IST2334: Web and Network Analytics Assignment

Group Members:

|  |  |
| --- | --- |
| Name | Student ID |
| Wong Yuen-Yi | 17112723 |
| Wong Wei Chean | 18000133 |

## Introduction

In this assignment, we have chosen an interesting data set of user interactions and uploads to Food.com, and subsequently performed three distinctive areas of analysis on the data. For this project, our aim is to:

* analyse the opinions of the users to determine their feelings about the recipes (eg. were they satisfied, or did they have any negative experiences?) through sentiment analysis,
* find out how certain characteristics about a recipe influences its rating, and
* perform market basket analysis on the dataset to study ingredient patterns across different recipes.

Through performing these analysis, our hope is to be able to extract meaningful information from the large set of data.

## Libraries and The Data Set

We will use several libraries to conduct our analysis for our project, namely:

* TidyVerse, which in itself includes useful packages for data analyses that we will be using like ggplot2, dplyr, tibble, and stringr,
* TidyText, for text mining purposes in the sentiment analysis section,
* ARules, for mining algorithms like APRIORI that are used in market basket analysis, and
* ARulesViz, for visualisation of association rules.

The data set used in this assignment is by Shuyang Li and Bodhisattwa Prasad Majumder, taken from [Kaggle](https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions?select=PP_recipes.csv). This data set contains over 180,000 recipes and 700,000 recipe review uploads to Food.com over the past 18 years. Here, we have multiple CSV files that contain data about the recipes on the site, user information, and user interactions.

We will load the relevant dataset files to conduct analysis, which are RAW\_recipes.csv and RAW\_interactions.csv. These are the files with information that we are mainly concerned with in our analysis.

In the recipes file, there are 231,637 recipe entries, each with 12 columns to describe them, such as recipe name, expected duration, date of submission, contributor, number of steps, and ingredients. The interactions file contains information about user interactions on the site, which consists of 1,132,367 entries and 5 columns, which describe the interactions by the user ID, recipe ID, date published, rating and review comments.

For our analysis, we are also able to join or merge these files to give us more context to produce purposeful results.

These files in the dataset contain large amounts of data; however, in this form, they do not provide any meaningful insights about the success of the website or recipes published on the website. As such, we have taken the opportunity to analyse this set of data to obtain indicative insights that will help us to understand better whether or not the website or certain recipes are successfully fulfilling the needs and wants of the users.

## The Analysis

In this section, we will explore three different types of analysis conducted on the Food.com dataset from Kaggle.

### Sentiment Analysis

As stated in the aim of our project, we intend to use the dataset to discover the general opinions of the users to determine their feelings and opinions about the recipes on Food.com. This can be done through sentiment analysis, or opinion mining, which is a technique used to extract, quantify or identify subjective information. We can apply sentiment analysis on the reviews of the recipes published on Food.com to effectively analyse the position or the thoughts of the users about the recipes on the website.

For this assignment, we will be performing sentiment analysis on 20 top-rated recipes and on 20 recipes with the the lowest rating. The analysis in this section is done with references to [this tutorial from datacamp.com](https://www.datacamp.com/community/tutorials/sentiment-analysis-R).

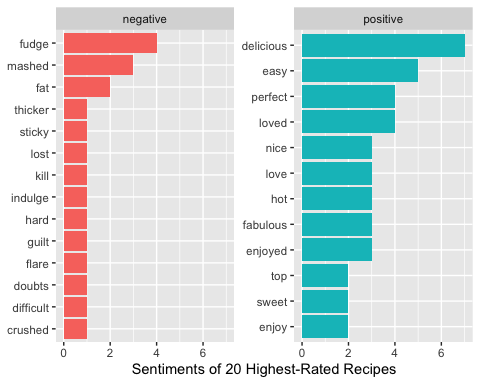
Before the lexicon is applied to the dataset, first we must perform some pre-processing works, such as removing stop words and special characters from the review texts. Stop words are common words that appear in natural languages which do not provide meaning to our data and what we woud like to find out from it. It is important that these are removed from our dataset because stop words and special characters may interfere with the performance of our code and integrity of the results of the sentiment analysis, as removing stop words can most likely significantly reduce the number of postings that the system must store and process. This will help to increase the efficiency of the analysis process.

We will be using a list of stop words from the Tidytext package. These stop words in the list will be removed from our top\_recipes and worst\_recipes data by using the anti\_join function after tokenizing the review texts by individual words.

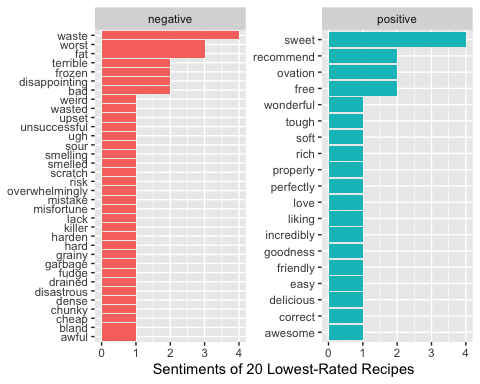
Next, we will use the [Bing lexicon](https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html) to get sentiments from top\_recipes and worst\_recipes. The Bing lexicon helps categorise the text by classifying words and their associations with two sentiments (negative and positive).

To visualise the data, we can plot the sentiment words against their frequency for top\_sentiments and worst\_sentiments using the ggplot2 library.

## Selecting by n



## Selecting by n



From the above, we can see that the number of positive sentiments for the 20 highest-rated recipes is much larger than the number of positive sentiments for the 20 lowest-rated recipes. In the same manner, the number of negative sentiments for the 20 lowest-rated recipes is larger than the number of negative sentiments for the 20 highest-rated recipes.

From the above, we can see that the number of positive sentiments for the 20 highest-rated recipes is much larger than the number of positive sentiments for the 20 lowest-rated recipes, where although there are more varying positive sentiment words used in the 1-star recipes, the frequencies of the positive sentiment words used in the 5-star recipe reviews are higher. For example, the word “delicious” was used a total of 7 times in the top 20 reviews, while it was only used once in the reviews of the 1-star recipes.

We can also note that the number of negative sentiments from the lowest-rated recipes is much larger than that from the highest-rated recipes. This is expected, as typically, users tend to give low ratings to recipes when their sentiment is generally negative such as bad experiences with the method, or the dish did not turn out as they expected. For example, there are 34 different negative sentiment words that appear in the 1-star recipe reviews, while only 14 different negative sentiment words appear in the 5-star recipe reviews.

From this portion, we validated that for recipes which received 5-star ratings, the number of positive sentiment words that appear in the reviews section is much greater than the number of positive sentiment words that appear in the reviews of 1-star recipes. In the same manner, the number of negative sentiment words that appear in the reviews of 5-star recipes is very much less than the number of negative sentiment words that appear in the reviews of 1-star recipes. This has proved that the rating feature on the website is an effective way to rank the recipe based on how well-received it is by the users. As such, we can conclude that the rating scale is able to provide rather accurate approximation of the users’ general opinion about a given recipe. This information is useful for the Food.com management to gauge the user experience of a particular recipe, and devise appropriate marketing strategies that can help to promote recipes that are popular. By doing so, they will be able to bring in more new users to the site and encourage existing users to return to their site to try more recipes.

### Correlation

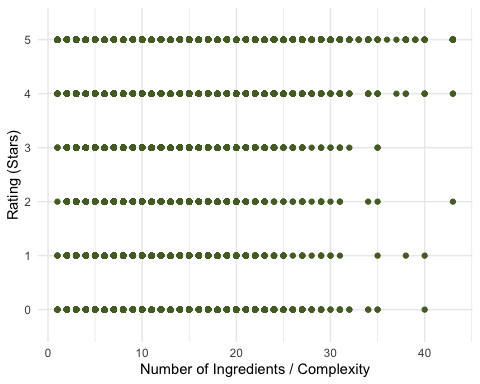
We will use this section to answer an interesting question about the data set:

Does rating of recipe and complexity of recipe have any correlation?

The analysis in this section is inspired by one of the notebooks found in the Food.com Recipe and Interactions Kaggle site. The [notebook](https://www.kaggle.com/etsc9287/food-com-eda-and-text-analysis) calculates the correlation between length of review and rating of recipe. In our case, we are looking for the correlation between rating and complexity of the recipes, but we do adopt similar techniques to obtain results.

For this, we will merge recipes and interactions to get the ratings associated with each recipe. To measure the complexity of the recipes, we will use the n\_ingredients, which represents the number of ingredients involved in the recipe. We have chosen to use this variable as the measurement for complexity because typically, the higher the number of ingredients, the more complicated the method of food preparation is.

We will group the data by their recipe ID, and find the correlation coefficient between number of steps and rating of recipe by using the cor() function.



## [1] -0.003767314

From the scatter plot, we can observe that the number of ingredients for recipes ranging from 0- to 5-stars generally fall between 1 to 40 ingredients. There does not seem to be a significant trend in number of ingredients across the recipes of different ratings. However, it is found that for 3-star recipes, the average number of ingredients is the lowest among the other ratings while the average number of ingredients for 5-star recipes is the highest.

The correlation coefficient calculated between the number of ingredients, which represents complexity, and the rating of the recipes is about -0.004. This negative value shows that the trend of the variables move in opposite directions; in other words, the number of ingredients decreases as the rating of the recipes increase. Yet, it must be noted that the value of the coefficient is very small. This tells us that the two variables have a very weak linear relationship. Thus, it can be said that there is almost no meaningful relationship between the complexity of the recipe and the rating.

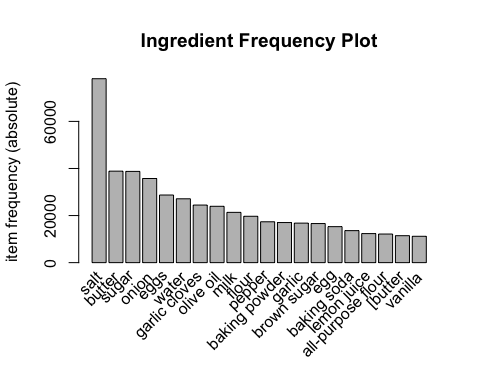
Calculating the correlation between complexity and ratings of the recipes showed that there is close to no correlation between the two variables, despite its small, almost insignificant negative value. This means that recipes uploaded on Food.com can have high ratings regardless of the complexity of the recipe, which we have measured by considering the number of ingredients called for by the recipe. This may be because Food.com attracts users of all cooking skill levels; from amateur home-cooks who want to master basic techniques to more advanced chefs looking to expand their repertoire. Hence, even the simplest or most tedious of recipes may be well-received by users. This insight can tell us that since their target audience is wide and from different skill levels, Food.com must ensure that recipes and other website content must be regularly updated for every skill-level group in order to effectively increase the web traffic.

### Market Basket Analysis

Market basket analysis is typically used to predict future purchase decisions of customers in a business setting. However, we will be applying the market basket analysis concept on our dataset to see ingredient patterns across the recipes uploaded on Food.com. For this analysis, we have referred to [this tutorial](https://www.datacamp.com/community/tutorials/market-basket-analysis-r#code), particularly the section on implementing market basket analysis and association rule mining in R.

To begin this process, we select the relevant columns of data from the tibble of recipes. Since we are only interested with the ingredients of the recipes, we select the ingredients column and use the na.omit() function to remove any null values from the set.

Following that, we perform some pre-processing of the selected data, that is to write the recipe ingredients to a CSV file and convert it to a transaction object. This allows us to easily count the number of individual ingredients as each of the items are separated by a comma. We are then able to visualise the data through a frequency plot of the ingredients. Below, a frequency bar chart of the 20 most common ingredients that appear in all the recipes on Food.com is plotted using the itemFrequencyPlot() function from the arules library.

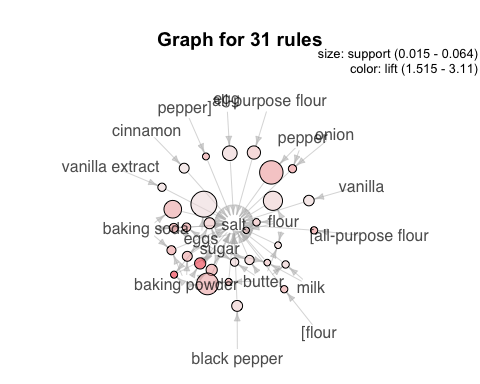


From the Ingredient Frequency Plot, we can see the top 20 common ingredients in the recipes from Food.com. From this we can note the ingredients that are most important to have at home, or “pantry staples” for cooking any sort of recipe.

Next, we move on to generate some association rules from the data using APRIORI algorithm. This is easily done in R by using the arules library. Through trial and error, we have found that the support and confidence thresholds to obtain around 30 rules is 0.015 and 0.5 respectively. Using these values, the apriori method and the plot() function, we can generate the rules and plot a graph of the rules.

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.5 0.1 1 none FALSE TRUE 5 0.015 1  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 3474   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[261274 item(s), 231638 transaction(s)] done [1.21s].  
## sorting and recoding items ... [90 item(s)] done [0.05s].  
## creating transaction tree ... done [0.17s].  
## checking subsets of size 1 2 3 4 done [0.04s].  
## writing ... [31 rule(s)] done [0.00s].  
## creating S4 object ... done [0.14s].

## lhs rhs support confidence coverage   
## [1] {onion,pepper} => {salt} 0.01956501 0.7921692 0.02469802  
## [2] {[all-purpose flour} => {salt} 0.01710427 0.7854877 0.02177536  
## [3] {baking soda,eggs} => {salt} 0.02163721 0.7739345 0.02795742  
## [4] {pepper} => {salt} 0.05768915 0.7699798 0.07492294  
## [5] {baking soda,sugar} => {salt} 0.02037231 0.7673171 0.02655005  
## [6] {baking powder,baking soda} => {salt} 0.02162426 0.7654340 0.02825098  
## [7] {pepper]} => {salt} 0.01626244 0.7605492 0.02138250  
## [8] {baking powder,sugar} => {salt} 0.02712422 0.7551683 0.03591811  
## [9] {baking powder,butter} => {salt} 0.01641354 0.7533188 0.02178831  
## [10] {baking powder,eggs} => {salt} 0.02380007 0.7405965 0.03213635  
## [11] {baking soda} => {salt} 0.04328737 0.7375506 0.05869072  
## [12] {[flour} => {salt} 0.01710859 0.7351141 0.02327338  
## [13] {baking powder} => {salt} 0.05355339 0.7282494 0.07353716  
## [14] {flour,sugar} => {salt} 0.01754462 0.6275479 0.02795742  
## [15] {milk,sugar} => {salt} 0.01546378 0.6241505 0.02477573  
## [16] {eggs,flour} => {salt} 0.01522634 0.6173639 0.02466348  
## [17] {all-purpose flour} => {salt} 0.03221406 0.6141564 0.05245253  
## [18] {eggs,sugar} => {salt} 0.02660185 0.6113095 0.04351618  
## [19] {butter,sugar} => {salt} 0.02225024 0.5786460 0.03845224  
## [20] {black pepper} => {salt} 0.02643349 0.5709091 0.04630069  
## [21] {flour} => {salt} 0.04708640 0.5537954 0.08502491  
## [22] {butter,milk} => {salt} 0.01738489 0.5533874 0.03141540  
## [23] {butter,flour} => {salt} 0.01584369 0.5512994 0.02873881  
## [24] {egg} => {salt} 0.03582745 0.5428441 0.06599953  
## [25] {butter,eggs} => {salt} 0.01982835 0.5427795 0.03653114  
## [26] {vanilla} => {salt} 0.02549236 0.5275147 0.04832540  
## [27] {vanilla extract} => {salt} 0.01970747 0.5228496 0.03769243  
## [28] {baking powder,eggs} => {sugar} 0.01672437 0.5204191 0.03213635  
## [29] {eggs} => {salt} 0.06399209 0.5160134 0.12401247  
## [30] {cinnamon} => {salt} 0.02374826 0.5109604 0.04647769  
## [31] {baking powder,salt} => {sugar} 0.02712422 0.5064893 0.05355339  
## lift count  
## [1] 2.348695 4532  
## [2] 2.328885 3962  
## [3] 2.294631 5012  
## [4] 2.282906 13363  
## [5] 2.275011 4719  
## [6] 2.269428 5009  
## [7] 2.254945 3767  
## [8] 2.238991 6283  
## [9] 2.233508 3802  
## [10] 2.195787 5513  
## [11] 2.186757 10027  
## [12] 2.179533 3963  
## [13] 2.159180 12405  
## [14] 1.860611 4064  
## [15] 1.850538 3582  
## [16] 1.830416 3527  
## [17] 1.820906 7462  
## [18] 1.812466 6162  
## [19] 1.715622 5154  
## [20] 1.692683 6123  
## [21] 1.641943 10907  
## [22] 1.640733 4027  
## [23] 1.634542 3670  
## [24] 1.609473 8299  
## [25] 1.609282 4593  
## [26] 1.564023 5905  
## [27] 1.550192 4565  
## [28] 3.110456 3874  
## [29] 1.529923 14823  
## [30] 1.514942 5501  
## [31] 3.027200 6283



It is established that salt is the most commonly used ingredient in this data set. By reading the generated rules top\_rules, it is found that, in more than half of the recipes found on Food.com, salt is used together with onion and pepper, all-purpose flour, baking soda and eggs, and other combinations of ingredients. Similarly, sugar is also used with ingredient combinations such as baking powder and eggs, and baking powder and salt in about 50% of the recipes. From this, we can infer that at least about 50% of recipes are recipes for baked goods, due to the fact that baking soda and baking powder are used for baking.

From plotted graph, we can determine the more significant ingredients of the lot, or rather, ingredients that are more heavily called for in recipes, by the appearances of the nodes in the graph. The larger the circle, the higher the support and the darker the pink, the higher the lift. In this case, for example, the probability that eggs appear together with salt have high support but low lift, as the size of the circle is relatively large compared to other ingredients and the circle is quite light in colour. High support tells us that this combination of ingredients appear frequently, while low lift indicates that the association between the two ingredients is not very strong.

Through conducting market basket analysis on the dataset, we learned that, unsurprisingly, salt is the most common ingredient in food recipes. We were also able to generate 31 association rules that showed the likelihood of a particular ingredient being used if another ingredient appears in the recipe. For example, it was found that in about 77% of recipes on Food.com that called for pepper also called for salt. By knowing this information, we are able to predict the different combinations of ingredients that people are likely to purchase together at a supermarket or grocery shop. Supermarkets and grocery shops can greatly benefit from knowing this sort of information, as they will be able to boost the effectiveness of their marketing models, promotional efforts and even store layouts. In this, they can adjust their product arrangements and offer purchase-with-purchase deals that will greatly increase customer satisfaction, which in turn would generate more profits.

## Lessons Learned

After performing the analysis, we have learned several interesting things from the data set. Out of the large amount of raw data that we had about the recipes and interactions on Food.com, we were able to extract meaningful information which can help to provide useful insights about the nature of the website. These insights can not only be used by Food.com management, but other relevant companies or organisations to reason what type of marketing tactics or business decisions should be made to increase the success of their services or products. For example, supermarkets or grocers can leverage on the information derived from the market basket analysis portion to adjust their marketing strategies and provide more satisfactory offers to their customers, which will help them to generate more profits and return customers.

One of the most important things that we have learned from this assignment is the real-world application of data analytics. We were able to apply some prominent concepts that we have learned in theory during lectures, and strongly believe that being able to implement these concepts on a real set of data helped us to fully understand and appreciate them. We also learned that making mistakes is okay - in programming, we must learn from our mistakes, and we are heavily supported not only by immediate friends and our lecturers, but also by online resources. Through this assignment, we were able to learn to detect and solve errors in R through independent research by referring and learning from community forums such as DataCamp Community and Stack Overflow.

For future projects involving data analysis, we will work on improving the quality of the results by using more techniques to clean up the data before processing it. One example would be to check for and remove outliers in the dataset before conducting any sort of analysis on it. On the same note, we believe that this is one of a few aspects that we should improve in our project.

## Reflections

### Wong Yuen-Yi

Through this assignment, I have gained invaluable knowledge about data analytics. As a Computer Science student with not much prior knowledge in data analytics, I feel that I can understand and appreciate the practice much more than before. During the course of this subject and especially through this assignment, I have realised the potential and beauty of data analytics, which can help us make sense of large amounts of data, all of which is abundant in this day and age. I have also learned to code in R, which I believe is a valuable skill to have, as it is widely used in statistical computing, data mining and data analysis. All in all, these skills that I have gained will definitely help me to grow into a more well-rounded and better developer in the future.

One thing that I have learned through completing this assignment is that some concepts such as market basket analysis are multipurpose and can be applied to many different situations. For instance, we have applied market basket analysis to obtain information about the most common ingredients across the recipes found on Food.com and the likelihood of certain ingredients to appear in recipes when another ingredient is used. In the initial stages of the assignment, I was skeptical that this would produce relevant results or give meaningful insights; however, I was pleasantly surprised to be able to prove myself wrong.

All in all, these lessons that I have learned have essentially opened my eyes to the value of data and data analytics, while also sparking my interest in the subject. I have also learned to use GitHub for the purpose of collaborating on this assignment with my groupmate, which I think will be a useful skill for future projects or assignments and even possibly in the workforce. All things considered, despite running into many errors that required fixing, ultimately I feel that this project is very rewarding and I had a positive experience from completing this assignment.

### Wong Wei Chean

I believe this project has been a fruitful experience for several reasons, namely, I was able to learn how to implement some important theories that I have learned from the lectures in this subject. Applying these analysis on a real world dataset helped me to grow as a developer because I now know how to properly analyse data and produce useful observations that can be used to drive successes in businesses and especially websites.

Throughout the process of doing this assignment, I encountered several unexpected situations where we had to adapt. For example, my groupmate and I had to change our initial plan for one of our analysis, which was to compare the ratings of vegetarian recipes against those of non-vegetarian recipes. However, we were not able to perform that analysis but we quickly found a replacement, which is the correlation between ratings and complexity of the recipes. Besides that, there were many instances where I encountered errors in my code, but as I seeked help from my groupmate, classmates and online resources, I was able to learn from others and solve the errors relatively quickly.

## References

1. Shuyang Li, “Food.com Recipes and Interactions.” Kaggle, 2019, doi: 10.34740/KAGGLE/DSV/783630.
2. D. Liske, “Tidy Sentiment Analysis in R”, DataCamp Community, 2020. [Online]. Available: <https://www.datacamp.com/community/tutorials/sentiment-analysis-R>. [Accessed: 27- Nov- 2020].
3. Minqing Hu and Bing Liu, ‘’Mining and summarizing customer reviews.’’, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, Aug 22-25, 2004.
4. E. Schacht, “🍕 Food.com EDA and Text Analysis 🌯”, Kaggle.com, 2020. [Online]. Available: <https://www.kaggle.com/etsc9287/food-com-eda-and-text-analysis>. [Accessed: 27- Nov- 2020].
5. H. Jabeen, “Market Basket Analysis using R”, DataCamp Community, 2020. [Online]. Available: <https://www.datacamp.com/community/tutorials/market-basket-analysis-r#code>. [Accessed: 27- Nov- 2020].