**Group 3 Project Journal**

**Soh Kai Jin U1740760K**

**Justin Chia Cheng Ze U1740466J**

**Chong Shu Chen U1740627B**

**Lim Rui Jia, Hannah U1640334B**

11th October 2019

Ideas:

1. Predict whether there are ingredients to cook the dish

* List out key ingredients present in your house
* Machine learning to recommend corresponding recipes
* Using big recipe dataset

1. Recipe summarisation (Dataset of bunch of recipes, give out summary of recipe)   
   Purpose: saving time

<https://github.com/rtlee9/recipe-summarization>

Pros: Found data, purposeful

Cons: Not as interesting :D but can consider if first idea is rejected

1. Given the ingredients of a recipe, predict what country’s cuisine is the dish

<https://www.kaggle.com/kaggle/recipe-ingredients-dataset>

<https://www.tutorialspoint.com/r/r_json_files.htm> (for reading json files into R)

Pros: Found dataset

Cons: Real life importance and applicability not significant

Allocated roles :

1. Report writing (initial phase, R markdown) = Hannah
2. Creating the output (Machine Learning) = Hannah, Kai Jin, Shu Chen

* By categories (Categories)
* By ingredients (Ingredients)
* Output is “Categories U Ingredients” (Like a Venn Diagram)

1. R Shiny = Justin

14th October 2019

Re-discussing theme.

**Suggestion**

Given ingredients, machine learning to predict the type of cuisine

Output: recommend potential food the person will like from another type of cuisine (?), based on similarity index

Purpose: Predict Origin of Cuisine

E.g. Japanese/ African/ Irish/ Greek

Output: Predicted cuisine

Input: Set of ingredients

Dataset

* Train -> create model, fit to whatever ingredients fed into the system
* Test -> used to check whether it is trained properly

2nd output: Recipe recommendation for another highly similar cuisine (bipartite)

18th October 2019

Due to time and feasibility constraints, we decided to forgo the second output which would involve a bipartite problem. Our group agrees that we should focus on machine learning for cuisine prediction.

Re-allocated workload and completely revamped IPIR.

New Allocated roles :

1. Report writing = Hannah(overall organisation and grammatical correction), Everybody(content)
2. Creating the output (Machine Learning) = Kai Jin, Shu Chen, Justin, Hannah
3. R Shiny = Justin

Officially started work on project.

4 November 2019

Meeting to clarify and agreed on possible removal of R shiny component due to time constraints.

Discussed the code for machine learning:

1. Preliminary data analysis
2. Cleaning of data - removal of punctuation, numbers, spaces. Reduced the word to its root.
3. Corpus and Word cloud
4. CART model

8 November 2019

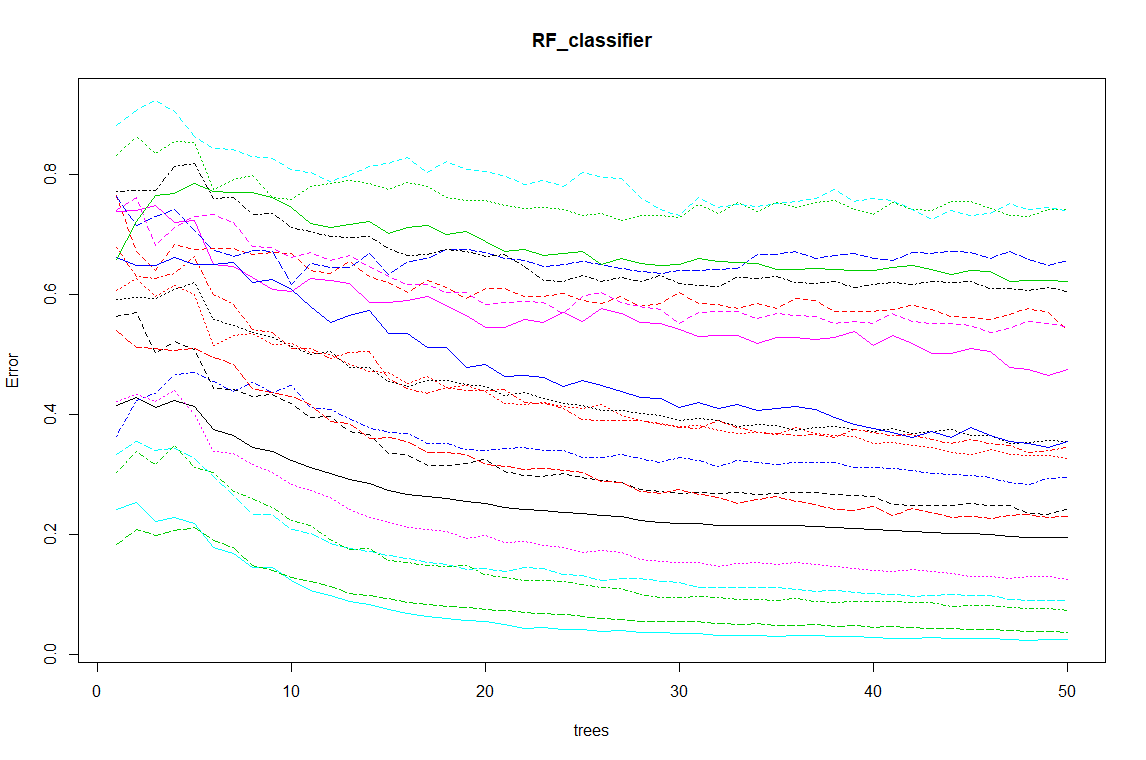
Meeting to consolidate code, debug errors and improve on the dataset. Preliminary images for graphs and the sort were re-evaluated and replaced with more aesthetically pleasing ones.

Due to our progress rate, R shiny was included in the end as per stated in the IPIR.

Achieved R shiny interface, however the ingredients appeared as truncated words, some with an additional “c”.

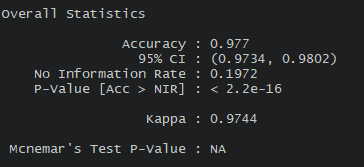
We returned back to earlier data processing steps, aiming to remove sparse terms in the dataset to clean it.

An issue with this dataset, which is commonly encountered during natural language processing, was the identification and removal of descriptives. In the Kaggle dataset, the manner in which the ingredient should be prepared is incorporated into the recipe. For instance, “diced carrots” instead of simply “carrot”.  To tackle this, we assumed that such adjectives appeared in a frequency lower than ingredient names. Our group then attempted to remove low frequency words using the *removeSparseTerms* function.

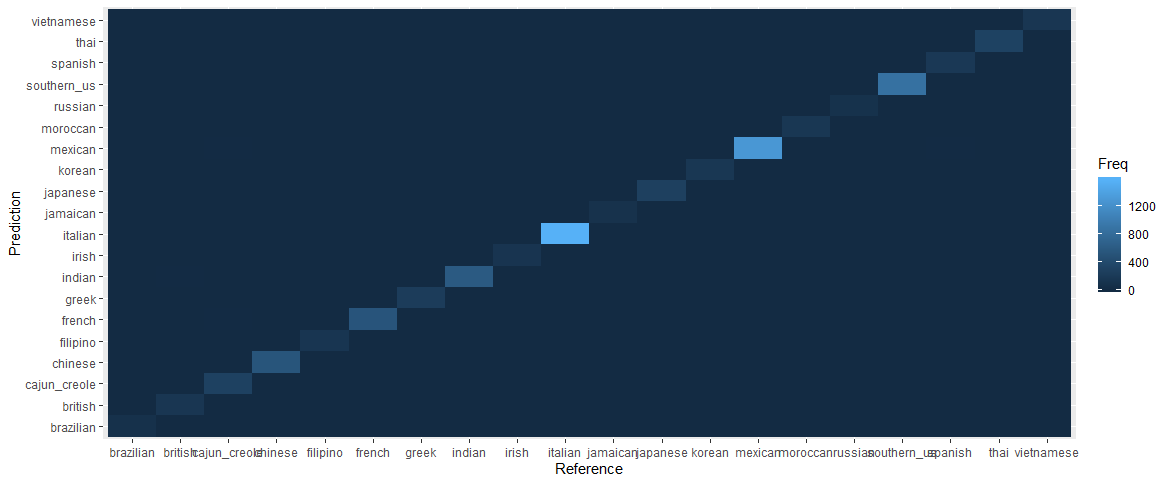
**

*Without sparsity removal. 50 trees*

Initially, we performed sparsity removal 50 trees, as shown below. However, we then decided to increase the number of trees to 100 to improve accuracy.



*Statistics after sparsity removal with 50 trees.*



*Heatmap of sparsity removed with 50 trees.*

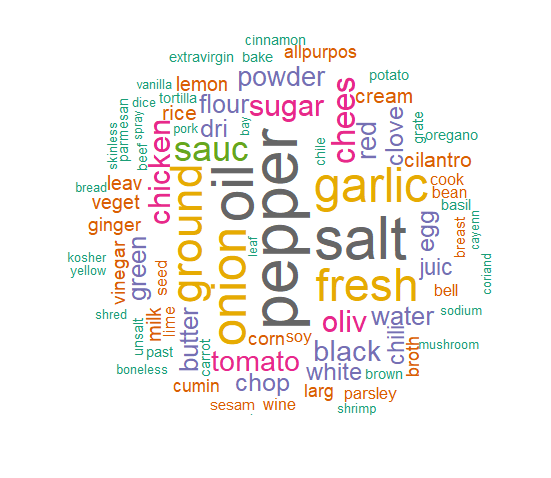
Parts for presentation were agreed upon and allocated as follows:

1. Introduction & Exploratory Data analysis - Hannah
2. Exploratory - Shu Chen
3. Machine Learning - Kai Jin
4. R shiny - Justin
5. Evaluation - Justin
6. Future Directions - Hannah

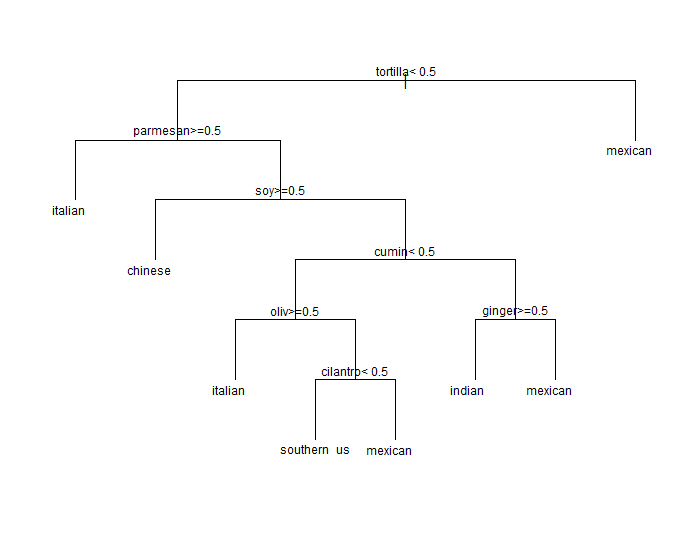
10 November 2019

Following the lecture on machine learning, we realised that we needed to balance our dataset by under-sampling recipes from overly represented cuisines such as Italian, Mexican and Southern US. This stems from the observation that the over-represented cuisines were more likely to the predicted than less represented cuisines. We found the average number of recipes per cuisine to be approximately 2000 recipes and proceed to undersample from abundant cuisines.

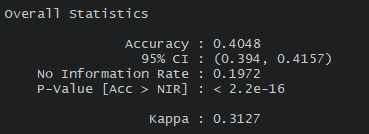
The following figures are **old figures** generated **prior to undersampling** of train.json.



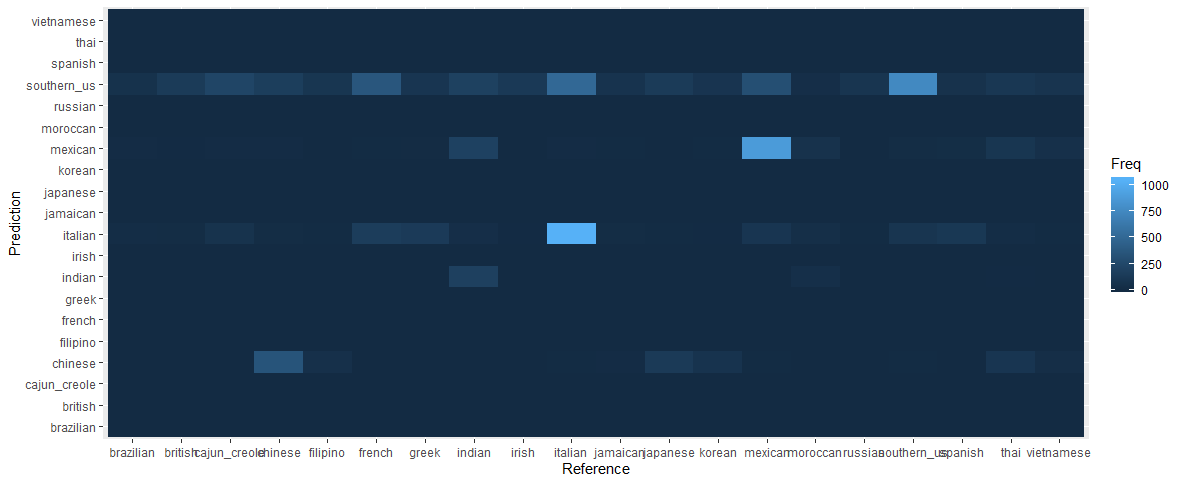
*Word cloud generated from the training data set after performing the stemDocument() function. Higher frequency words are depicted in the center while low frequency words are positioned at the periphery.*



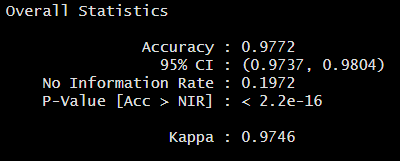
*CART model for classifying the type of cuisines according to the percentage of a certain ingredients within the recipes of that particular cuisine.*



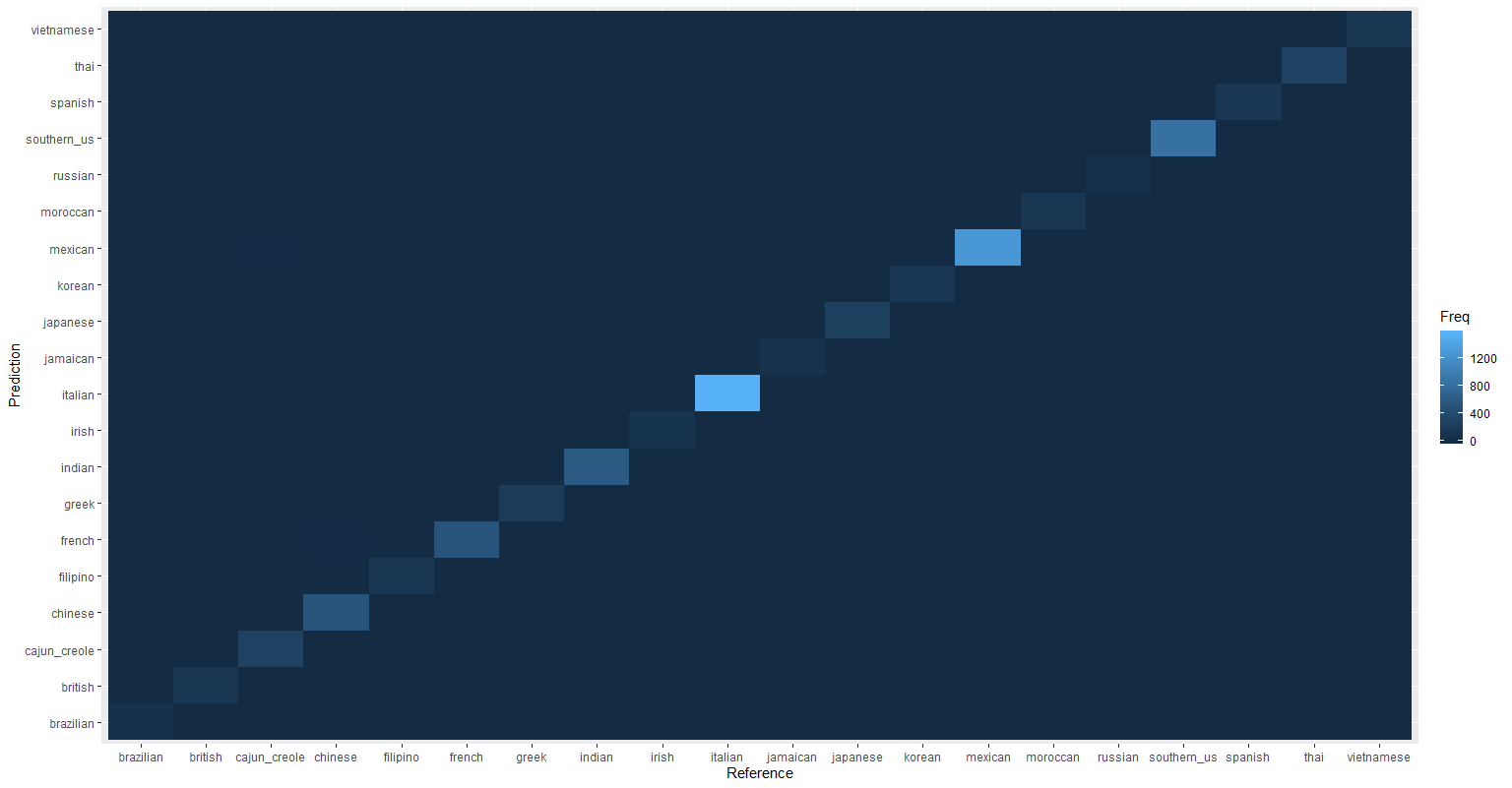
*Statistics of the decision tree*



*The multiclass confusion matrix comparing predictions provided by the model in the y-axis against actual type of cuisine shown in the x-axis. Note that Italian cuisine has highest frequency.*

**

*Statistics of the random forest model*

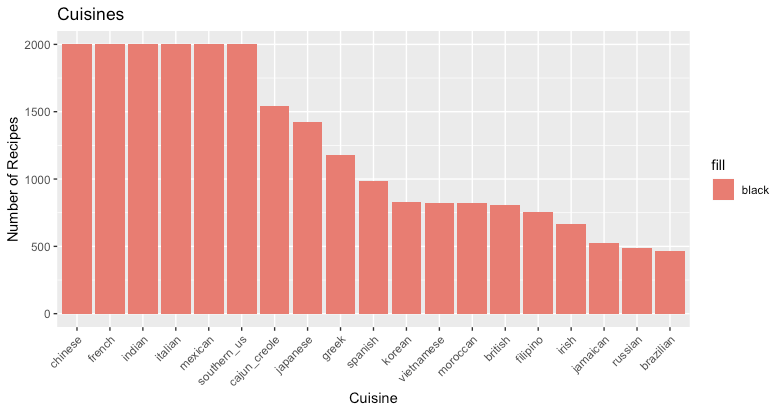
**

*The multiclass confusion matrix comparing the prediction of the cuisine generated by the random forest algorithm in the y-axis against the actual cuisine as the reference in the x-axis.*

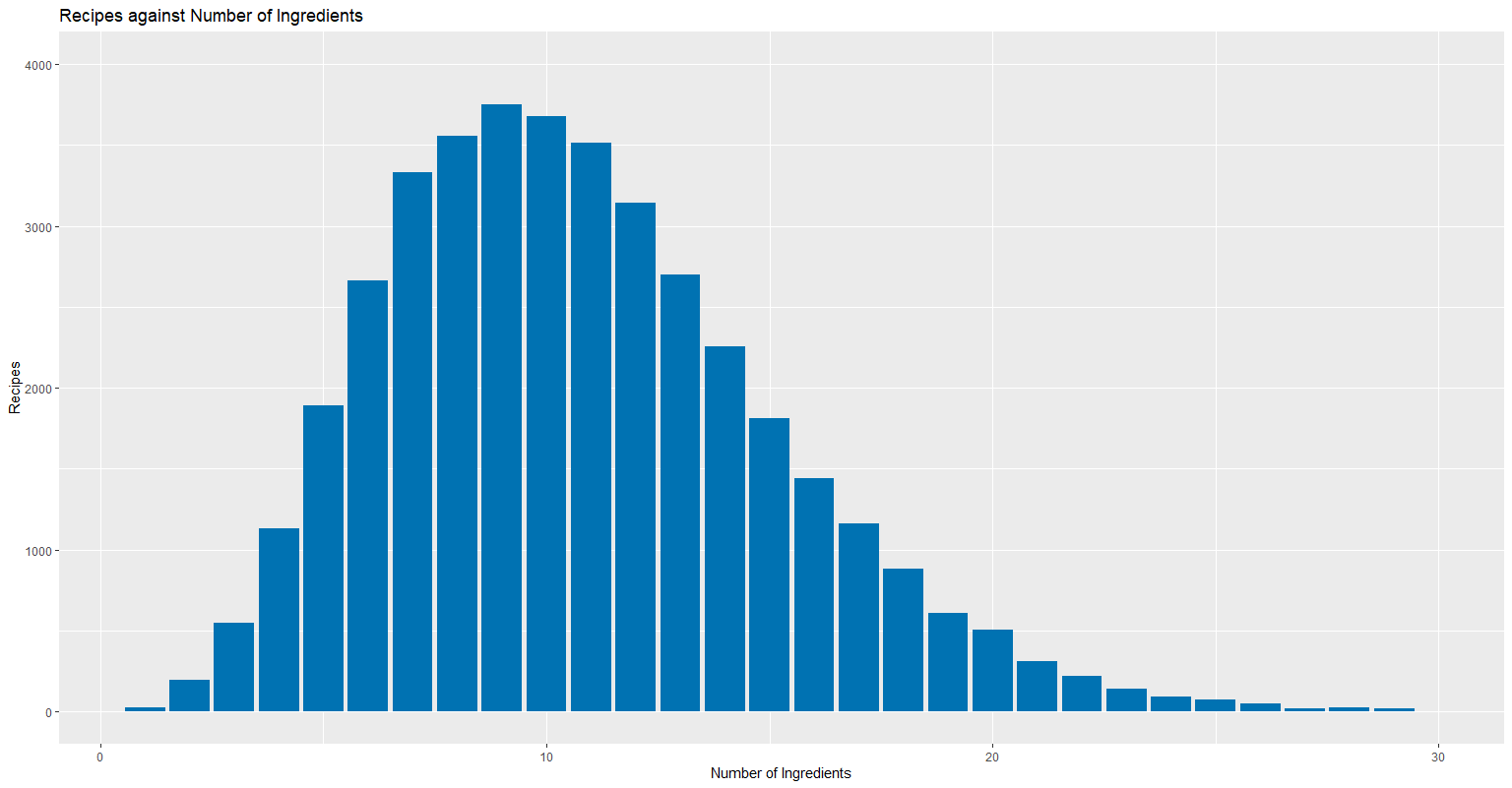
13 November 2019

Another meeting to discuss some discrepancies in the dataset. There were NAs when trying to balance the dataset. Realised that thai cuisine was not included some figures and had to rectify issues stemming from earlier parts.

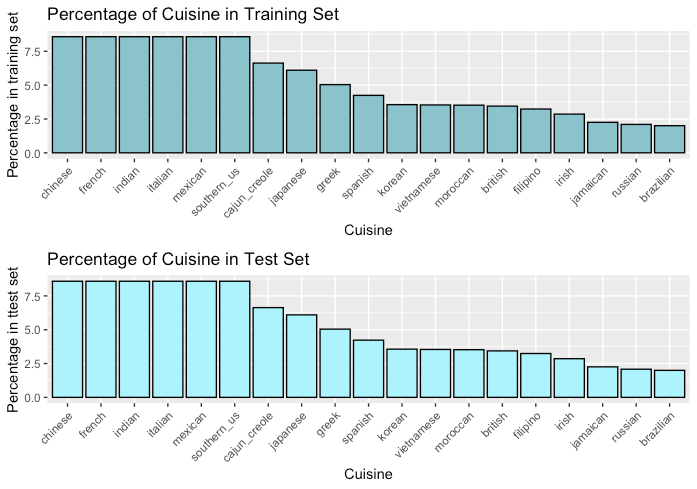
Below are the **older figures without thai cuisine**.

****

*Figure 4. Histogram of the number of recipes of each type of cuisine from the train.json dataset, with the top 6 cuisines (Chinese, French, Indian, Italian, Mexican, Southern\_US) undersampled to balance the different classes.(need to add in thai)*



*Histogram of the number of recipes with the corresponding number of ingredients. This histogram was generated with the original train.json dataset.(need to add in thai)*

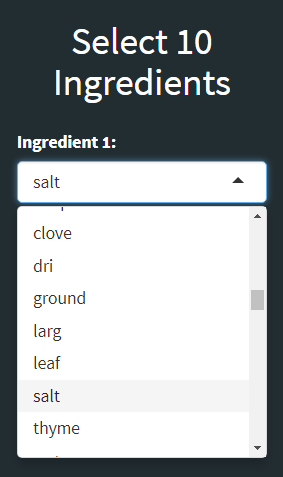


*Barchart of the percentage of each cuisine in both training set and test set (need to add in thai)*

14 November 2019

Presentation slides finalised.

Following the presentation of our report, we decided to revisit and change a few things in RShiny app and remove adjectives from the ingredients list.



*The initial R Shiny interface drop down box utilising the list of of 86 ingredients. Note the truncated ingredient names and the presence of descriptives such as ground.*

In order to subvert the issue of descriptives appearing in the dropdown box, we chose to manually remove descriptions from the list of 86 ingredients. Root words from the *stemdocument()* function were also manually recorrected into actual full ingredient words. The final list of 57 ingredients was used to update the R Shiny dashboard.

The final code was consolidated and commented on.

17 November 2019

Few more future directions added into the report, regarding the determination of the quantity of ingredients to use.

20 November 2019

Minor edits done to report. Group report finalized.