Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

**How to improve restaurant reputation even just one star using Yelp user reviews !**
**" The path from being normal to good, and good to perfect !"**

Just to recap the previous step : I am working on 2 sub datasets "review" and "business". I did some data wrangling on "business" dataset and made the new sub dataset "restaurant_eng", which just includes restaurants in few English-spoken countries - USA, Canada, Britain. I also did some initial data wrangling on "review" dataset and made the new sub dataset "review_restaurant_eng" which just consists of users' reviews of restaurants in some English-spoken cities.

After data wrangling, now it's time to do some exploratory data analysis. I separated this part into two main sub parts : 1- EDA on "restaurant_eng" 2- EDA on "review" dataset.

***1 - EDA on "restaurant_eng"***

In order to do EDA, I checked the columns of datasets once again. The columns of "restaurant_eng" dataset that I think may give me some useful insight in the path of this project are: ***City, State, Stars, Review_count, Country, Food_type***.

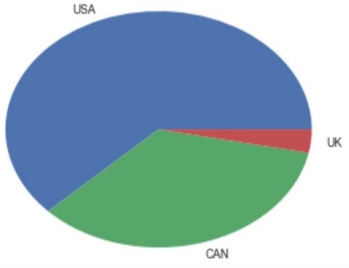
Most seen "Cities" in the restaurant dataset

| | |
|---|---|
| Toronto | 6424 |
| Las Vegas | 5533 |
| Phoenix | 3361 |
| Montréal | 2981 |
| Charlotte | 2309 |
| Pittsburgh | 2039 |
| Scottsdale | 1334 |
| Cleveland | 1278 |
| Edinburgh | 1246 |
| Mississauga | 1223 |

Most seen "State" in the restaurant dataset

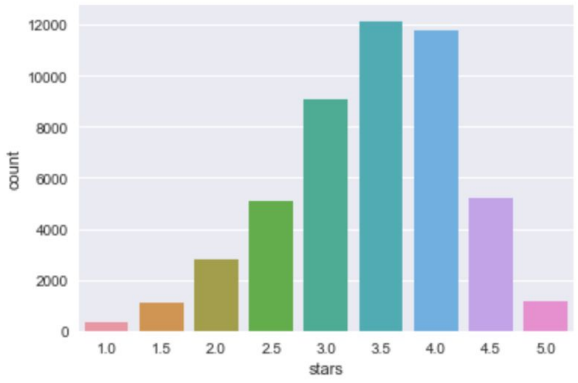
| | |
|---|---|
| ON | 12291 |
| AZ | 9802 |
| NV | 6680 |
| QC | 4528 |
| OH | 4521 |
| NC | 3622 |
| PA | 3396 |
| WI | 1387 |
| EDH | 1210 |
| IL | 577 |

As you can see most of the reviews came from Canada and US cities like Toronto, Las Vegas, Phoenix, Montreal and etc. Ontario, Arizona, Nevada, and Quebec have the most reviews amongst other states. Amongst top 10 cities and top 10 states, there is just one city , Edinburg, and one state , EDH, from Britain. The distribution of restaurants in different countries is shown below:



Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

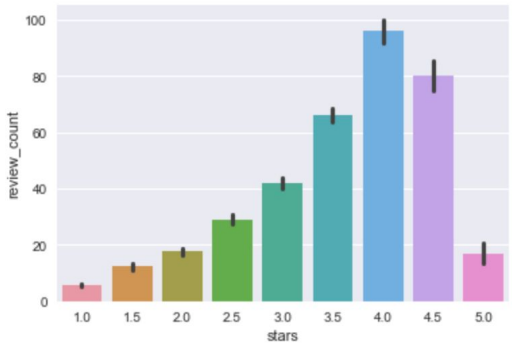
Exploration on "Stars" column:




According to the left chart, "restaurants ratings" has left-skewed distribution.
Most restaurants have either 3.5 or 4 average ratings.
Also based on this chart, users tend to give more perfect rating (4.5 , 5) than poor rating (1 , 1.5).
Later, I will dig more into rating column to see whether other factors like "country" or "type of food" have impact on this general distribution or not.

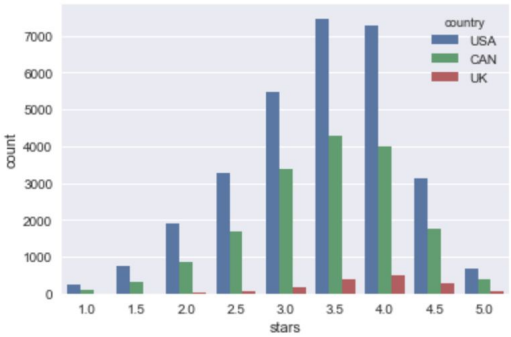
| stars | |
|---|---|
| 1.0 | 5.620896 |
| 1.5 | 12.256198 |
| 2.0 | 17.632184 |
| 2.5 | 28.971750 |
| 3.0 | 41.974925 |
| 3.5 | 66.016231 |
| 4.0 | 95.941821 |
| 4.5 | 80.192841 |
| 5.0 | 16.641093 |

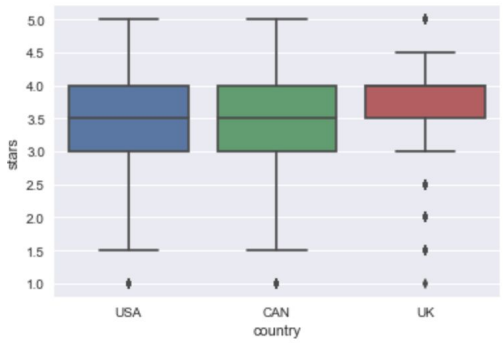
average "review_count" across different "star"




| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| country | | | | | | | | |
| CAN | 16819.0 | 3.426333 | 0.769820 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| UK | 1549.0 | 3.765655 | 0.686447 | 1.0 | 3.5 | 4.0 | 4.0 | 5.0 |
| USA | 30196.0 | 3.395069 | 0.803634 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |

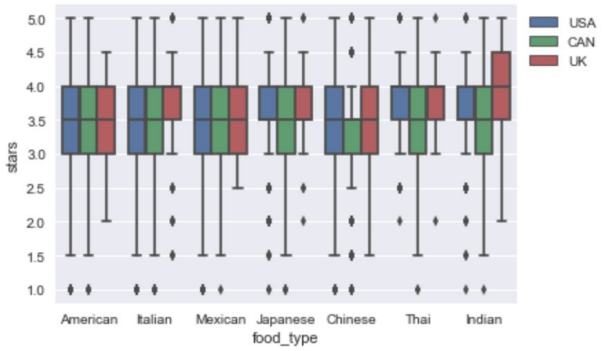
According to the above table (and below chart), it seems that people in UK gave better ratings to restaurants than USA and Canada. In the next step, I will explore more in depth to see ***whether "country" affect the restaurant rating or not. (Is it statistically significant or not)***



Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

| food_type | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| American | 8401.0 | 3.402928 | 0.713942 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Chinese | 3318.0 | 3.286468 | 0.689724 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Indian | 1209.0 | 3.579404 | 0.690748 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Italian | 3788.0 | 3.478221 | 0.773531 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Japanese | 1972.0 | 3.582657 | 0.662095 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Mexican | 3604.0 | 3.380827 | 0.768168 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |
| Thai | 989.0 | 3.582406 | 0.707131 | 1.0 | 3.0 | 3.5 | 4.0 | 5.0 |




Different types of food have relatively similar ratings distributions (in general and across different countries as well). They also have very close average ratings (range : 3.29 - 3.58) . In next step, I will also check ***whether the difference of ratings across different type of food is significantly difference or not. (it also can be checked whether "country" factor can impact "type of food" rating or not.***

Exploration on "review_count" column :

| | |
|---|---|
| count | 48564.000000 |
| mean | 60.896611 |
| std | 155.282867 |
| min | 3.000000 |
| 25% | 7.000000 |
| 50% | 19.000000 |
| 75% | 56.000000 |
| max | 7361.000000 |

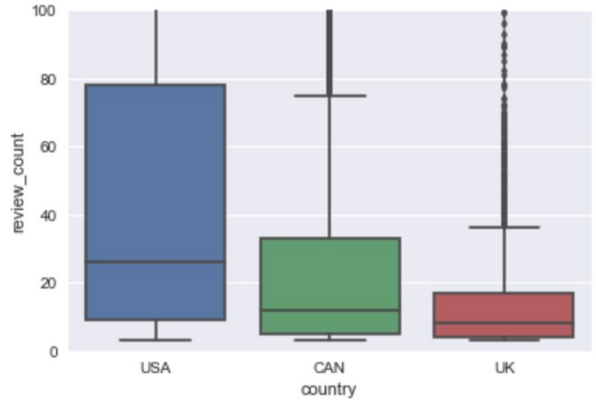
"review_count" statistics

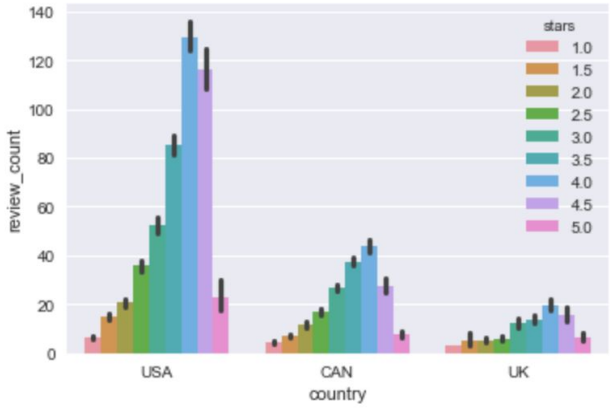
There is no missing value in review_count column. In other words, each restaurant has at least 3 reviews.
75% of all 48564 entries (restaurants) have 7 or more reviews. In the next steps, I am exploring this column , "review_count", to check whether there is a significant difference in review_count across different countries and types of food or not. This exploration can be very insightful since it can show the level of interaction between customers and restaurants across different countries and types of food.

| country | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| CAN | 16819.0 | 30.961532 | 59.316654 | 3.0 | 5.0 | 12.0 | 33.0 | 1953.0 |
| UK | 1549.0 | 14.883150 | 20.463048 | 3.0 | 4.0 | 8.0 | 17.0 | 266.0 |
| USA | 30196.0 | 79.930686 | 189.287147 | 3.0 | 9.0 | 26.0 | 78.0 | 7361.0 |

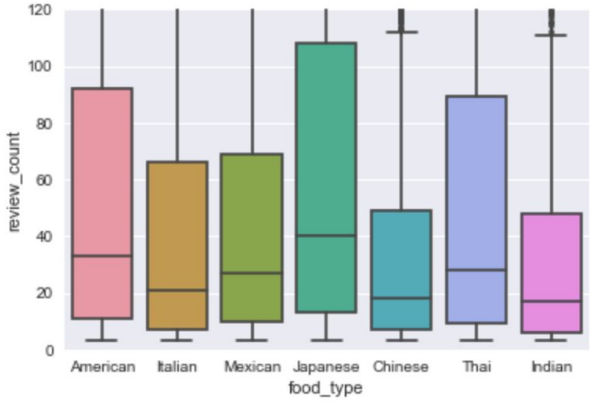
"review_count" across different countries

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : “Improving Restaurant Reputation Using Yelp User Reviews”
Part 2: Storytelling

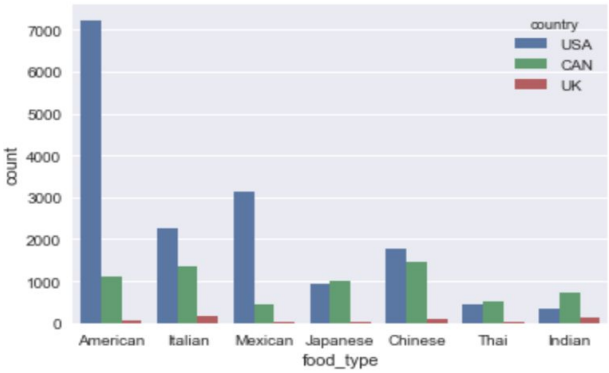


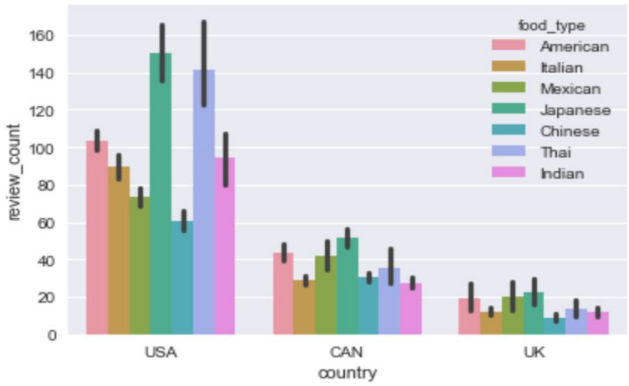




It is clear that, American, Canadian and then British people, respectively, write more reviews for restaurants on Yelp. While all of these three countries have very similar distribution of number of reviews across restaurants’ rating.




In general, restaurants with Japanese, American and Thai foods have the highest average number of reviews while restaurants with Chinese and Indian foods usually get the least reviews.

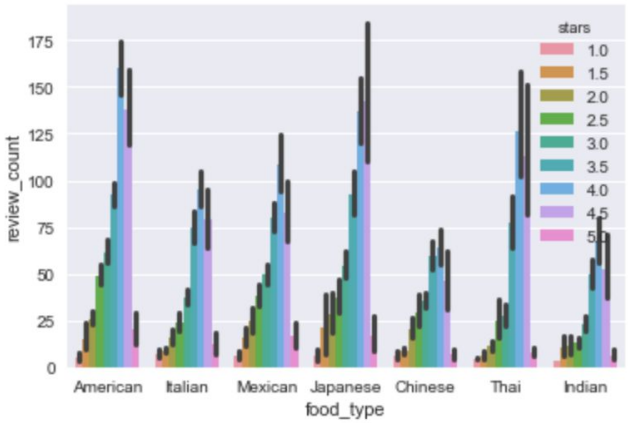



According to the above charts, the distribution of number of reviews across types of food in these three countries are slightly different, especially about Mexican food which in US this type of food has one of the lowest average “review counts” while in Canada and Britain, Mexican food has one of the top average “review_count”.

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

Finally, different types of food have very similar distributions of "review_count" across "rating". In other words, it does not matter what the type of food is, the restaurants with average rating of 4 and 4.5 star have the highest number of review counts compared to the restaurant with very high and very low ratings.

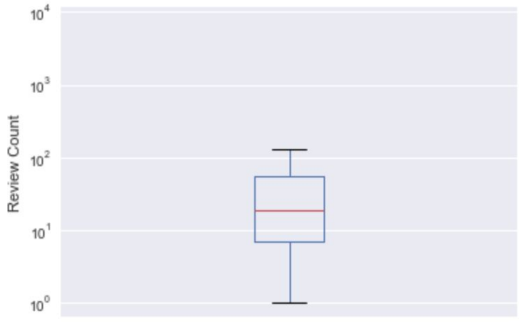


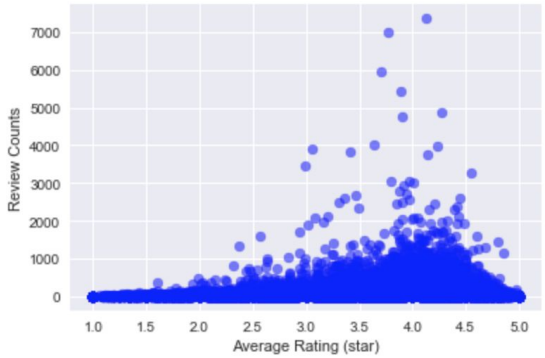

### *2 - EDA on "review_restaurant_eng"*

In order to do EDA on "review_restaurant_eng" dataframe, first I checked several rows of this dataframe to think of what possible columns worth exploration:

| | review_id | user_id | business_id | stars | date | text | useful | funny | cool |
|---|---|---|---|---|---|---|---|---|---|
| 0 | v0i_UHJMo_h... | bv2nCi5Qv5v... | 0W4lkclzZTh... | 5 | 2016-05-28 | Love the st... | 0.0 | 0.0 | 0.0 |
| 1 | vkVSCC7xljj... | bv2nCi5Qv5v... | AEx2SYEUJmT... | 5 | 2016-05-28 | Super simpl... | 0.0 | 0.0 | 0.0 |
| 2 | n6QzIUObkYs... | bv2nCi5Qv5v... | VR6GpWIda3S... | 5 | 2016-05-28 | Small unass... | 0.0 | 0.0 | 0.0 |
| 3 | MV3CcKScW05... | bv2nCi5Qv5v... | CKC0-MOWMqo... | 5 | 2016-05-28 | Lester's is... | 0.0 | 0.0 | 0.0 |
| 4 | IXvOzsEMYti... | bv2nCi5Qv5v... | ACFtxLv8pGr... | 4 | 2016-05-28 | Love coming... | 0.0 | 0.0 | 0.0 |

The columns that my project main purpose is relied on are: **"text"**, and **"stars".** However other columns such as **business_id, date, useful, funny,** and **cool** can give us some useful insights as well.
One of the initial area to explore can be the number of reviews for different restaurants (business_id) and its distribution across average rating, the related charts are shown below:



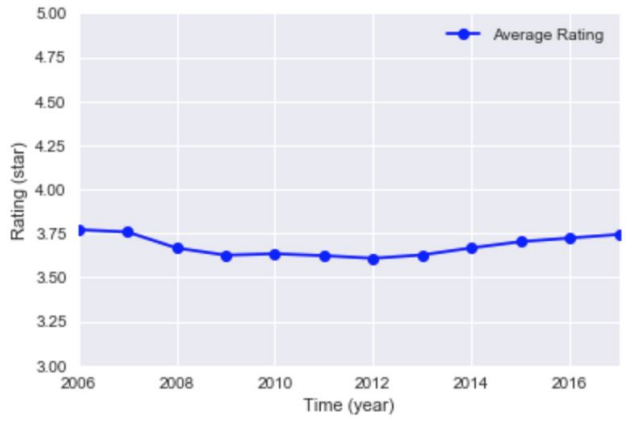

The chart on the left shows that at least 30% of entries have more than 10 user review. Review counts across average rating has left-skewed distribution with the picked around 4 star.

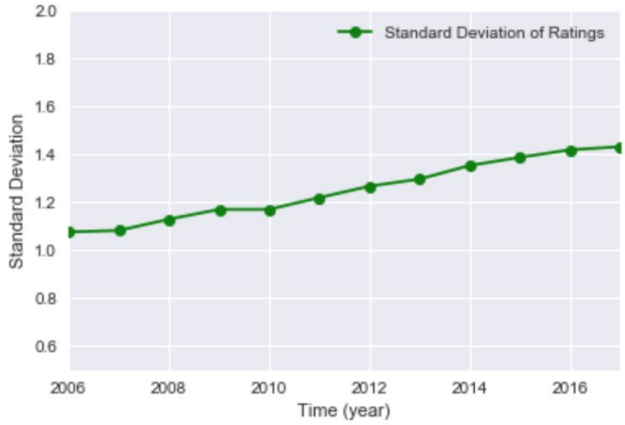
Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : “Improving Restaurant Reputation Using Yelp User Reviews”
Part 2: Storytelling

| date | count | mean | std |
|---|---|---|---|
| 2004 | 8 | 4.125000 | 0.640870 |
| 2005 | 470 | 4.010638 | 0.896507 |
| 2006 | 2846 | 3.770555 | 1.076212 |
| 2007 | 12137 | 3.757683 | 1.082474 |
| 2008 | 32252 | 3.666749 | 1.129546 |
| 2009 | 58544 | 3.626144 | 1.170284 |
| 2010 | 110318 | 3.634167 | 1.170615 |
| 2011 | 170275 | 3.623239 | 1.219496 |
| 2012 | 202222 | 3.608084 | 1.267207 |
| 2013 | 268200 | 3.627405 | 1.296896 |
| 2014 | 389124 | 3.667397 | 1.353523 |
| 2015 | 515831 | 3.702470 | 1.387453 |
| 2016 | 580414 | 3.723277 | 1.418872 |
| 2017 | 615094 | 3.743738 | 1.431998 |

Average and Standard Deviation of restaurants‘ ratings through years

There are not enough entries in the first two years (2004, 2005), so let's just look at the "mean" and "standard deviation" of the last 12 years (2006-2017) :




According to the above table and charts, we noticed an interesting fact and that is through time restaurants' ratings tend to be distributed around the same average **(around 3.65)** every year, however by passing time the rating distribution get more divers **(standard deviation rose almost 32 percent)**.
In other words, restaurants tend to get more diverse ratings by passing time or users tend to be more attentive when it comes to give ratings to restaurants.
This finding can be a good motivation for restaurants' owners to pay attention to their customers' reviews and ratings and related analyses.

Now let’s check whether reviews’ features (cool, funny, and useful) can tell us something about the review rating or not:

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

| | funny | | | cool | | | useful | | |
|---|---|---|---|---|---|---|---|---|---|
| | count | mean | sum | count | mean | sum | count | mean | sum |
| stars | | | | | | | | | |
| 1 | 341467 | 0.636963 | 217502.0 | 341467 | 0.257079 | 87784.0 | 341467 | 1.364021 | 465768.0 |
| 2 | 289638 | 0.588614 | 170485.0 | 289638 | 0.382864 | 110892.0 | 289638 | 1.311154 | 379760.0 |
| 3 | 417075 | 0.523031 | 218143.0 | 417075 | 0.583691 | 243443.0 | 417075 | 1.148570 | 479040.0 |
| 4 | 809859 | 0.481549 | 389987.0 | 809859 | 0.740233 | 599484.0 | 809859 | 1.170537 | 947970.0 |
| 5 | 1099695 | 0.327923 | 360615.0 | 1099695 | 0.559142 | 614886.0 | 1099695 | 0.933763 | 1026854.0 |

Based on the above table, it seems that ***reviews with lower rank were found more funny to the readers than review with higher rate***. Also, ***reviews that have higher rank (4,5) are found more cool by other readers than reviews with lower rates (1 and 2 star)***.

Finally, it seems that there is a negative correlation between review rating and being recognized useful by readers. In other words, ***reviews with lower ratings seem to be more useful for other readers***.

Now let's dive into the most important column of this dataset for my project : **"text"**
In order to do some analysis on text, first I need to apply some text pre-processing techniques. Here it is a sample of "text" entry in the "df_review_restaurant_eng" dataset:

```
"Super simple place but amazing nonetheless. It's been around since the 3
0's and they still serve the same thing they started with: a bologna and
salami sandwich with mustard. \n\nStaff was very helpful and friendly."
```

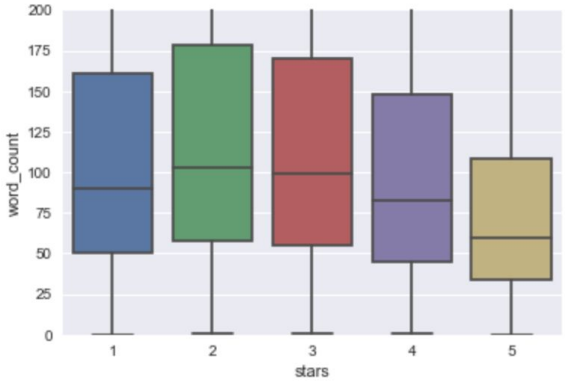
Lower casing of all words, punctuations removal, stopword removal, frequent and rare words removal, spelling correction, and lemmatization are some of the common basic pre-processing techniques that have been done on this column through different stages.
Initially, I added the new column as **"word_list"** which is the list of words of each review's text in the dataset. Along with creating this list, I applied two mentioned techniques (*'lower casing' and 'punctuations removal'* ) in this stage.

One of the most basic features we can extract here is the number of words (new column added: **"word_count"**) in each review and check its distribution across review rating:

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

| stars | mean | count |
|---|---|---|
| 1 | 126.131922 | 341467 |
| 2 | 136.876805 | 289638 |
| 3 | 130.166157 | 417075 |
| 4 | 113.260666 | 809859 |
| 5 | 87.592629 | 1099696 |

word count across rating




It seems users tend to write longer reviews if they give low rate to restaurants compared to cases that high rates are given to a restaurant. In next step, I will check whether this difference is statistically significant or not.

For the next stages, since my machine could not execute some commands for the original dataset, I had to reduce the size of dataset and I did some sampling of original dataset to make it smaller.

```
df_review_restaurant_eng_small = df_review_restaurant_eng.sample(frac=0.005 , replace=False , random_state=1 )
```

| | |
|---|---|
| 5 | 5546 |
| 4 | 3987 |
| 3 | 2125 |
| 1 | 1696 |
| 2 | 1435 |

This new dataset is called "df_review_restaurant_eng_small" and it has around 15K entries distributed with the similar rate of main dataset.

Number of reviews across different rating

After creating smaller dataset, two other preprocessing techniques have been done on the "text" column : "stopword removal" and "lemmatization" , and then I looked for the most frequent words in users' reviews and here was the result :

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

| | |
|---|---|
| food | 11734 |
| place | 9587 |
| good | 9480 |
| get | 8431 |
| go | 7866 |
| order | 7725 |
| great | 6747 |
| come | 6620 |
| service | 6193 |
| like | 6179 |
| time | 5980 |
| one | 4716 |
| try | 4600 |
| make | 4463 |
| back | 4374 |

One of the area that can be explored in the next stage could be "**the possible difference in number of occurance of this words' list in different rating categories".**

The below table which shows the frequency of common words in each rating group may not show us very convincing evidence. However, there are some points that may capture some attentions and can be the foundation of some other focused research on the weight of specific words in each rating group. For instance, the extreme positive words such as "great", "amaze", "best" and "love" have a significant existence in 5 star ratings even compared to 4 star. Presence of word "wait" in the 1 star and 2 star restaurants may show the importance of "Service Time" especially in order to not getting bad ratings for restaurant.

**1 star**

| | Word Freq | Percentage |
|---|---|---|
| food | 1564 | 1.431251 |
| order | 1470 | 1.345230 |
| get | 1368 | 1.251887 |
| go | 1271 | 1.163121 |
| place | 1077 | 0.985587 |
| come | 1026 | 0.938916 |
| time | 888 | 0.812629 |
| service | 841 | 0.769618 |
| say | 837 | 0.765957 |
| like | 786 | 0.719286 |
| us | 769 | 0.703729 |
| ask | 755 | 0.690917 |
| take | 753 | 0.689087 |
| one | 716 | 0.655228 |
| back | 709 | 0.648822 |
| would | 700 | 0.640586 |
| wait | 648 | 0.592999 |
| even | 641 | 0.586593 |
| never | 607 | 0.555479 |
| tell | 578 | 0.528941 |

**2 star**

| | Word Freq | Percentage |
|---|---|---|
| food | 1446 | 1.439536 |
| order | 1151 | 1.145855 |
| get | 1000 | 0.995530 |
| good | 984 | 0.979602 |
| go | 963 | 0.958695 |
| place | 945 | 0.940776 |
| like | 884 | 0.880049 |
| come | 785 | 0.781491 |
| time | 738 | 0.734701 |
| service | 700 | 0.696871 |
| one | 597 | 0.594331 |
| say | 573 | 0.570439 |
| would | 559 | 0.556501 |
| really | 536 | 0.533604 |
| us | 525 | 0.522653 |
| back | 517 | 0.514689 |
| try | 458 | 0.455953 |
| take | 458 | 0.455953 |
| make | 457 | 0.454957 |
| wait | 450 | 0.447989 |

**3 star**

| | Word Freq | Percentage |
|---|---|---|
| food | 1987 | 1.351526 |
| good | 1953 | 1.328400 |
| get | 1467 | 0.997830 |
| place | 1447 | 0.984227 |
| order | 1395 | 0.948857 |
| like | 1253 | 0.852271 |
| go | 1238 | 0.842068 |
| come | 1086 | 0.738680 |
| service | 989 | 0.672702 |
| time | 939 | 0.638693 |
| would | 886 | 0.602643 |
| really | 806 | 0.548228 |
| great | 751 | 0.510818 |
| one | 746 | 0.507417 |
| try | 708 | 0.481570 |
| make | 706 | 0.480210 |
| restaurant | 615 | 0.418313 |
| price | 602 | 0.409471 |
| pretty | 597 | 0.406070 |
| back | 591 | 0.401989 |

**4 star**

| | Word Freq | Percentage |
|---|---|---|
| good | 3436 | 1.434793 |
| food | 2867 | 1.197192 |
| place | 2629 | 1.097809 |
| get | 2317 | 0.967525 |
| great | 2155 | 0.899878 |
| go | 1965 | 0.820538 |
| order | 1862 | 0.777528 |
| come | 1757 | 0.733682 |
| like | 1727 | 0.721155 |
| service | 1569 | 0.655178 |
| time | 1558 | 0.650584 |
| try | 1364 | 0.569575 |
| really | 1329 | 0.554959 |
| one | 1242 | 0.518630 |
| make | 1191 | 0.497334 |
| also | 1104 | 0.461005 |
| would | 1088 | 0.454323 |
| nice | 1071 | 0.447225 |
| back | 1069 | 0.446389 |
| love | 1026 | 0.428434 |

**5 star**

| | Word Freq | Percentage |
|---|---|---|
| food | 3870 | 1.496149 |
| place | 3489 | 1.348854 |
| great | 3258 | 1.259549 |
| good | 2572 | 0.994340 |
| go | 2429 | 0.939056 |
| get | 2279 | 0.881066 |
| service | 2094 | 0.809544 |
| come | 1966 | 0.760059 |
| time | 1857 | 0.717920 |
| order | 1847 | 0.714054 |
| love | 1728 | 0.668048 |
| try | 1649 | 0.637507 |
| make | 1609 | 0.622042 |
| best | 1586 | 0.613151 |
| delicious | 1555 | 0.601166 |
| like | 1529 | 0.591114 |
| back | 1488 | 0.575264 |
| one | 1415 | 0.547042 |
| amaze | 1348 | 0.521139 |
| also | 1210 | 0.467788 |

"20 Common Words in each restaurant rating group"

The above tables just give us very general information, so I will apply two advanced text processing techniques (***Ngrams , TF-IDF***) in order to get better sense of the occurence of words or sequence of words in user reviews.

N-Grams' results can be very tricky and be different by using N (number of words in each group of words), also different results may occur if we either keep or remove stop words. Here are some results:

### *I. N Grams (keep stop words, N = 5)*

Here I used **N=5.** I chose relatively bigger number as N to reduce the impact of keeping stop words and get better insight of the users' reviews.

| *Entire Document (all rating restaurants)* | |
|---|---|
| 5-Word Sequence of Words | Occurrence |
| Can't wait to go back | 97 |
| If you are looking for | 62 |
| The best i've ever had | 60 |
| The quality of the food | 52 |
| We will definitely be back | 48 |
| I can't wait to go | 47 |
| If you're looking for a | 47 |
| This is a great place | 46 |
| I will definitely be back | 46 |
| Nothing to write home about | 45 |
| Just the right amount of | 43 |
| Been here a few times | 42 |
| This is one of the | 42 |
| Is one of the best | 41 |
| At the end of the | 41 |
| To give it a try | 41 |
| Give this place a try | 41 |
| Was one of the best | 40 |
| This was my first time | 40 |
| It was my first time | 39 |

According to above table, we can find some common expressions or equivalent of them in the entire restaurants' user reviews such as ***"Go Back", "If You Are Looking For", "Right Amount", "Can't Wait to Go", "Give It a Try", "My First Time", "The Best"*** .

Although this expressions may not give us obvious and direct insight of reviews, but some interesting points can be extracted from them:

- Some users tell their opinions based on their own **future behavior** (Go back / Not go back)

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : "Improving Restaurant Reputation Using Yelp User Reviews"
Part 2: Storytelling

- Users tend to give their direct **suggestions** to others (If you are looking for)
- Some users show their **excitement** in the reviews (Can't wait to go / The best)
- **First time visits** are very crucial to shape users' minds. (My first time)

Now let's look at 5-Grams of each rating group :

| 1 Star | 2 Star | 3 Star | 4 Star | 5 Star |
|---|---|---|---|---|
| Wanted to like this place | I really wanted to like | Nothing to write home about | This is a great place | Can't wait to go back |
| I will never go back | Wanted to like this place | The food was good but | If you are looking for | The best I've ever had |
| Would not recommend this place | Out of my way to | Go out of my way | Is a great place to | I can't wait to go |
| An hour and a half | Really wanted to like this | The food is good but | The best I've ever had | We will definitely go back |
| We had to ask for | The quality of the food | The quality of the food | I will definitely be back | Was one of the best |
| We will not back | I felt like I was | The food was pretty good | Just the right amount of | I will definitely be back |
| Would never recommend the place | Give this place a try | The best part of the | Can't wait to go back | Can't wait to come back |
| We will never go back | The food was good But | If you are looking for | Been here a few times | If you are looking for |
| We will never eat here again | We were the only ones | Out of my way to | I would definitely go back | You will not be disappointed |
| The quality of the food | Go out of my way | At the end of the | Get what you pay for | Was out of this world |
| It was my first time | I won't be going back | On the other hand was | Never had a better meal | Just the right amount of |
| This place used to be | The best part of the | I am not a fan | The staff is friendly | Had in a long time |

Extracting some insight from the above table needs very precise attention. Some of the interesting details that may need further investigations can be:

- In 1 star reviews, the **negative direct recommendations** (Would never recommend) are very obvious. Also other **strong negative words** and expressions are very obvious too (Never go back, Never eat again) compared to 2 star rating.
- One interesting common feature in 3 star rating group is, having **at least the minimum quality** of food **but** there are some issues with the place or service (So it seems that the first priority is usually "*food quality*" which makes sense). In other words, the reviewers of this group address positive and negative points of restaurants at the same time. ("On the other hand" , "But" are seen relatively a lot)
- The 4 star reviews are the first time to see **positive direct recommendations** a lot (If you are looking for), it also shows a lot of **positive future behavior** (Will be back), there is a interesting difference between 4 star and 5 star ratings which **more excitement and passion** is seen in 5 star reviews. (Can't wait to go back ). Also **more positive encouragement and suggestions** ( You will not be disappointed) and **exaggerated and extreme** words are seen in 5 star (Best I've ever had/ Out of this world/ Had in a long time) .

## *II. N Grams (remove stop words)*

Here I used the smaller number as N (N=3 ), since I removed the stop words and that change makes common expressions smaller. The process was similar to the previous part, and the results that I got for the most common expressions in each rating group were very similar to the previous part.

***TF - IDF***

This method reduces the weights of more common words like (the, is, an etc.) which occurs in all document. This is called as **TF-IDF i.e Term Frequency times inverse document frequency.**

**A. Approach 1 (each word is analysed just in the documents are present - Excluding 0s TF-IDF)**

If the impact of words whenever they are present in each document is the subject of investigation, the 0s of each column (word) should be excluded. In this approach, there is a very small but important factor in analysis of this technique, especially when the number of document is relatively high that is getting ***high TF-IDF score for unique words (like name of people, restaurant, etc) Also the misspelled words usually gets high TF-IDF scores. (below figure)*** By having this point in mind, the output of this technique should be taken care before making any conclusion.

```
The index is: 19153  , the tf_idf mean score is: 0.9989891800035118 and the word is: limitword
The index is: 20294  , the tf_idf mean score is: 0.8657876347190105 and the word is: maurice
The index is: 3135  , the tf_idf mean score is: 0.844926741420074 and the word is: authentictasting
The index is: 37002  , the tf_idf mean score is: 0.8275019447166703 and the word is: бя
The index is: 17203  , the tf_idf mean score is: 0.7926016486957184 and the word is: interiorsservicefood
The index is: 9083  , the tf_idf mean score is: 0.7888118435599033 and the word is: dakota
The index is: 36110  , the tf_idf mean score is: 0.7801111036923897 and the word is: whoop
The index is: 24233  , the tf_idf mean score is: 0.7777072738679452 and the word is: pedro
The index is: 3807  , the tf_idf mean score is: 0.7756987556836809 and the word is: bbqjust
The index is: 28361  , the tf_idf mean score is: 0.7672117162085086 and the word is: salsitas
The index is: 36641  , the tf_idf mean score is: 0.7410315910527158 and the word is: yayas
```

One other interesting point is some words that are written with excitement and have same letters together have high tf_idf average scores in the whole dataset. Words such as "wooooow " , "yasyas" , "aaaaaamazing" , and etc are amongst these words.

If every review is analysed individually, we can get the most influential words in each document. Since the dataset is relatively big, I just brought some of them here:

```
Top words in document 280
        Word: smile, TF-IDF: 0.32263
        Word: atmosphere, TF-IDF: 0.17527
        Word: tasty, TF-IDF: 0.17477
Top words in document 281
        Word: bingo, TF-IDF: 0.10334
        Word: play, TF-IDF: 0.06188
        Word: ratchet, TF-IDF: 0.05348
Top words in document 282
        Word: seriously, TF-IDF: 0.19182
        Word: pico, TF-IDF: 0.18721
        Word: carnitas, TF-IDF: 0.18325
Top words in document 283
        Word: glendale, TF-IDF: 0.13151
        Word: vitoshad, TF-IDF: 0.08026
        Word: pizzasthe, TF-IDF: 0.0766
Top words in document 284
        Word: iced, TF-IDF: 0.10974
        Word: stpatricks, TF-IDF: 0.09898
        Word: sighting, TF-IDF: 0.09898
Top words in document 285
        Word: montadito, TF-IDF: 0.14369
        Word: chicharr├│n, TF-IDF: 0.13715
        Word: mofongo, TF-IDF: 0.13251
Top words in document 286
        Word: decide, TF-IDF: 0.19678
        Word: difficult, TF-IDF: 0.19061
        Word: prepared, TF-IDF: 0.15114
```

In general, unique names and then some extreme positive or negative words (amazing, awful, etc) have the high TF-IDF scores in each review.

**B. Approach 2 (words are analysed based on the whole document - Including 0s TF-IDF)**

In this approach, the whole rows (whether the specific word is present in each document or not) are under investigation and the "Average TF-IDF score" is calculated on the whole dataset (including 0s).

```
The index is: 32992  , the tf_idf mean score is: 0.12470247704911087 and the word is: the
The index is: 2322  , the tf_idf mean score is: 0.08672071554212421 and the word is: and
The index is: 35651  , the tf_idf mean score is: 0.06422218090118853 and the word is: was
The index is: 33423  , the tf_idf mean score is: 0.05834787342320686 and the word is: to
The index is: 17365  , the tf_idf mean score is: 0.04604597333908818 and the word is: is
The index is: 22756  , the tf_idf mean score is: 0.043702902356594533 and the word is: of
The index is: 17401  , the tf_idf mean score is: 0.04315999291816829 and the word is: it
The index is: 13270  , the tf_idf mean score is: 0.039186341695144254 and the word is: for
The index is: 35761  , the tf_idf mean score is: 0.037834196360432136 and the word is: we
The index is: 16763  , the tf_idf mean score is: 0.03537075871103319 and the word is: in
The index is: 13163  , the tf_idf mean score is: 0.03331553325133301 and the word is: food
The index is: 33129  , the tf_idf mean score is: 0.03139284794347962 and the word is: this
The index is: 5588  , the tf_idf mean score is: 0.030396385558006524 and the word is: but
The index is: 21772  , the tf_idf mean score is: 0.030320069802536535 and the word is: my
The index is: 36292  , the tf_idf mean score is: 0.03017469413238613 and the word is: with
The index is: 14470  , the tf_idf mean score is: 0.02910971648984088 and the word is: good
The index is: 36780  , the tf_idf mean score is: 0.028295269763590414 and the word is: you
The index is: 32966  , the tf_idf mean score is: 0.027828797174542633 and the word is: that
The index is: 14754  , the tf_idf mean score is: 0.027550479407554056 and the word is: great
The index is: 33059  , the tf_idf mean score is: 0.027451047847110093 and the word is: they
The index is: 15127  , the tf_idf mean score is: 0.0269325789869546 and the word is: had
The index is: 24882  , the tf_idf mean score is: 0.026390709685571775 and the word is: place
The index is: 35941  , the tf_idf mean score is: 0.025873341522061548 and the word is: were
The index is: 22981  , the tf_idf mean score is: 0.025779343852543955 and the word is: on
The index is: 22465  , the tf_idf mean score is: 0.0256307587153162 and the word is: not
The index is: 2695  , the tf_idf mean score is: 0.02394986468701806 and the word is: are
The index is: 15464  , the tf_idf mean score is: 0.023933974324525393 and the word is: have
The index is: 29237  , the tf_idf mean score is: 0.022683473452009 65 and the word is: service
```

In this approach, on top of the list (highest TF-IDF score) are mostly common words (the, and, was, etc.) and then some adjectives (great, good, amazing, etc) and then some food names (chicken, pizza, etc.).

### C . Approach 3 (TF - IDF on modified text (after stop words removal) )

Finally, this approach can diminish the impact of stop words, and here is the result:

```
The index is: 11498  , the tf_idf mean score is: 0.038565460113140 54 and the word is: food
The index is: 12660  , the tf_idf mean score is: 0.033575532446043645 and the word is: good
The index is: 21996  , the tf_idf mean score is: 0.032752556769533224 and the word is: place
The index is: 12917  , the tf_idf mean score is: 0.03142011323068666 and the word is: great
The index is: 12400  , the tf_idf mean score is: 0.02662651555885503 and the word is: get
The index is: 12586  , the tf_idf mean score is: 0.026594884561317735 and the word is: go
The index is: 25632  , the tf_idf mean score is: 0.026257795534241624 and the word is: service
The index is: 20522  , the tf_idf mean score is: 0.025262039323639494 and the word is: order
The index is: 6834  , the tf_idf mean score is: 0.023422424573149685 and the word is: come
The index is: 29032  , the tf_idf mean score is: 0.022637699255719304 and the word is: time
The index is: 16793  , the tf_idf mean score is: 0.020868541563612293 and the word is: like
The index is: 3120  , the tf_idf mean score is: 0.018762153443669896 and the word is: back
The index is: 29653  , the tf_idf mean score is: 0.01847357964431506 and the word is: try
The index is: 17145  , the tf_idf mean score is: 0.018226723768376182 and the word is: love
The index is: 23531  , the tf_idf mean score is: 0.017639831753698786 and the word is: really
The index is: 20338  , the tf_idf mean score is: 0.017403477293770547 and the word is: one
The index is: 17467  , the tf_idf mean score is: 0.016825611942174863 and the word is: make
The index is: 31833  , the tf_idf mean score is: 0.016290269827321933 and the word is: would
The index is: 6124  , the tf_idf mean score is: 0.016210407766239884 and the word is: chicken
The index is: 9699  , the tf_idf mean score is: 0.016026384733398978 and the word is: eat
The index is: 3819  , the tf_idf mean score is: 0.01597781522282413 and the word is: best
The index is: 8419  , the tf_idf mean score is: 0.015343906177937008 and the word is: delicious
The index is: 22725  , the tf_idf mean score is: 0.015327694366840656 and the word is: price
The index is: 21954  , the tf_idf mean score is: 0.015239935526626145 and the word is: pizza
The index is: 24070  , the tf_idf mean score is: 0.015235743852000513 and the word is: restaurant
The index is: 19530  , the tf_idf mean score is: 0.014520740848286157 and the word is: nice
The index is: 30979  , the tf_idf mean score is: 0.014504462752334987 and the word is: wait
The index is: 2054  , the tf_idf mean score is: 0.014238109636776249 and the word is: always
The index is: 2027  , the tf_idf mean score is: 0.014149142025256937 and the word is: also
The index is: 30360  , the tf_idf mean score is: 0.01404969993768703 and the word is: us
The index is: 11994  , the tf_idf mean score is: 0.014022613068799681 and the word is: fry
The index is: 28231  , the tf_idf mean score is: 0.013755574340146254 and the word is: take
The index is: 11890  , the tf_idf mean score is: 0.013362276788455 26 and the word is: friendly
The index is: 27243  , the tf_idf mean score is: 0.013302745374052821 and the word is: staff
The index is: 2078  , the tf_idf mean score is: 0.013248511289463777 and the word is: amaze
```

The above table tells us the important words in the whole review dataset. Same analysis can be done on the separate group of ratings in order to recognize the differences in different rating groups.

```
------ The words with highest TF-IDF scores in 1 star group -------
The index is: 3405  , the tf_idf mean score is: 0.041854608341795135 and the word is: food
The index is: 5788  , the tf_idf mean score is: 0.03887033478151762 and the word is: order
The index is: 3637  , the tf_idf mean score is: 0.034325159203172625 and the word is: get
The index is: 3678  , the tf_idf mean score is: 0.033724593866598415 and the word is: go
The index is: 6191  , the tf_idf mean score is: 0.03059672978683831 and the word is: place
The index is: 1976  , the tf_idf mean score is: 0.0293429328508784 and the word is: come
The index is: 7268  , the tf_idf mean score is: 0.028765711202300235 and the word is: service
The index is: 8269  , the tf_idf mean score is: 0.026756117660533194 and the word is: time
The index is: 8665  , the tf_idf mean score is: 0.02450464556984513 and the word is: us
The index is: 7126  , the tf_idf mean score is: 0.0233933645039022 and the word is: say
The index is: 4827  , the tf_idf mean score is: 0.023220392555815452 and the word is: like
The index is: 1005  , the tf_idf mean score is: 0.02281029700156039 and the word is: back
The index is: 8042  , the tf_idf mean score is: 0.022724834833080288 and the word is: take
The index is: 8839  , the tf_idf mean score is: 0.022662547433539862 and the word is: wait
The index is: 5750  , the tf_idf mean score is: 0.022191009410048028 and the word is: one
The index is: 5551  , the tf_idf mean score is: 0.02184937083960551 and the word is: never
The index is: 877  , the tf_idf mean score is: 0.021739686233436892 and the word is: ask
The index is: 3050  , the tf_idf mean score is: 0.021271840722107506 and the word is: even
The index is: 9090  , the tf_idf mean score is: 0.020924370490551507 and the word is: would
The index is: 3700  , the tf_idf mean score is: 0.020178016685414493 and the word is: good
```

```
------ The words with highest TF-IDF scores in 2 star group -------
The index is: 3394  , the tf_idf mean score is: 0.041631589349644 95 and the word is: food
The index is: 5849  , the tf_idf mean score is: 0.033556548754031384 and the word is: order
The index is: 3710  , the tf_idf mean score is: 0.03244108705169513 and the word is: good
The index is: 6282  , the tf_idf mean score is: 0.03097473488459214 and the word is: place
The index is: 3638  , the tf_idf mean score is: 0.030614388480889248 and the word is: get
The index is: 3691  , the tf_idf mean score is: 0.029991428275045504 and the word is: go
The index is: 7334  , the tf_idf mean score is: 0.02736778900173042 and the word is: service
The index is: 4841  , the tf_idf mean score is: 0.02667646404107184 and the word is: like
The index is: 8354  , the tf_idf mean score is: 0.02599812439262411 and the word is: time
The index is: 1935  , the tf_idf mean score is: 0.0258312078436424 and the word is: come
The index is: 6743  , the tf_idf mean score is: 0.02017176960006998 and the word is: really
The index is: 869  , the tf_idf mean score is: 0.02004495068581224 and the word is: back
The index is: 5795  , the tf_idf mean score is: 0.01982978375767665 and the word is: one
The index is: 9194  , the tf_idf mean score is: 0.019383310605123945 and the word is: would
The index is: 1719  , the tf_idf mean score is: 0.019213793734960558 and the word is: chicken
The index is: 8781  , the tf_idf mean score is: 0.019034099446287724 and the word is: us
The index is: 7197  , the tf_idf mean score is: 0.01890254472623203 and the word is: say
The index is: 8943  , the tf_idf mean score is: 0.018844752298337555 and the word is: wait
The index is: 8526  , the tf_idf mean score is: 0.01818743136436674 and the word is: try
The index is: 8172  , the tf_idf mean score is: 0.018108856903981525 and the word is: taste
```

```
------ The words with highest TF-IDF scores in 3 star group -------
The index is: 4314  , the tf_idf mean score is: 0.04074412459663194 and the word is: food
The index is: 4716  , the tf_idf mean score is: 0.039860122215878965 and the word is: good
The index is: 8088  , the tf_idf mean score is: 0.03184141960272601 and the word is: place
The index is: 4625  , the tf_idf mean score is: 0.029209428074165374 and the word is: get
The index is: 7554  , the tf_idf mean score is: 0.028614399279113793 and the word is: order
The index is: 4693  , the tf_idf mean score is: 0.02661670883940662 and the word is: go
The index is: 6190  , the tf_idf mean score is: 0.026155312781065393 and the word is: like
The index is: 9454  , the tf_idf mean score is: 0.025867692182684303 and the word is: service
The index is: 2465  , the tf_idf mean score is: 0.02411654907137067 and the word is: come
The index is: 10830  , the tf_idf mean score is: 0.022983111040025362 and the word is: time
The index is: 4816  , the tf_idf mean score is: 0.0221632769373433 and the word is: great
The index is: 11861  , the tf_idf mean score is: 0.02115735173312847 and the word is: would
The index is: 8675  , the tf_idf mean score is: 0.020870490304784554 and the word is: really
The index is: 8343  , the tf_idf mean score is: 0.018615912412276833 and the word is: price
The index is: 11053  , the tf_idf mean score is: 0.018390715924718665 and the word is: try
The index is: 8334  , the tf_idf mean score is: 0.018159046127848886 and the word is: pretty
The index is: 7498  , the tf_idf mean score is: 0.018025523686217924 and the word is: one
The index is: 2201  , the tf_idf mean score is: 0.017514671018122645 and the word is: chicken
The index is: 6445  , the tf_idf mean score is: 0.017169476731370904 and the word is: make
The index is: 4471  , the tf_idf mean score is: 0.01681662683043951 and the word is: fry
```

Data Science Capstone Project
Springboard Bootcamp, San Francisco
Reza Taeb
Spring 2018
Title : “Improving Restaurant Reputation Using Yelp User Reviews”
Part 2: Storytelling

```
------ The words with highest TF-IDF scores in 4 star group -------
The index is: 6382  , the tf_idf mean score is: 0.04163136173883538 and the word is: good
The index is: 5811  , the tf_idf mean score is: 0.036848799516895646 and the word is: food
The index is: 6523  , the tf_idf mean score is: 0.034251472500542834 and the word is: great
The index is: 10948  , the tf_idf mean score is: 0.03323662006155973 and the word is: place
The index is: 6243  , the tf_idf mean score is: 0.02741528684824884 and the word is: get
The index is: 6345  , the tf_idf mean score is: 0.02583011917383694 and the word is: go
The index is: 12739  , the tf_idf mean score is: 0.02527558677124518 and the word is: service
The index is: 10207  , the tf_idf mean score is: 0.023343616733586356 and the word is: order
The index is: 3403  , the tf_idf mean score is: 0.023151961468794596 and the word is: come
The index is: 14506  , the tf_idf mean score is: 0.022503083807181766 and the word is: time
The index is: 8418  , the tf_idf mean score is: 0.021572093557275705 and the word is: like
The index is: 11732  , the tf_idf mean score is: 0.020069763236839983 and the word is: really
The index is: 14826  , the tf_idf mean score is: 0.019808994021065625 and the word is: try
The index is: 9728  , the tf_idf mean score is: 0.019330452720525852 and the word is: nice
The index is: 8594  , the tf_idf mean score is: 0.019132635704502944 and the word is: love
The index is: 3042  , the tf_idf mean score is: 0.01777053257857148 and the word is: chicken
The index is: 1527  , the tf_idf mean score is: 0.017539076130726722 and the word is: back
The index is: 10119  , the tf_idf mean score is: 0.017480775291666542 and the word is: one
The index is: 11329  , the tf_idf mean score is: 0.01729746554988776 and the word is: price
The index is: 15873  , the tf_idf mean score is: 0.016813147090390782 and the word is: would


------ The words with highest TF-IDF scores in 5 star group -------
The index is: 5775  , the tf_idf mean score is: 0.04104586070793748 and the word is: food
The index is: 6482  , the tf_idf mean score is: 0.040721225848869516 and the word is: great
The index is: 10984  , the tf_idf mean score is: 0.03685056880587934 and the word is: place
The index is: 6357  , the tf_idf mean score is: 0.030509974982444395 and the word is: good
The index is: 12774  , the tf_idf mean score is: 0.028967830732228413 and the word is: service
The index is: 6318  , the tf_idf mean score is: 0.026692235941517307 and the word is: go
The index is: 8555  , the tf_idf mean score is: 0.025357750442189228 and the word is: love
The index is: 6225  , the tf_idf mean score is: 0.02400885239036771 and the word is: get
The index is: 1813  , the tf_idf mean score is: 0.023580433661669065 and the word is: best
The index is: 3399  , the tf_idf mean score is: 0.023264761755542807 and the word is: come
The index is: 14470  , the tf_idf mean score is: 0.022717656316880153 and the word is: time
The index is: 4231  , the tf_idf mean score is: 0.021715983667886 and the word is: delicious
The index is: 914  , the tf_idf mean score is: 0.021190845975573897 and the word is: amaze
The index is: 1451  , the tf_idf mean score is: 0.02073149636478149 and the word is: back
The index is: 10224  , the tf_idf mean score is: 0.020431673990848883 and the word is: order
The index is: 14804  , the tf_idf mean score is: 0.020217635553904573 and the word is: try
The index is: 901  , the tf_idf mean score is: 0.019379459180167573 and the word is: always
The index is: 8724  , the tf_idf mean score is: 0.018614337976849065 and the word is: make
The index is: 8386  , the tf_idf mean score is: 0.017707684911565772 and the word is: like
The index is: 5973  , the tf_idf mean score is: 0.017675304434091918 and the word is: friendly
```

By looking at the above tables, we can observe that although some words (food, place, get, order, service) are seen in all rating groups, but some words are just in one group or their place is going up or down in each rating group. The results are very similar to the ones that we got from previous parts. For example, the very positive words such as “delicious” , and “friendly” are seen just in 5 star group and “wait” is seen in group 1 and 2 star.