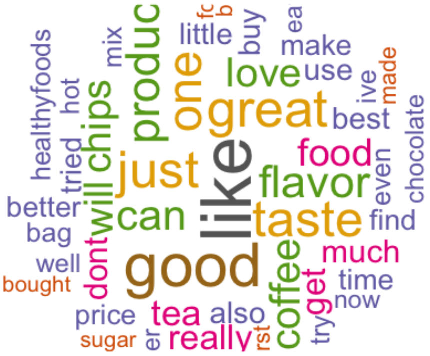
Group 8 : Assignment 4

Group Members : Yi-Hsuan Tseng, Chia-Ling Ni, Mihir Kungulwar, Tejaswini Edupuganti

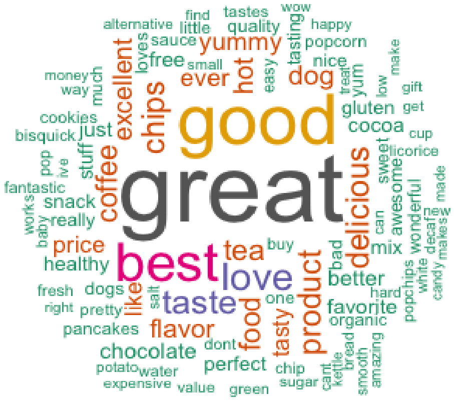
<table>
<tr><th>Questions</th><th>Team member names</th></tr>
<tr><td>What are some insights gained from exploratory data analysis? (Use visuals)</td><td>

- **Plot 1 : Word Cloud of reviews textual data**



- Widely used positive word : ‘Good’
- Most popular products in reviews : tea, chips, coffee, chocolate
- Other positive words : great, best, better, well

- **Plot 2 : Word Cloud of reviews title data**

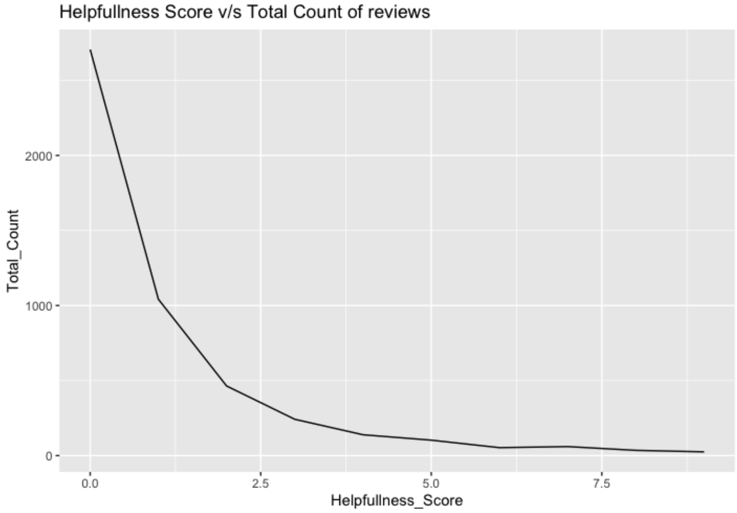


- Most popular words in review title : great, good, best
- Adjectives that describe the food : delicious, sweet, yummy, healthy, tasty, smooth
- Most popular food items in review titles : tea, cocoa, chips, coffee, popcorn, candy, pancakes, licorice, popchips

Insight from plot 1 and 2: From the most popular words in review texts and titles, we can know the overall customer feedback toward the store and food.

</td></tr>
</table>

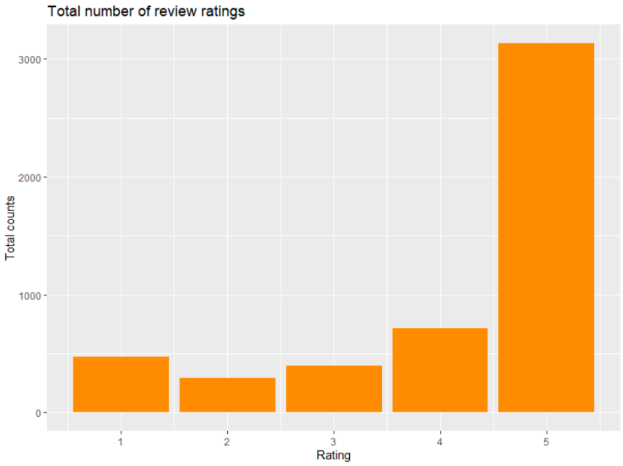
- **Plot 3: Total review count v/s helpfulness score**



- Most reviews have a helpfulness score between 0~2.
- Total of reviews having 0 helpfulness score is about 2500
- There are less reviews having higher helpfulness score

Insight from plot 3: Even though some reviews don't get any feedback from other customers, we can find out that there are still people actually reviewing those reviews and considering the reviews helpful when they're checking out the store or the restaruant.

- **Plot 4: Total number of review ratings**



- Majority or reviews have a rating of 5
- There are more reviews having a rating of 4 than 1,2 or 3
- More reviews have a rating of 1 than 2 or 3

Insight from plot 4: We can see that the majority of the customers leave good ratings toward the store, so most of the customers are satisfied with the food. However, for those below average-rating 3, rating of 1 is more than 2, which means there's some customers extremely unhappy and we need to further examine what or why customers are happy or unhappy.

Describe steps taken to prepare the data for sentiment analysis and reasoning for those steps.	**For topic modeling, dtm:** 1) Convert the required text column into a collection of text documents. This is our primary structure called Volatile Corpus. **Reasoning** : This makes managing the documents or the text easier. Corpus helps us with handling the delimiters, tokens, frequencies etc. 2) Strip off the numbers, punctuations and white spaces. **Reasoning** : These entities act as noise in the analysis which in turn hinder the results. Ex : The frequency of the white spaces might be the highest and can be counted as a word. 3) Converted every text to lowercase : **Reasoning** : For uniformity, helps match the words with the dictionary words better and capital letters take up more space thus slowing down the algorithm. 4) Strip off stop words. **Reasoning** : These are prepositions, conjunctions and articles used to form a sentence, but they do not add any value to the analysis of the final sentiment. So removing them makes the algorithm efficient and helps it converge faster. 5) Convert the Corpus into a Term Document Matrix **Reasoning** : Helps identify the relationship between terms and documents. It represents document vectors in matrix form in which the rows correspond to the terms in the document, columns correspond to the documents in the corpus and cells correspond to the frequency of the terms. 6) Sort the matrix in descending order to get the most frequently occurring word. **For sentiment analysis:** The data preparing steps are mostly the same as topic modeling and dtm. But we use tibble instead of character vectors, we need to use other methods to clean up and prepare the data. Converting every text to lowercase, stripping off the numbers, punctuations, urls and removing stop words and curse words are all needed for data preparation. The curse words are stripped off for sentiment analysis because we don’t want to give a sentiment analysis report to the manager and stakeholders that might have curse words.

<table>
<tr>
<td>Describe the difference and/or similarities observed using different lexicons in conducting sentiment analysis. Describe the overall sentiment of consumers toward the store.</td>
<td>

We choose **‘nrc’** and **‘afinn’** lexicon to do the sentiment analysis.

- **Differences and/or Similarities using these two lexicons:**

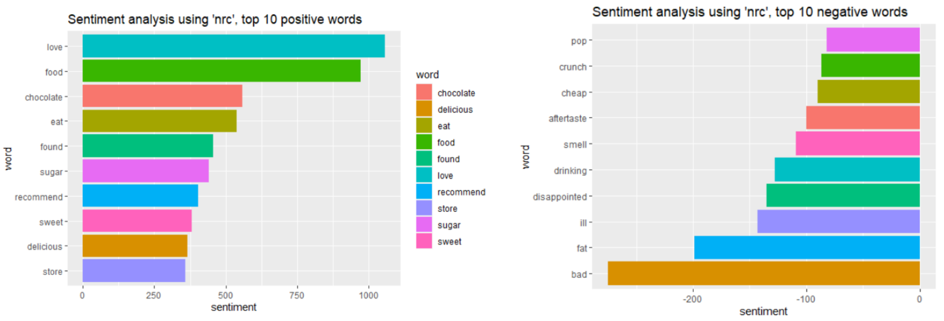
From the overall sentiment analysis score, we can notice that every lexicon has its own words, but there’re common words from both lexicons.

‘nrc’ lexicon assigns words with different sentiments, we need to first filter out the “positive” and “negative” sentiments from its own lexicon for further sentiment analysis. Then we could separate the positive and negative sentiments and get sentiment scores from corresponding words.

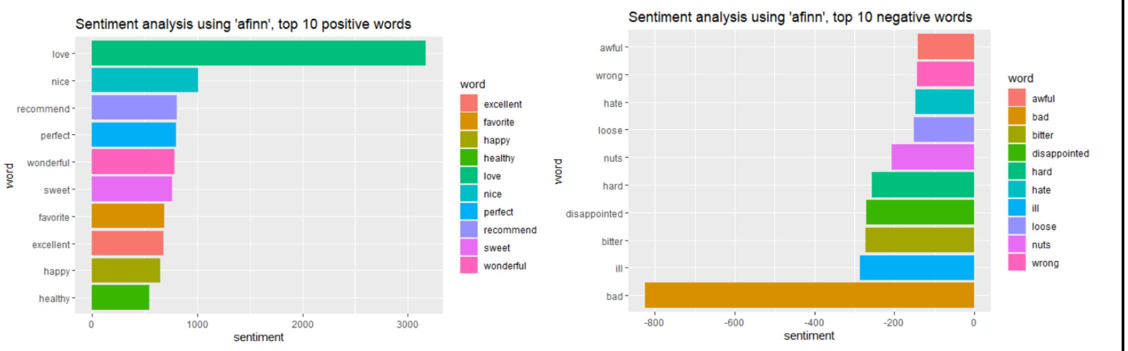
However, **‘afinn’ lexicon** already assigned each word with different weighted sentiment values. For example, if a word is positive, then the value will be a positive number, and if that word is strongly positive, the value is even higher than other positive words, vice versa for the negative words. Thus the result between the positive words and negative words may have higher differences with this lexicon.

We showed the top10 positive and top10 negative word plots from both lexicons as below. From the plots, we can see that ‘love’ is the most positive word and ‘bad’ is the most negative word from our reviews no matter in which lexicon. There’re also other common words from both lexicons such as ‘recommend’ and ‘sweet’ from top10 positive, ‘disappointed’ and ‘ill’ from top10 negative. However, we can also see that we will get more differences within ‘afinn’ lexicon sentiment analysis due to weighted values.

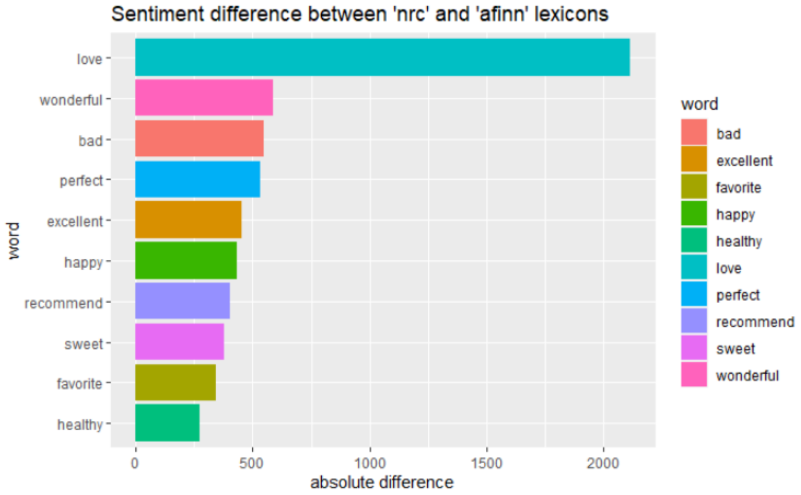
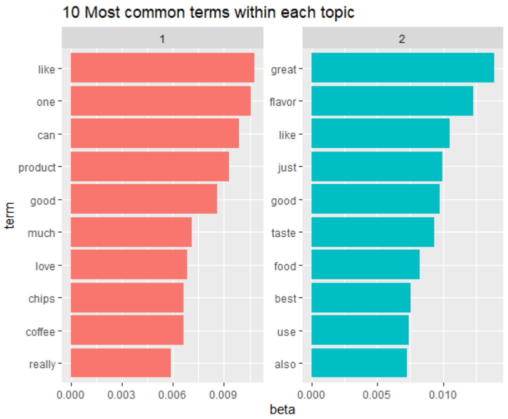
Plot 5-1: ‘nrc’ lexicon sentiment analysis, top10 positive and top10 negative



Plot 5-2: ‘afinn’ lexicon sentiment analysis, top10 positive and top10 negative

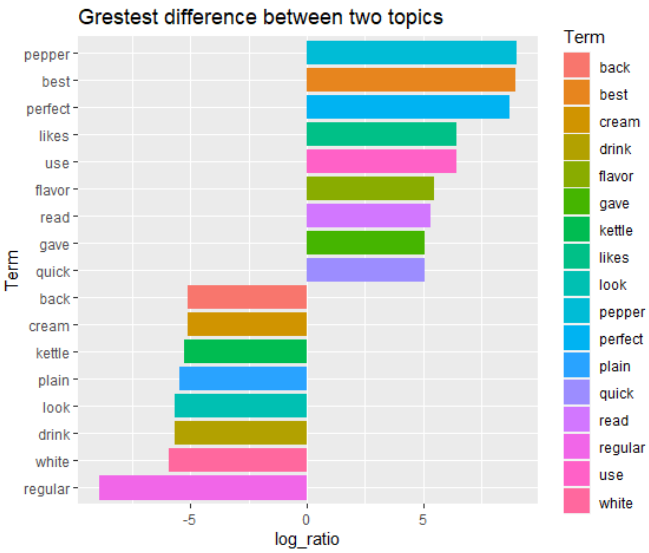


</td>
</tr>
</table>

	- **Plot 6: sentiment difference between 'nrc' and 'afinn' lexicons**  From the common words to each sentiment and discriminate the most between the sentiments. We can see that the majority of the common words are positive words, also the most positive word, 'love', and the most negative word, 'bad', still have relatively higher differences because 'afinn' lexicon weighted each word's value. - **Overall Sentiment: more positive than negative** With 'nrc' lexicon, the total sentiment score is 10949, and 'afinn' is 15113. Even though the total scores are different, we can definitely see that the overall consumer sentiment toward the store is more positive than negative.
Describe insights gained from your analysis using topic modeling with a discussion of what the results mean and how you would describe the two main topics observed in the reviews.	In the topic modeling part, we will separate the analysis into 2 parts. First part will focus on using both beta (per-topic-per-word probabilities) and gamma (per-document-per-topic probabilities) to identify two main topics that appear to be persistent in the reviews. The second part will implement four scoring algorithms to find the perfect number of topics for the LDA model. **<u>Part I</u>** - **Plot 7: Top 10 most common terms of each topic(beta)** 

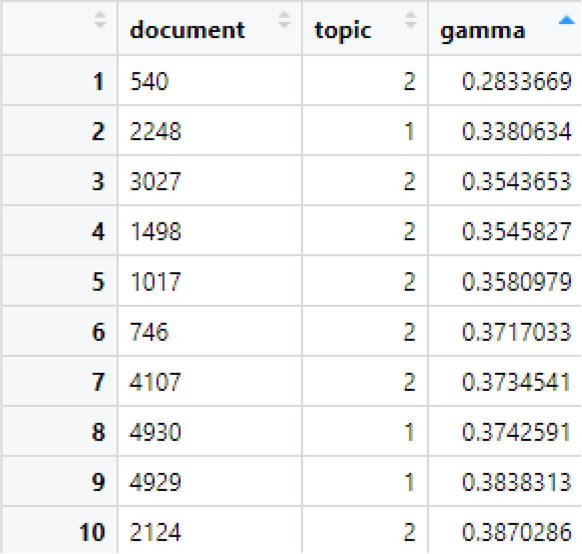
The plot shows the most common terms within the two topics that were extracted from the reviews. The most common words in topic 1 include “chips”, and “coffee”, suggesting that this topic might represent snacks and drinks. In topic 2, it includes “flavor”, “taste”, and “food”, suggesting that this topic could focus on the seasoning of meals from the reviews.

- **Plot 8: Finding the greatest difference between 2 topics(beta)**



By comparing to plot 7, this plot is used to find the words with the greatest differences between the two topics. The words which are more common in topic 2 (log_ratio>5) are “pepper” and “flavor”, and for topic 1 (log_ratio<-5) are “cream” and “drink”. This plot helps confirm that the two topics algorithm identified were snacks drinks and the seasoning of food.

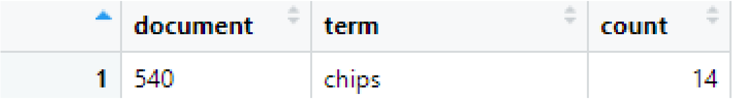
- **Table 1-1: Document-topic probabilities(gamma)**



	document	topic	gamma
1	540	2	0.2833669
2	2248	1	0.3380634
3	3027	2	0.3543653
4	1498	2	0.3545827
5	1017	2	0.3580979
6	746	2	0.3717033
7	4107	2	0.3734541
8	4930	1	0.3742591
9	4929	1	0.3838313
10	2124	2	0.3870286

We examine the per-document-per-topic probabilities - gamma by ascending order, to estimate the proportion of words from each document which are generated from the specific topic. For instance, document 540 estimates only 28% of words which were generated from topic 2. In other words, most of the words in document 540 are generated from topic 1. To check again the results, we create another Table 1-2 to see the most common word in document 540.

- **Table 1-2: Check the most common words in particular document**



	document	term	count
1	540	chips	14

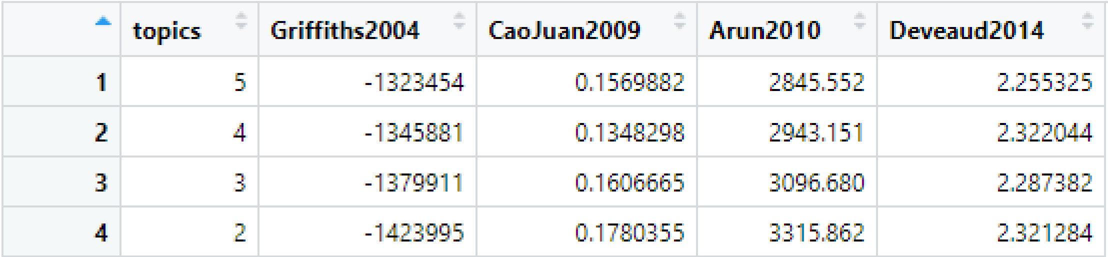
Based on the most common words in document 540, we can see it corresponds to plot 7- Top 10 most common terms of each topic(beta). “Chips” is the most common term in topic 1, so it means the algorithm is correct to place the document in topic 1.

(End of Part I)

Part II

We first set the number of topics in a range of 2-5 to run the four scoring algorithms. Result shows the number of topics and corresponding values of metrics below.

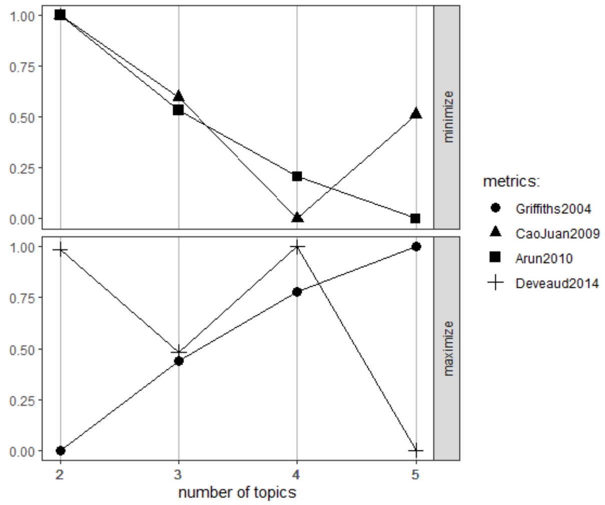
- **Table2: Check the number of topics for LDA model (range2-5)**



	topics	Griffiths2004	CaoJuan2009	Arun2010	Deveaud2014
1	5	-1323454	0.1569882	2845.552	2.255325
2	4	-1345881	0.1348298	2943.151	2.322044
3	3	-1379911	0.1606665	3096.680	2.287382
4	2	-1423995	0.1780355	3315.862	2.321284

Noting that for "CaoJuan2009" and "Arun2010" lower values are better, for “Griffiths2004” and "Deveaud2014" higher values are better. It’s hard to see which topic segmentation is better to describe the data in table1, so we plot a graph to see if there are some interaction points to better analyze the results.

- **Plot 9: To easy analyze results from Table1**

	 **Minimization:** CaoJuan2009 [4] Arun2010 [5] **Maximization:** Deveaud2014 [4] Griffiths2004 [5] From this plot, we can conclude that the optimal number of topics is in the range of 4-5. However, from our perspective, this dataset has the most common words in reviews as sentiment words, like “good”, “love”, “great”, “like”, and “best”, so each topic content may be quite overlapping if we continue making more segments on the topics.
How can the Document Term Matrix be utilized to conduct additional analyses of use to the firm given other data available in the dataset?	The term-document matrix is a two-dimensional matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j. It keeps a track of all the words and their associated frequencies across all the documents. It represents document vectors in matrix form in which the rows correspond to the terms in the document, columns correspond to the documents in the corpus and cells correspond to the frequency of the terms. The firm can utilize this information to identify the relationship between the customer and their reviews. A DTM can help consolidate the vast information into a concise word cloud that the firm can leverage to find out if the reviews from the customers are good or bad. Also, which of it’s food items are more popular or under severe scrutiny depending on the sentiment. **<u>Additional data</u>** : The same analysis can be extended if the data is more accurate and covers a lot of other variables such as Gender of the reviewer, age, date and time, if the review was changed from a positive sentiment to a negative sentiment i.e the higher rating was changed to something low and vice versa.

	Also, the same can be synced up with twitter data for example. This can help the firm identify how the social media is reacting to the items on the menu and if it is what we expected from the reviews data.
Evaluate the usefulness of your analysis in terms of how helpful it is to inform marketing strategy for the company. In addition, make recommendations for how the company could use these results to improve its marketing strategy.	**Evaluation :** With sentiment analysis, we can identify the reviews containing positive or negative sentiments, then determine what customers are happy or unhappy about. In addition to that, it's also critical for us to identify if the negative issue is a pervasive issue. If some negative reviews are common reasons for certain food, we might need to consider further responses to address the issue. For example, should we find alternative products for it or how to improve the quality etc. The same evaluation could also be applied for positive reviews, knowing what makes customers happy and discussing marketing strategy based on it. Moreover, topic modeling can help a business better equipped to choose an approach that is suited to their target and help the process of sifting through large volumes of text data. By conducting the modeling, it enables the viewer to discover hidden thematic structures in text. For example, we may find topics that associated reviews to specific food cuisines (Drink, Snack, Asian Food, Western Food) in this dataset. By allocating different reviews of topic modeling will give a sense to restaurants to explore similar customer preferences. **Recommendation :** Discount campaigns for most popular items on the menu to increase the reach of the product into an untapped market and maintain its popularity in the current market. Examples : tea, chips, coffee. Stitch other dataset from an external vendor or internal repositories that tells us more about the review, like demographics of the person who wrote the review, time of review, if the rating was updated from positive to negative or vice versa. This will help the firm to take a deeper dive into understanding the experience of their customers. For example : Segmentation of popular items and reviews based on gender, age, preferences etc. Identify new audiences, broaden ad placement channels and effectively allocate advertising budgets. For example, tea and coffee might be popular amongst older age groups, but soft drinks might be popular for younger age groups. Arrange ad placements to push these products accordingly.