THE WORLD BANK

Industrial Project: An alternative approach to PPP

Group 2: Mayank Loyalka, Preet Oza, Adhyatma Gautam and Tzuchi Chiu

Table of Contents

INTRODUCTION

Our business objective was to create an alternative approach to traditional PPP (purchasing power parity) method under ICP program of the world bank, for the category of house food consumption goods and items, by eliminating the human factors utilizing different machine learning and text mining techniques available, performing real time analysis on the data, and presenting all our findings into an interactive user interface (UI).

APPROACH

Data Selection

To achieve the business objective, we wanted to test our approach on a smaller subset of food data and successfully scale it to all the food categories in the future. For this we decide to run the analysis on bakery products. This process was further categorized into three levels.

Level 1: Classification of bakery vs non bakery item.
Level 2: Classification of biscuit vs non-biscuit.
Level 3: Further categorization of biscuit vs non-biscuit based on their name.

Process Pipeline

Our process pipeline can be divided into 3 phases as follows:

1. Data collection and data preparation
2. Model selection
3. Model prediction

For phase 1, we built on the web scraper of Cal Poly DxHub, which scrapes the product name from websites and then stores the data in DynamoDB. On the extracted data we performed the following steps Data Cleaning, Building N Grams and Computing TF-IDF.

For phase 2, to identify the best classification algorithm to differentiate between products we used the computed TF-IDF to train machine learning models like Decision trees (bagging trees), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Logistic regression, Neural Network and Random Forest. Furthermore, to understand which metric will determine the algorithm's success, we tested all the models against the accuracy and ROC (Receiver operating characteristics) metric.

For phase 3, After testing all the machine learning models successfully for each level, we used the model with the highest roc and accuracy values. We integrated the respective chosen models in our user interface to make future products predictions.

METHODOLOGY & ANALYSIS

Level 1 Introduction

Once we had a good understanding of what our data set was, we started to work on our Level 1 modeling. The goal here was to create a model that just by looking at a product name can classify if an item is Bakery or Non-Bakery. Since we already had data on Bakery products, we needed some data on Non-Bakery products as well. With the help of Non-Bakery item data, we can gain a better understanding of our Bakery data and create better models. So, we decided to include Rice data as our non-Bakery data just for our level 1 modeling purpose. When we started to perform data understanding. For Level 1 we decided to perform Text analytics. To perform text analytics, we had to make tokens of our product name and then performed the following 4 major steps

Step 1 - Data Cleaning: We had to remove the following words as they do not carry any prediction power.

- Stop words: Stop words are a set of commonly used words in any language. For example, in English, "the", "is" and "and", would easily qualify as stop words.
- Punctuations, Numbers and Symbols: Punctuation marks like (,?," , Symbol and numbers like 1,2,3 needed to be removed.
- Rice and Cookie: Upon further analysis we noticed that words like Cookie and Rice were the most frequent words in each category. Being the most frequent words, they were the most dominant attribute in making predictions. We decided to remove them as we did not want to create a model that just relied on 1 word for making predictions.

Step 2 - Identifying the most frequent words: Our next goal was to identify the most within each category and create a document frequency matrix using them

- Bakery Product: In Bakery product we had 200 different words like Chocolate, Sugar, Sandwich, Chip which appeared a lot (Figure 1).
- Rice Product: In Rice product we had 300 different words like Brown, Grain, Chicken, Vegetables which appeared a lot (Figure 2)

Step 3 - Creating n grams: We created N grams with N=1 and N=2. This helped us to create a model that will make better predictions. After that we computed their prediction power with the help of TF-IDF.

Step 4 - Singular Value Decomposition: Once we calculated TF- IDF we performed Singular Value Decomposition to reduce our attributes from 1900 to 300. Once we had a reduced number of attributes, we ran 6 different machine learning algorithms (Figure 3) and tested all of them against ROC and accuracy metric. We decided to go with Random Forest as it had an accuracy and ROC of 89%.

Level 2 Introduction

Our Level 1 model was able to successfully classify products like Oreo Biscuit, Maria Biscuit, Chocolate ice cream into Bakery products. However, we needed to come up with further classification of those products. Since almost 60% of our data contained details about Biscuit products, we decided to make further classification of those products as Biscuit or Non-Biscuit.

Level 2 Algorithms

Our goal for Level 2 was to create a model that just by product name can classify if a product is Biscuit or Non-Biscuit. To perform text analytics, we made tokens of our product name and performed the following 4 major steps:

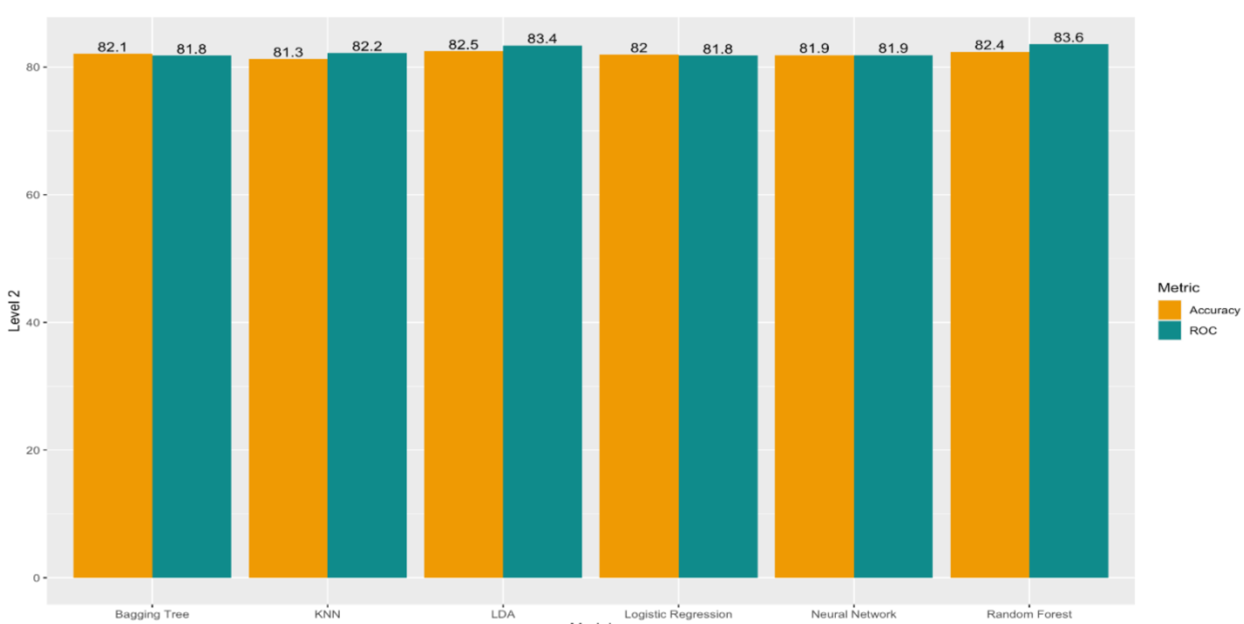
Step 1 - Data Cleaning: Since we were already using the same data set, we did not have to go over the same data cleaning process again.

Step 2 - Identifying the most frequent words: Our next goal was to identify the most within each category and create a document frequency matrix using them

- Biscuit Product: In Biscuit products we had 150 words like Chocolate, Sugar, Sandwich, Chip which appeared a lot (Figure 4).
- Non-Biscuit Product: In Non-Biscuit product we had 100 words like ice, chocolate, dough, cream which appeared a lot (Figure 5).

Step 3 - Creating N grams: We created N grams with N=1 and N=2. This helped us to create a model that will make better predictions. After that we computed their prediction power with the help of TF-IDF (Figure 6).

Step 4 - Singular Value Decomposition: Once we calculated TF-IDF we performed Singular Value Decomposition to reduce our attributes from 1500 to 300. We ran 6 different classification models (As depicted Below) and tested them against the ROC and accuracy metric. We choose Random Forest as the most suitable model because of its high accuracy value



Level 3 Introduction

Our level 2 model was able to classify danish biscuits, Oreo biscuits, and maria biscuits as Biscuits but it was important for us to come up with a way to make further classification of those products. To do that we decide to take an unsupervised machine learning approach for Level 3. Also, Level 3 was the most important from the client perspective, as they will eventually use it to find comparable products. Those products within the same classification will be suitable for calculating the Purchasing Power Parities (PPP).

Level 3 Algorithms

For level 3 the main goal here was to create a model where we'd simply be able to use product names to make our classifications. We first decided to implement clustering analysis to understand our data better. By implementing k means clustering, we would be able to classify products based on the names we gave our clusters. We explored outputs for many numbers of "k" clusters to see if the differences between each cluster would increase. In addition, we also needed to ensure that the cluster names made sense for our classification models. We discovered that the optimal number of clusters was 5. After running cluster analysis, based on our cluster word frequency we named cluster 1 as chocolate, cluster 2 as chocolate chip, cluster 3 as creamme, cluster 4 as peanut butter, and cluster 5 as sugar(Figure 7).

Once we were able to make 5 easily understandable clusters. We decided to now perform supervised machine learning. To perform text analytics, we had to make tokens of our product name and perform the following 4 major steps.

Step 1 - Data Cleaning: Since we were already using the same data set, we did not have to go over the same data cleaning process again

Step 2 - Identifying the most frequent words: Our next goal was to identify the most frequently used words within each cluster and create a document frequency matrix using them.

- Cluster 1 - Chocolate: Among the 100 most frequent words some of them were sandwich, dough, cream, shortbread (Figure 8)
- Cluster 2 - Chocolate Chip: Among the 110 most frequent words some of them were dough, cream, ice, milk (Figure 9)
- Cluster 3 - Creamme: Among the 90 most frequent words some of them were vanilla, fill, sandwich (Figure 10)
- Cluster 4 - Peanut Butter: Among the 80 most frequent words some of them were bar, Jelly (Figure 11)
- Cluster 5 - Sugar: Among the 120 most frequent words some of them were frost, fill, Bake (Figure 12)

Step 3 - Creating n grams: We created N grams with N=1 and N=2. This helped us to create a model that will make better predictions. After that we computed their prediction power with the help of TF-IDF.

Step 4 - Singular Value Decomposition: Once we calculated TF-IDF we performed Singular Value Decomposition to reduce our attributes from 1200 to 300 Once we had a reduced number of attributes, we ran 6 different machine learning algorithms and tested all of them against ROC and accuracy metric. We decided to go with Random Forest as it had an accuracy of 89% and roc of 88%. (Figure 13)

USER INTERFACE OVERVIEW

To culminate what we have accomplished for this project, we created a user interface (UI) that allows for the World Bank to easily classify products for all 3 levels. Our user interface requires the user to upload a CSV file and it would display the dataset and the 3 levels of classification. In addition to this, we built our models to output probabilities that would also be displayed for the respective product in each level.

FUTURE OVERVIEW

One of the biggest challenges that the world bank faced was taking language into account while making predictions. In future we believe 2 AWS services like AWS Rekognition and AWS Translate will help to solve this problem. After both the services will be incorporated this is what the new workflow will look like

1. By using the AWS Rekognition "Text in image" features, we can upload the product's packages images that we want to analyze on the Rekognition, and AWS Rekognition will identify the text on the image and display the detected texts on the image and transform them to the texts. Although they can detect the words in different languages, but it doesn't automatically translate the foreign languages into English, there are still some non-English words on the package.

2. With the AWS Translate, which is a text translation service that uses advanced machine learning technologies to provide high-quality translation on demands. So, after recognizing the texts on the product's package, we are going to use the AWS Real-Time translation service and copy paste the non-English words to the platform, and then with this smart service it will translate the detected language to English which we can eventually feed those English words into our model.

CONCLUSION

Our goals for the project will be creating: Fast, Smart, and Automated process for PPP calculation. In this case, when the World Bank wants to find the products that are comparable across countries, they can simply drag and drop product labels into AWS Rekognition that will automatically recognize product descriptions and translate them into different languages if needed. Then using the User Interface, it will help to identify the following:

- Is the product a Bakery item or Non-Bakery item?
- If the product is a Bakery item, then is it a Biscuit or Non-Biscuit item?
- Making further classification of that product

Going forward, we hope that the User Interface designed in incorporation with AWS services will simplify the data collection process to meet the goals of the ICP program. This will be a big milestone for the World Bank and their ICP project to leap forward. We are very much looking forward to the World Bank to use our technology to better analyze the countries' GDPs and purchasing power in the future to reach their bigger goal.

GLOSSARY

ROC	ROC curve, also known as Receiver Operating Characteristics Curve, is a metric used to measure the performance of a classifier model. IT depicts the rate of true positives with respect to the rate of false positives, therefore highlighting the sensitivity of the classifier model.
ACCURACY	Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions.
TF-IDF	In information retrieval Tf-idf, is short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
TOKENS	Tokens are the individual units of meaning you're operating on. This can be words, phonemes, or even full sentences. Tokenization is the process of breaking text documents apart into those pieces.
DOCUMENT FREQUENCY MATRIX	A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.
SINGULAR VALUE DECOMPOSITION	The goal of SVD is to find the optimal set of factors that best predict the outcome.
N GRAMS	They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward (although you can move X words forward in more advanced scenarios)
BAGGING TREE	Bagging classification model uses a simple approach of improving the estimate of one by combining the estimates of many. Bagging constructs n classification trees using bootstrap sampling of the training data and then combines their prediction to produce a final metadata.
KNN	K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).
LDA	Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes.
LRM	Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.
NEURAL NETWORK	A neural network is a series of algorithms that recognize underlying relationships in a set of data through a process that imitates the way the human brain operates. The artificial neural network (ANN) assimilates data in the same way the human brain processes information.
RANDOM FOREST	Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

APPENDIX

Figure 1.

Figure 2.

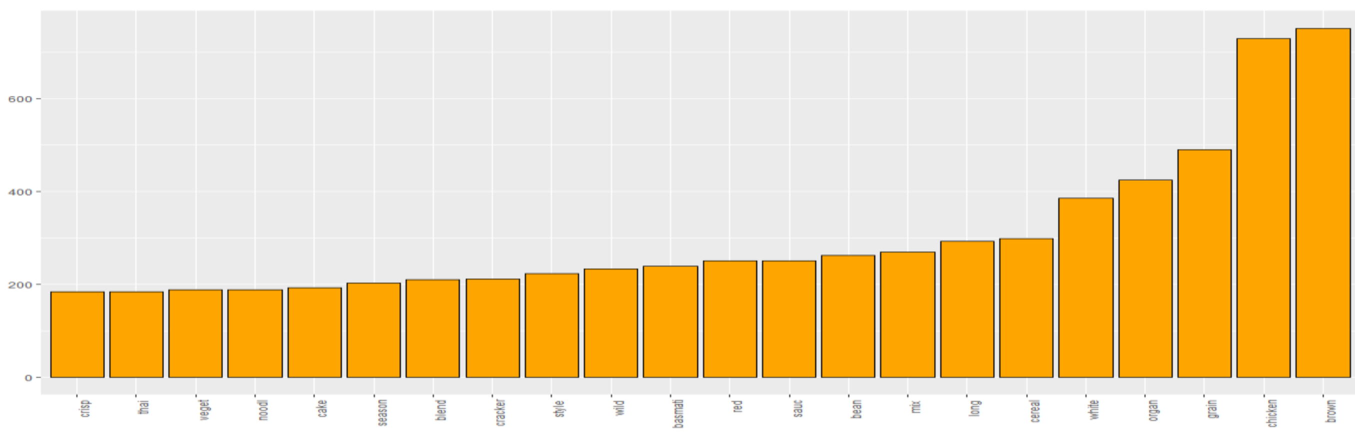


Figure 3.

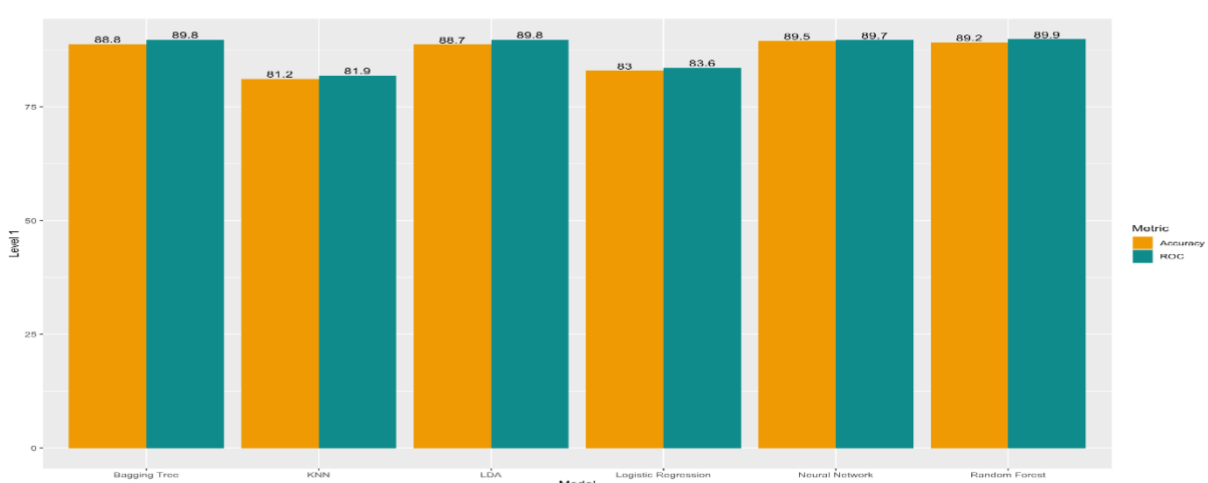


Figure 4.

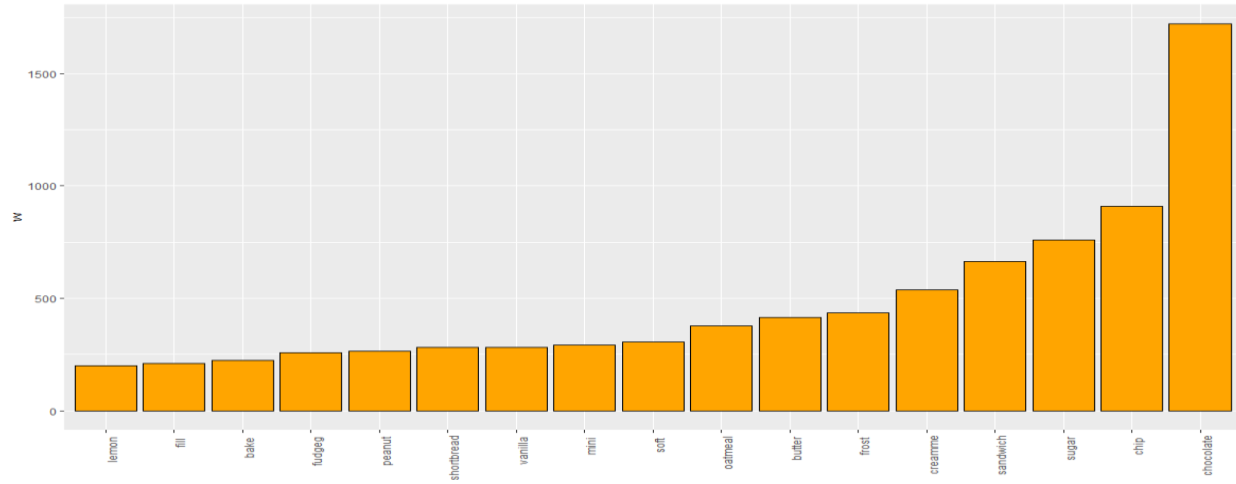


Figure 5.

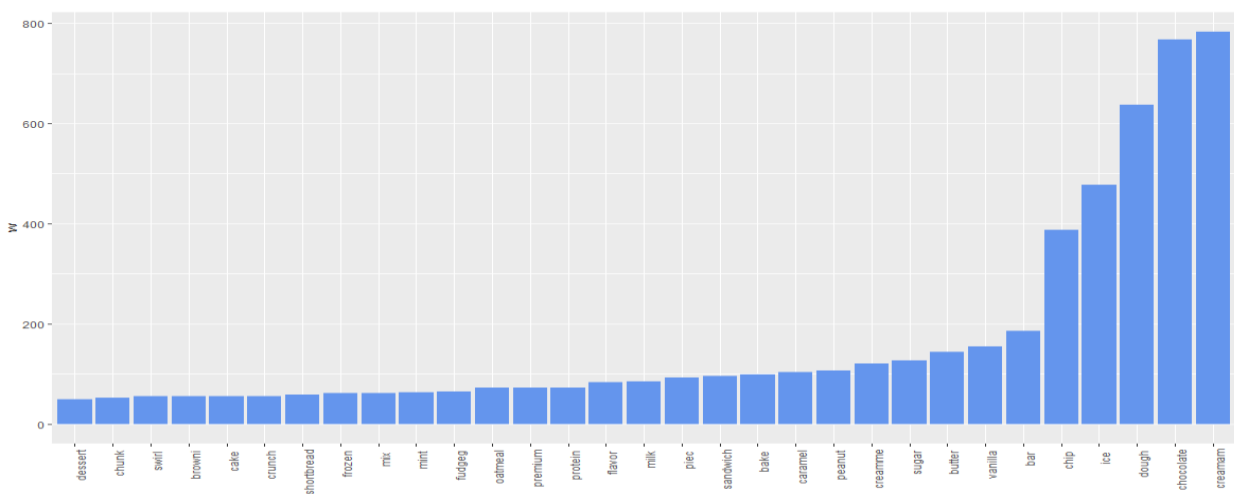


Figure 6.

```
 [1] "cooki"          "cream"            "ice"
 [4] "cream"          "cooki"            "dough"
 [7] "piec"           "wrap"             "dark"
[10] "chocol"         "coat"             "cooki"
[13] "crumbl"         "top"              "white"
[16] "candi"          "ribbon"           "premium"
[19] "ice"            "cream"            "bar"
[22] "cooki"          "dough"            "crunch"
[25] "cooki_cream"    "cream_ice"        "ice_cream"
[28] "cream_cooki"    "cooki_dough"      "dough_piec"
[31] "piec_wrap"      "wrap_dark"        "dark_chocol"
[34] "chocol_coat"    "coat_cooki"       "cooki_crumbl"
[37] "crumbl_top"     "top_white"        "white_candi"
[40] "candi_ribbon"   "ribbon_premium"   "premium_ice"
[43] "ice_cream"      "cream_bar"        "bar_cooki"
[46] "cooki_dough"    "dough_crunch"
>
```

Figure 7.

	cooki	oreo	vanilla	chocol	creme	fill
1	1.036694	0.03020756	0.034099333	0.11545589	0.000000000	0.020941438
2	1.146444	0.00167364	0.058577406	1.36903766	0.016736402	0.015899582
3	1.063551	0.03177570	0.271028037	0.24859813	1.125233645	0.164485981
4	1.055363	0.02422145	0.006920415	0.28027682	0.103806228	0.038062284
5	1.074627	0.00000000	0.044776119	0.01865672	0.004975124	0.001243781

	butter	peanut	oatmeal	raisin	shortbread	chip
1	0.031134173	0.002779837	0.07561156	0.036323202	0.060229800	0.019088213
2	0.008368201	0.004184100	0.01506276	0.000000000	0.005020921	0.958995816
3	0.005607477	0.000000000	0.01308411	0.001869159	0.005607477	0.003738318
4	1.290657439	1.204152249	0.01384083	0.003460208	0.000000000	0.148788927
5	0.006218905	0.002487562	0.01368159	0.000000000	0.007462687	0.003731343

	sandwich	gourmet	caramel	lemon	mini	white
1	0.060044477	0.030948851	0.030578206	0.02557450	0.03613788	0.023906597
2	0.067782427	0.025104603	0.045188285	0.00167364	0.05941423	0.040167364
3	0.583177570	0.005607477	0.007476636	0.10093458	0.03738318	0.003738318
4	0.131487889	0.010380623	0.010380623	0.00000000	0.03806228	0.003460208
5	0.004975124	0.012437811	0.003731343	0.02985075	0.04228856	0.016169154

	bake	soft	ice	bite	fudg	dough
1	0.03521127	0.038176427	0.07005189	0.02390660	0.043365456	0.047813195
2	0.05690377	0.036820084	0.13054393	0.03514644	0.019246862	0.265271967
3	0.00000000	0.001869159	0.01308411	0.01121495	0.059813084	0.001869159
4	0.10034602	0.055363322	0.05536332	0.02768166	0.093425606	0.093425606
5	0.04601990	0.054726368	0.02238806	0.00000000	0.007462687	0.049751244

	sugar	doubl	cream	bar	bakeri	frost
1	0.000000000	0.015381764	0.11452928	0.023906597	0.021312083	0.01037806
2	0.010041841	0.070292887	0.17405858	0.057740586	0.009205021	0.00334728
3	0.009345794	0.046728972	0.01308411	0.028037383	0.005607477	0.02803738
4	0.010380623	0.006920415	0.05536332	0.076124567	0.020761246	0.00000000
5	1.077114428	0.000000000	0.01243781	0.004975124	0.043532338	0.49129353

Figure 8.

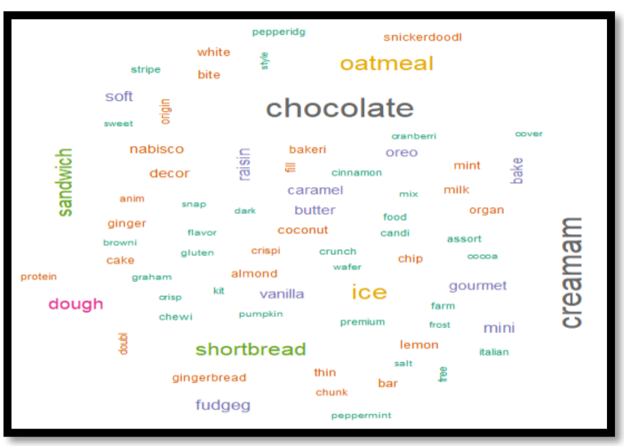


Figure 9.

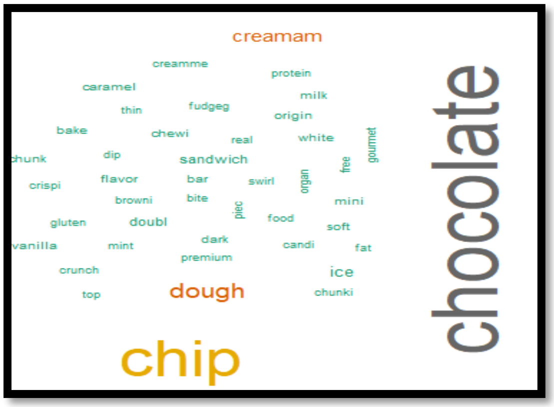


Figure 10.

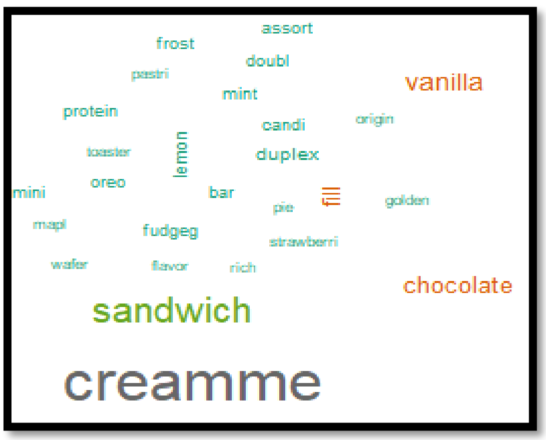


Figure 11.

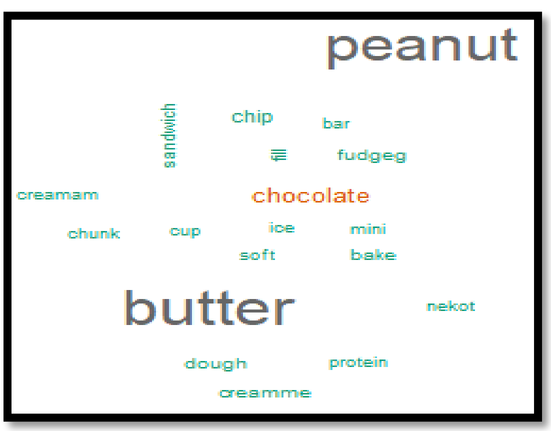


Figure 12.

Figure 13.