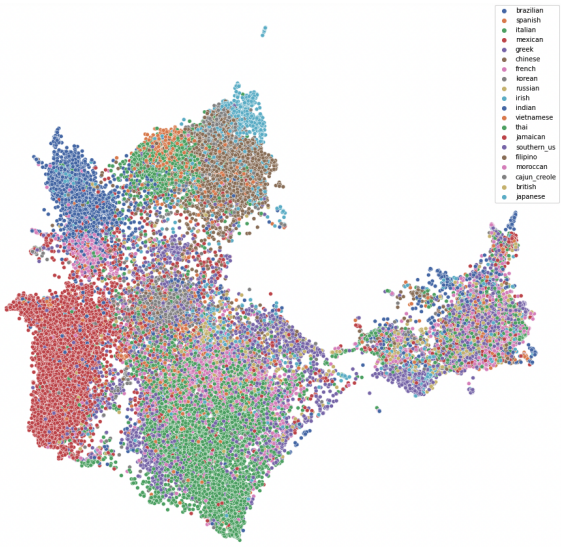
Executive Summary of the May 2022 Boot Camp Project: Cuisine Predictor
Team Clementine: Zhi Jiang, Jimin Kim, Jason Lee, Yumeng Li, Arpan Pal

Problem Statement: Cuisines vary significantly across the countries. One of the key characteristics that distinguish different types of cuisines is the ingredients used in each dish. The goal of the project is to identify cuisine origin using only the ingredients. Accurately associating ingredients to cuisines will be a valuable tool to better understand the characteristics of each type of cuisine as well as creating an algorithm that can recommend recipes using different combinations of raw ingredients.

Data Description: We use a dataset provided by Kaggle (competition: “What's Cooking?”) that contains cuisine origin and ingredients of roughly 40,000 recipes. There are 20 types of cuisines in total and the number of unique ingredients is roughly 7,000.

Data Processing: The initial process is to vectorize the ingredient list to use it in the machine learning models. This processing achieves two main objectives. First, we convert natural language into more analyzable numerical values. Second, we lower the dimensionality of the ingredients from 7,000 unique values to 100-300 dimensions. We utilize the Word2Vec algorithm that uses a neural network model to achieve this goal. Alternatively, we also try principal component analysis (PCA) from the one-hot encoding of the data.



Classification method: We use vectorized data to train our model. We explore multiple machine learning models: Neural network, Support Vector Machine, Random Forest, and XGBoost. We find that the approach using the Fine Tuning SVC in combination with the Word2Vec-300 and PCA gives the highest accuracy of 80.01% when predicting the cuisine on a validation set.

Value: Identifying cuisines using raw ingredients will be valuable to companies that build recipe databases. It helps an easier categorization of recipes, and can be a key instrument to creating a recipe recommendation tool for the end users. Word2Vec can still vectorize a new ingredient that is never seen in the training set.