited to the task of text classification, as they take a sequence of variables as an input. As such, they are able to incorporate past details in every prediction. The first layer of our Neural Network used the afore-mentioned GloVe embedding to transform the sequences of text. Next, we used a dropout layer of 50% to combat overfitting. Then we utilized a Long Short Term Memory layer followed by a Gated Recurrent Unit layer. These are currently the most popular recurrent neural network mechanisms in the field of natural language processing and text classification.

```
## ('Accuracy:', 0.8605324643810041)
```

Results

Finally, let's discuss the results of our classification. The history of model training has been plotted with regards to classification accuracy and categorical cross-entropy. An overall test accuracy of around 87% is achieved, which vastly outperforms all other models. To see a more nuanced accuracy metric, investigate the confusion matrix. Clearly, the categories of Five and One are predicted with a relatively high accuracy. Categories Two, Three, and Four are slightly less accurate. This is a model limitation, but it is one we believe will be difficult to overcome, due to the subjective nature of reviews. It is often difficult to quantify what the difference between a Three and a Four is. To see this explanation in practice, let's investigate individual cases of misclassification.




```
## ('Predicted', 3)
```

```
## ('Actual', 4)
```

```
## The service was pretty quick. We ordered the 20 pack with 2 fries. The bottom buns were soggy
----otherwise, the burgers were pretty good. Would come again if in the area.
```

First, we see a case where the prediction does not match the actual star value. Note the combination of positive and negative text. This is exactly what makes it so difficult to classify reviews of a mixed emotion.

```
## ('Predicted', 5)
```

```
## ('Actual', 1)
```

```
## 四川担々麺がピリ辛で美味しかっただけど、友達はあごの手術を受けたせいで口を大きく開けないんだ。そのため「锅贴」っていうそもそも焼き餃子みたいなやり方で、型は細長いものを注文した。けど、持ってきたものは水餃子をちょっとだけ焼いたもので、美味しくないし、友達が食べる時もめちゃくちゃ不便。店員に聞いたら:あなた以前食べたもんはどうでもいい、うちはこういうやり方だ！ってその態度にびっくりしました。この店を避けた方がいい。
##
## 朋友下颌骨受伤不能张大口吃东西，所以选了长形的锅贴方便食用，结果端上来一盘水饺煮好后微煎一下的东西，做不出锅贴就别拿煎饺糊弄吧，服务员理直气壮说：我们这儿的锅贴就是这样！态度差到令人咂舌。
```

Next, we look at a review which was predicted to be a Five, but in actuality was a One. We immediately see the problem. This review is in Japanese. Since our model was not trained on Japanese characters, this will be input as a string of 200 zeroes. Clearly our model has language based limitations. This can be solved by training on a larger dataset.

```
## ('Predicted', 2)
```

```
## ('Actual', 5)
```

```
## Pro:
## 1. So many options to choose from. Knowing what you'd like to eat prior to the arrival would help a lot. Especially if you have very limited time.
##
## 2. Food tastes great.
##
## 3. Size per portion is just enough. Some might need more than 1 order.
##
## Con:
## 1. Cleanliness needs some more attention. Especially on dining table. I had to ask a bottle of cleaning chemical to the cashier since the table was so dirty & I had a 2 YO daughter with us. She gave us a cleaning rag instead.
##
## 2. Lacking of basic knowledge of cleanliness/sanitation.
## What concerned me the most, is that when she gave me the cleaning rag, it was passed over our food. That's a huge No No in food industry.
##
## 3. Man's restroom was not lockable. I accidentally opened the door when a kid was using it. I suggested the kid to lock the door when he uses the restroom. Yet, when I tried to lock the door, it wasn't successful.
##
## What if there was a pedophile taking advantage on that kid since you fail to maintain the basic safety toward your patrons. God forbids on that. Restroom need to be check hourly & documented for the cleanliness, functions, & its safety.
##
## 4. Restroom was not cleaned.
```

Finally, we look at a review which was predicted to be One, and was actually a Five. We can see certain positive words. However, this review is overwhelmingly negative and repeats concerns of cleanliness numerous times. This may well be a case of user error. User misclassification is one area that we intended to explore, and so this is a potential avenue for future exploration.

To expand upon that, a future application of our model as displayed above may be to identify misclassified reviews. The gravity of the misclassification is extreme. In the future, reviews which display such a grave error might be flagged as misclassified or fraudulent. This is the direction that we hope to move in with our project.

Assumptions

Let's discuss the assumptions we made in answering this question. We assumed that 1% of the data and the data provided by yelp is a great representation of all the restaurants in Canada and America. This was a necessary assumption, as training a model on the entirety of 5,000,000 reviews would take an inordinate amount of time on a student's laptop. Additionally, this is reasonable since most people will react in a similar way towards the restaurants they like and those that they hate. Additionally, while one may question the validity of such an assumption, the results demonstrate the soundness. We obtained a really good result from our last model with an accuracy of around 87% and we have tested it using randomly and independently collected reviews online from yelp and in most cases it reacted appropriately. Clearly this assumption was necessary and didn't compromise the overall results of our model. In the future it may be desirable to retrain the model and obtain more appropriate values for training and test error.

An additional assumption inherent to our model was the validity of each review's actual classification. As we saw in the few cases where our model incorrectly predicted the level of stars, there is a non-trivial number of reviews which may have been misclassified by the user. In these cases, manual inspection is required. But for the scope of this project we do not have the manpower or time to carry out such an investigation.

Limitations

In our project, one of the most significant limitations will be the hardware problem. Our computers are not able to run the full 100% data. It will crash R Studio due to the fact that the memory is not big enough, and it will be functionally useless if used in a different language and software due to the sheer size of the dataset. For this project we were forced to sample only 1% of the data from the original dataset. Using this subsampled data, it still takes a long time to run the code and the computers' temperature increases dramatically. Although the sample size is only around 45k, the number of rows of the dataset after tokenization for a unigram model still exceeded 1 million rows. Since this is only a sample, there might be some bias and we will not be able to collect all the information in its most complete form. if we had a high performance computer, or even access to a vGPU of some kind we would have a more accurate prediction for the model. With 1% of the data, we are able to achieve an accuracy of approximately 87% so with 100% of the data, I believe we could have an accuracy that will be even better.

Another limitation will be that the provided review dataset only has restaurants in several cities. There are no big cities like Los Angeles and New York and no small cities such as Reno and Champaign. So it might not represent all of the restaurants for Canada and America. We should have a larger datasets with restaurants all over the place. This is not a limitation we can overcome independently, but rather one which must be addressed by Yelp in another iteration of this data competition. Yelp is most likely not going to do this. There is no incentive for them to make this information more readily available, so we will have to make do with what data we are given.

Future Progress

In the future, we could have a lot more to do with the current datasets. We are only able to input the restaurant categories for the wordcloud. We might be able to make an app and users can input the restaurant name and see the most used words for that restaurant using unigram, bigram and trigram tokens so they are able to have a

glimpse about the restaurant's condition and their popular food. We could also incorporate some other languages when training our neural network model which was misclassified and shown in the presentation slide.

The nature of our future investigations are all directly connected to the limitations and assumptions of our project. Obviously given the opportunity we would like to train on data from multiple business types in multiple cities and multiple languages. This may decrease the accuracy of our model in the short term due to increased variation in inputs. However it will broaden the applicability of our model. In our estimation, a slight sacrifice in accuracy is more than worth the increase in utility.

Additionally, new methods and algorithms in neural networks and natural language processing are introduced constantly. The state of the industry may change in a way that can improve our model.

In conclusion, the state of this model and project is obviously in flux. It is a work in progress, and should continue to evolve pending the continuing advancements of natural language processing and neural networks. We will re-evaluate this model as times change. In the present, however, the classification of Yelp reviews undertaken in this project can be considered a moderate success.

Citations

1. *Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation*
2. *https://www.cnn.com/travel/article/world-best-food-cultures/index.html (https://www.cnn.com/travel/article/world-best-food-cultures/index.html)*
3. *https://www.bbc.com/food/cuisines (https://www.bbc.com/food/cuisines)*
4. *https://edav.info/leaflet.html (https://edav.info/leaflet.html)*