What shall we cook tonight?

by Hanying Ji, Anshuma Chandak, Lan Wen, Xueying Ding, Qianhui Wan

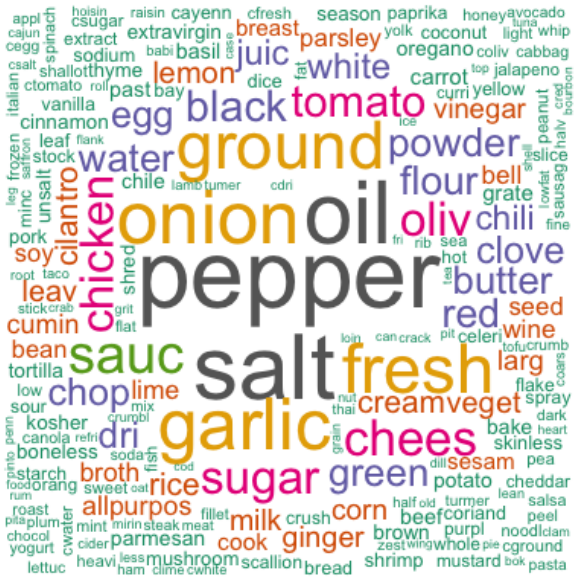
Food is an innate part of any culture or region. Every cuisine has some unique ingredients or some ingredients that are used in almost all of its dishes.

If you visit Korea, the markets would be sprawled with kimchi and the smell of squids. The colorful and aromatic spice markets of India indicate the natural use of diverse spices in the Indian cooking. In this project, we predict the category of a dish's cuisine given a list of its ingredients. We are using Yummly's data which is arranged by cuisine, dish ID, and its ingredients. We started off with 6849 ingredients and used the **bag-of-words model** to reduce the number of ingredients to 2000. Aiming to combine and use the knowledge from other projects in this course, and build a product that has high functionality and usability, we have divided this project into two parts:

The first part is to use different algorithms (**Random Forests, XGBoost, SVM, Logistic Regression, Decision Tree, KNN**) to predict the category of a dish and aim at improving the accuracy the prediction. Following is our algorithms summary, from which we can see that **Logistic Regression** outperforms others under this situation:

Classifier	Parameter	Parameter Explaination	Test Accuracy
XGBoost	max.depth = 10 eta = 0.1 nthread = 2 nround = 50	maximum depth of the tree step size shrinkage number of parallel threads used to run xgboost number of rounds for boosting	**67.83%**
Decision Trees	cp = 0.001 usesurrogate = 2	complexity parameter how to use surrogates in the splitting process	**55.55%**
SVM RBF Kernel	c = 100 gamma = 0.001	penalty parameter of the error term kernel coefficient for "rbf"	**77.63%**
Logistic	**penalty = 'l2'**	**norm used in the penalization**	**78.51%**
KNN	k = 7	number of neighbors	**63.43%**
Random Forest	n_estimators = 500 max_depth = 140 max_features = 40	number of trees in the forest maximum depth of the tree number of features to consider	**75.87%**

The following word clouds show the different cuisines, and the assortment of ingredients in our data set (the size reflects the number of recipes.



In our project, we wish to apply a superior model to predict a specific cuisine given various different ingredients we put in. Image when it's dinner time, you open the refrigerator and see there are kimchi, romaine lettuce inside. You wonder what to cook for tonight. So the shiny app could provide you a useful cooking recommendations given the ingredients you have. From the six algorithms we used in previous part, logistic regression model performs the best. For example, if you have kimchi and romaine lettuces as your inputs, you will get a predicted "Korean" cuisine result.

We also introduced Pearson Correlation algorithm to calculate the similarities between all 20 different cuisines in our dataset. So if you get a most relevant recommendation given the ingredients you have, the app could also provide other two most similar cuisines for your consideration.